PROCEEDINGS

MTA
THE MILITARY TESTING ASSOCIATION

27th Annual Conference of the

MILITARY TESTING ASSOCIATION

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
SAN DIEGO, CALIFORNIA

Coordinated by the
NAVY PERSONNEL RESEARCH & DEVELOPMENT CENTER

SAN DIEGO, CALIFORNIA
21 - 25 OCTOBER 1985

UNITED STATES NAVY

Volume I

PROCEEDINGS

27TH ANNUAL CONFERENCE

of the

MILITARY TESTING ASSOCIATION

Coordinated by

the

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER

SAN DIEGO, CALIFORNIA

21-25 OCTOBER 1985

VOLUME I

FOREWORD

The papers presented at the Twenty-Seventh Annual Conference of the Military Testing Association came from the business, educational, and military communities, both foreign and domestic. The papers reflect the opinions of their authors only and are not to be construed as the official policy of any institution, government, or branch of the armed services.

CONTENTS

VOLUME I

CONTENTS

VOLUME II

XXI

p. xxiii

OPENING SESSION OF THE 27TH ANNUAL MILITARY TESTING ASSOCIATION CONFERENCE

21 October 1985

The 27th Annual Conference of the Military Testing Association was hosted by the Navy Personnel Research and Development Center (NPRDC). The Conference was held at the Bahia Hotel in San Diego, California, 21 through 25 October 1985. A total of 169 paper and symposium presentations were given during 57 sessions structured into three concurrent tracks. Conference attendance was 326.

CALL TO ORDER: Dr. Martin F. Wiskoff, Head, Manpower and Personnel Laboratory, NPRDC, and MTA Chairman called the Conference to order at 1300, 21 October. Dr. Wiskoff then introduced the MTA President, CAPT Howard S. Eldredge, Commanding Officer, NPRDC.

WELCOME: CAPT Eldredge officially welcomed the attendees to the Conference and San Diego. He discussed the importance of personnel systems to missions of the armed forces and emphasized how they represent the key to military superiority. CAPT Eldredge then expressed the hope that the formal presentations and the informal interchanges the attendees would experience during the week would go far to further personnel research and development.

KEYNOTE: Prior to the introduction of the keynote speaker, Brigadier General Caleb J. Archer, Commander, U.S. Military Entrance Processing Command, Dr. Wiskoff gave a brief summary of the General's career. Dr. Wiskoff noted that General Archer has held a wide variety of important command and staff positions. In 1966 he served in Vietnam commanding the 212th Military Police Company and later was Chief of Physical Security for the U.S. Army Vietnam. General Archer also served on the faculty of the Army Command and General Staff College and has served as Provost Marshal of the 3rd Infantry Division in Wurzburg, Germany and Commander of the 793rd Military Police Battalion, Nuremburg, Germany. General Archer has also served in Washington, DC as the Army Deputy Chief of Staff for Personnel. He has been the Provost Marshal of the Army Field Artillery Center at Fort Sill, Oklahoma. His most recent assignments have included Commander, Western Region Recruiting Command at the Presidio of San Francisco, Commandant of the Army Military Police School and Deputy Commanding General of the Army Military Police and Chemical Training Centers, Fort McClellan, Alabama

Dr. Wiskoff then asked the attendees to give a warm welcome to General Archer. (General Archer's address follows in its entirety.)

THE U. S. MILITARY ENTRANCE PROCESSING COMMAND AND ITS TESTING PROGRAM FOR THE AIDS VIRUS

Brigadier General Caleb J. Archer
U. S. Military Entrance Processing
Command

CAPT Eldredge, Dr. Lancaster, Dr. Wiskoff, ladies and gentlemen.

It is a distinct privilege and honor for me to be invited to speak to this group of researchers and professionals. Although I've only been the Commander of the U. S. Military Entrance Processing Command (USMEPCOM) for three months, I've seen already what an important part research plays in our command and, in fact, the whole Armed Forces. Your efforts in developing the various forms of testing are very important to us. As past Commandant of the Military Police School I can also assure you of the importance of other areas of personnel research and development such as your efforts involving women in the Armed Services.

At this time we in USMEPCOM are faced with the complex issue of the Acquired Immune Deficiency Syndrome (AIDS). It is an issue that will require immense effort to see the problem to a successful solution. We'll need your help.

What I would like to do today is first to tell you a little bit about the Command and then to talk about the mission we have received in USMEPCOM to test all new applicants to the Armed Forces for the HTLV-III antibody related to the AIDS virus.

USMEPCOM reports directly to the Department of Defense through Dr. Steve Sellman and Dr. Lancaster's office, and up to LTG Chavarrie. We have a very direct route through Headquarters DCSPER because they are the executive agent for processing for the Armed Forces. Our biggest customers are the four recruiting commands - Army, Navy, Air Force, and Marine Corps - and the fourteen training centers - eight Army, three Navy, 1 Air Force, and 2 Marine Corps. We qualify applicants for this complex system.

We have 68 Military Entrance Processing Stations (MEPS) throughout the U. S. and overseas and two substations - one in Guam and one in Alaska. This structure is broken down into three sectors - the Western Sector under Marine Corps COL Bill Stroup, headquartered in San Francisco; the Central Sector under Army COL James Tyler, headquartered in Chicago; and the Eastern Sector under Air Force COL Linda Sendt at Fort Meade, Maryland. We are a joint staff agency with about 49%, Army, 21% Navy, 19% Air Force, and 11% Marine Corps. The service affiliations of our commanders reflect these service percentages. However, we have 24 female commanders out of a total of 68 because all of our Navy commanders are female. This is because the Navy is seeking command billets for this level of female officer and, fortunately for us, we are able to take advantage of the situation.

Our screening process is based on the Armed Forces Vocational Aptitude Battery (ASVAB). We administer the ASVAB to about one million applicants at our testing sites each year and we administer an additional million in high schools. So we give the test about two million times each year.

We give about 540,000 physical exams at our stations a year but next year that will rise to 700,000 because we are taking on the responsibility of examining for the Army National Guard starting 1 December.

In addition to testing, USMEPCOM interviews applicants to detect law violations and other disqualifying past conduct.

Out of the total, the selection process pares the number down to about 380,000 that actually join the Armed Forces through the Military Entrance Processing Stations (MEPS) per year. All but about 5% of the total enter the Delayed Entry Program. This program allows the participants to delay entrance for up to a year. Currently the program has about 130,000 in it but this number varies up to about 150,000. So we are always working with the future in mind in our enlistment process.

For our production tests, the ones we give to the million applicants, we use the ASVAB 11, 12, and 13 forms which many of you have worked on. For the school test, administered in the high schools and junior colleges, we use ASVAB 14. With the help of Dr. Lancaster and others, we have just got testing specialists in all 81 of our stations. These specialists help market ASVAB to the schools. We market it as the excellent counseling tool it is but, of course, it also has great value to the recruiters. It allows the recruiters to direct their attention to the high quality youngsters the military needs to enlist.

We have some difficulty getting into some of the high schools to test. We give the ASVAB in about 79% of the schools but some of them test only three to five examinees so the totals don't add up very fast. We only test about 13% of the available high school juniors and seniors in the country. Endorsements from state agencies, city officials, and school administrators help the program, but the strong feeling that any outside activity detracts from the education process plays against our initiatives.

We also administer special purpose tests at the time of enlistment for the services - tests which many of you helped develop - and my staff and I participate in the planning of Department of Defense research.

I now want to give you an example of the kind of research most of our commanders grasp at -- specifically that which gives the commander a product in the near term. When we marketed the ASVAB to the schools in July 1984, we thought the test would take only three hours to administer but with the inclusion of the instructions the time was increased to about three hours and twenty minutes. Now this does not sound like much of a problem but all the testing periods had been blocked into three hours. The school bells rang and students got up and wandered off. We had many incompleted tests and test security was compromised. Some of the schools quit the program and the recruiters were becoming upset.

A research program was quickly implemented by the Command with Navy trainees to see if shortened instructions would still be effective. It was determined that both an experimental and control group did equally as well and that shortening the instructions would have no impact on test validity. So we were able to shorten the test back to three hours and continue with the high school testing program.

Another example of research with great value to USMEPCOM is the Computerized Adaptive Testing - Armed Services Vocational Aptitude Battery (CAT-ASVAB) program which will be discussed thoroughly this week. CAT is an example of basic research which has matured to the practical level; we will be putting it to use in the near future.

I won't cover the CAT-ASVAB milestones since they will be discussed in depth during the week's sessions. However, the one I am interested in is when we declare victory and actually start CAT-ASVAB testing in the MEPS. In the meantime it is going to be a burden to get CAT-ASVAB evaluated, especially for the recruiters who must bring in applicants and tell them they will have to take 6 hours of tests. We are going to pay the price to stay on schedule. Already we are getting resistance from the recruiter force but CAT-ASVAB will be operationalized.

What I would like to do now is to switch to a discussion of our effort to test for the AIDS virus. Most of us believed up until a few months ago that this was a disease that affected homosexual men, intravenous drug users, and, tragically, hemophiliac patients but did not have much impact on the population in general. But I can tell you now that an estimated 1.2 million Americans are currently carrying the antibody. Occurrence of the disease is estimated to be doubling every year. Young people at enlistment age will be the most affected. In the past members of the Armed Forces have been six times as likely as the general population to contract venereal disease. If this trend holds for AIDS we as researchers and leaders must be very concerned about this disease and its potential for disruption of the Armed Services.

As most of you know the Department of Defense made a decision on 30 August to test all new applicants for the HTLV-III antibody. A positive test result indicates that the individual has come in contact with the AIDS virus some time in the past. It does not indicate that the individual has the virus now nor does it mean that the individual has the disease now. However, recent tests show that a very high percentage of people testing positive for the antibody do carry the virus at that time.

The testing program has already begun. It started on 10 October in the training centers and on 15 October in the MEPS. Based on this test we will deny enlistment to applicants who test positive to the antibody.

In implementing this program I wanted to maintain the current processing flow because it is critical to manning the force. We wanted to minimize the impact on shipping. Actually, as I mentioned, we don't have many straight shippers. Almost everybody we enlist goes into the Delayed Entry Program or back to Reserve or National Guard units. Only about 5% are straight shippers. They can't ship any longer. They must wait at least one day for the test results to come back. We wanted to be very accurate in the preparation of our specimens with all the traffic involved. We had some difficulty in testing initially but hopefully that has been corrected.

Applicants to the Delayed Entry Program are currently being tested at the training centers in order to allow the MEPS to continue their normal processing flow. However, by 1 October the MEPS will take over the responsibility for all AIDS testing except for active duty personnel. As you may know the plan is to begin testing the active duty force in the future.

Finally we wanted to be very careful in notifying the applicants that they were carrying the antibody. We decided to return them to the MEPS and tell them face to face as opposed to sending them a registered letter which we studied as an alternative.

To develop our plan we worked with a number of the MEPS commanders with advice from the service recruiting commands. We could have gone with a decentralized system very easily but the cost would have run to $16 or $18 per test. This high cost and the concern over quality control forced us to a national contract. One lab which would pick up nationwide and overseas and do the lab tests in one location. This would allow Walter Reed and other institutions to perform very specific quality control procedures.

We then had all of our medical Noncomissioned Officers in Charge (NCOICs) come into Chicago so we could train them very carefully. We also brought in all our medical officers - each of the 70 MEPS has one full time chief medical officer. Some of the larger ones have two. These are augmented with 400 fee based physicians. The average age of medical officers is 60. We had four over the age of 80. Just getting them to Chicago was an effort. But they really were enthusiastic about the program. The medical officers and the MEPS commanders have given the program their full support.

USMEPCOM began testing 15 October; the Navy, Air Force and Marine training centers started 1 October; and the Army training centers started 10 October. We had already completed over 17,000 tests by last Friday and we are moving forward.

How many positives are out there? It is emerging data; I don't know yet. I know the studies we did at Fort Benning showed 8 positives per 1000 for 1,400 tested. We don't know what the numbers are going to be in this program but we hope it is going to be less than 8 per 1,000.

The contract went to Damon Corporation Medical Labs in Needham, MD. The cost was $4.41 per test as opposed to the average of $16 to $18 for decentralized testing. All tests will go to Dallas by overnight express and the results electronically relayed to us each day before two o'clock. We test by doing the ELISA Screening Test twice. If the specimen is tested positive twice it is sent to Chicago for the more specific Western Blot Confirmatory Test. And if this test proves positive only then do we call the individual in to inform him or her of the results.

With the ELISA Screening Test we get the negatives back in one day. So if we pick up by noon, by noon the following day the results are back. This allows us to ship all our applicants except those who were ELISA positive. It takes three days to get the results of the Western Blot Confirmatory Tests back to the MEPS.

Issues of confidentiality are obviously a great concern for us. At the time the individual comes in for the physical exam we tell them that they are going to be tested for the HTLV-III antibody. They sign an acknowledgement form which amounts to an informed consent so it is not a complete surprise if they are called back with a positive result.

The release of the information will be a problem. We are handling it through the Office of the Secretary of Defense General Counsel. All requests are being sent there for disposition. Specific requests may go through Central Disease Control in Atlanta. We don't know at this time. We won't be indiscriminately releasing any information concerning individuals.

The notification process entails sending a letter to the individual to return to the MEPS about a medical problem. The recruiter who made the initial contact with the individual and his or her family will go to the home and pick up the applicant and return with the applicant to the MEPS. The individual will be informed of the positive results by the chief medical officer with the commander present. After notification a second blood sample will be drawn. We are doing that to be 100% positive. The individual will then be returned home by the recruiter. We will inform the individual of the results of the second test by registered mail, return receipt requested, or by telephone; probably both. If the second test comes back negative - we don't expect to have any of these - we have to research every aspect of the case to find out why.

A second letter is actually handed to the individual after the interview telling him or her that they have the antibody, but it doesn't mean they have AIDS; that they may or may not contract the disease; and that they should see their private physician to seek further advice about their health.

We also have a fact sheet developed at our request by the doctors at Walter Reed. The fact sheet is very good and is used by own people to overcome the concern they have about the disease. Put yourself in the place of the recruiter who has to pick up an applicant and then come maybe three hundred miles in six hours to the MEPS and then return home with the individual. This is a very difficult task. The fact sheet helps calm the fears of all involved.

We have also developed a workshop program to give the recruiter training in how to handle these applicants. In developing this training we used social workers, psychiatrists, and suicide counselors. They came up with some do's and don'ts to reinforce and give confidence to the recruiters before they have to deal with this situation. We have included in this training a list of questions and answers which address possible problems the recruiter could encounter while bringing in an applicant.

We have about 7,000 recruiters and all of them will be faced with transporting a ELISA positive applicant at one time or another. We are taking this workshop to the recruiting schools where it will become part of their training.

This is the end of my prepared remarks. Thank you for your time and attention.

Rule Space: A Model for Identifying Erroneous Rules of Operation

Maurice Tatsuoka

University of Illinois at Urbana-Champaign

Introduction. It is well known that a student's total score on a test does not tell the whole story--in fact it often tells very little--about the student's achievement level and even less about the kinds of incorrect notions he or she has about the subject matter being tested. To get something close to the whole picture, we need to examine the student's response pattern, which is a vector of 1's and 0's, representing right and wrong, respectively, on the successive items, like [1,0,0,1,1,1,1,0,...,0,1,1].

However, since the vector will have as many elements as there are items on the test, it is clear that we'd be hard put to make much sense out of it unless the test is quite short. What is needed, therefore, is some way to summarize the information contained in a response-pattern vector by means of some kind of numerical index.

Suppose that an achievement test has been given to a sizable group of students, and the facility level of each item (i.e., the percentage of the group that got each item right) has been determined. Suppose, further, that the items have been rearranged from the easiest to the hardest in recording the 1's and 0's (for right and wrong) in the successive response-pattern vectors. What sort of vectors would we expect to find predominantly in the set of vectors for the whole group? From the way the items have been arranged, we should find most vectors to have more 1's toward the left and more 0's toward the right-hand part--representing a pattern of responses in which more of the easier items were passed and more of the harder ones were failed. The "ideal" pattern would be one in which all the 1's precede all the 0's. Such a vector may be called a "Guttman vector," in analogy with what are called Guttman scales in attitude scaling. Of course, Guttman vectors would be found rarely if at all in any set of real response-pattern vectors. Most of the vectors would show some 0's interspersed among the predominant 1's in the left part and some 1's among the 0's in the right-hand part.

In several vectors, we may find a rather random assortment of 1's and 0's throughout; in a few, we may even find a predominance of 0's occupying the left part (corresponding to the easier items) and more 1's in the right-hand (i.e., the harder-items) part. Both these types of response vectors would have to be regarded as anomalous or "aberrant," because the first type shows just about the same proportions of the easier and the harder items being passed, while in the second type more of the harder items are passed than the easier items. Thus an index that shows the extent to which a given response vector approaches a Guttman vector or the opposite extreme in which all the 0's precede all the 1's (a "reverse Guttman vector") could serve as a measure of the "typicality" or "atypicality" of that vector. Such an index was, in fact, developed at an early phase of our project, and it was called the "Norm Conformity Index" (NCI). However, this index has the defect that it defines "typicality" with reference to an "ideal" standard or norm that is virtually unrealized in practice.

Extended Caution Indices. The above circumstance led us (mainly Kikumi Tatsuoka) to seek a measure of typicality/atypicality that is based on a probabilistic model, namely Item Response Theory (IRT). In highly simplified terms, what our model does is first useIRT with the one- or two-parameter logistic function to estimate the student parameter θ and the item parameter a (or parameters a and b) for each student and item in a data matrix. Then, for each student, the IRT-based probability for passing each item is computed, and a matrix $(P_{ij}(\hat{\theta}_i))$ is constructed. This could be called an IRT-based theoretical data matrix, with the 0-1 entries replaced by $P_{ij}(\hat{\theta}_i)$'s. Finally, an index analogous to the NCI is computed for each student, representing the extent to which his/her observed response vector deviates from the corresponding vector of probabilities based on IRT, relative to the extent to which the group's average response vector deviates from that IRT-probability vector.

Actually, several such indices, called "Extended Caution Indices" (ECI) can be defined, depending on just what IRT-based quantities are used to replace the 1's and 0's of the Guttman vectors and their means over students. Five such ECI's were defined and discussed by K. Tatsuoka and Linn (1983).

It turned out that the one that was called ECI4, when standardized, served the best as an index for detecting anomalous response patterns in a group. (The standardization is for the purpose of making the values comparable over different θ's.) This was denoted by ζ and has been used exclusively in all our subsequent work. It also has the convenient property that, prior to being standardized, it is interpretable as a linear mapping function:

$$\begin{aligned} f(\underset{\sim}{x}) &= (\underset{\sim}{P}(\theta) - \underset{\sim}{x})'(\underset{\sim}{P}(\theta) - \underset{\sim}{T}(\theta)) \\ &= \Sigma_{j=1}^{n} (P_j(\theta) - x_j)(P_j(\theta) - T(\theta)) \\ &= -(\underset{\sim}{P}(\theta) - \underset{\sim}{T}(\theta))'\underset{\sim}{x} + (\underset{\sim}{P}(\theta) - \underset{\sim}{T}(\theta))'\underset{\sim}{P}(\theta) , \end{aligned}$$

which associates with each response pattern $\underset{\sim}{x}$ a real number $f(\underset{\sim}{x})$. Here

$$\underset{\sim}{T}(\theta)' = [T(\theta), T(\theta), \ldots, T(\theta)]'$$

is a vector whose n elements are all equal to

$$T(\theta) = (1/n)\Sigma_{j=1}^{n} P_j(\theta) ,$$

which is the mean, over items, of the IRT-based response probabilities for a fixed θ.

The expected value and variance of $f(\underset{\sim}{x})$ for fixed θ, denoted $f_\theta(\underset{\sim}{x})$, was shown by K. Tatsuoka (1985) to be

$$E(f_\theta(\underset{\sim}{x})|\theta) = 0$$

and

$$Var(f_\theta(\underset{\sim}{x})|\theta) = \Sigma_{j=1}^{n} P_j(\theta)Q_j(\theta)[P_j(\theta) - T(\theta)]^2$$

respectively.

Rule Space. With the standardized mapping function

$$\zeta = f_\theta(\underset{\sim}{x})/[Var(f_\theta(\underset{\sim}{x})|\theta)]^{1/2}$$

defined above, we can now map each student's response pattern $\underset{\sim}{x}$ into a point (θ, ζ) in a two-dimensional space with abscissa $\hat{\theta}$ and ordinate ζ. It was shown that the maximum likelihood estimate of θ, MLE $\hat{\theta}$, and ζ are uncorrelated. This space is called "Rule Space," because--if a student consistently uses some specific rule of operation (or algorithm) in solving all the items on a test--each rule R will yield a unique response pattern $\underset{\sim}{x}_R$ (or just $\underset{\sim}{R}$ for short). Thus the correct rule will yield the response pattern [1,1,1,...,1], and each incorrect rule will yield some specific permutation of 1's and 0's (including all 0's, of course).

Consequently, the rules that yield response patterns leading to the same MLE $\hat{\theta}$ will all be mapped into points that lie on a straight line perpendicular to the $\hat{\theta}$ axis. (Such response patterns will have the same number, Σx_j, of 1's in the case of the one-parameter logistic model, and the value of the sufficient statistic, $\Sigma a_j x_j$, will be equal in the case of the two-parameter model.)

Thus, what ζ does is to pull apart students who have the same total score (or the same sufficient statistic for θ) on a test. Students whose response patterns are typical of their group (and hence whose ζ values are small) will be represented by points close to the $\hat{\theta}$ axis, while those whose response patterns are unusual for the group (large ζ values) will be mapped into points that are above the $\hat{\theta}$ axis--the more unusual their pattern, the farther up their points will be.

But now comes the rub. The neat mapping indicated above was predicated on the assumption that each student uses some specific rule consistently throughout the test. But of course this will not be true in practice. There are bound to be some random departures from the use of a constant rule in a few of the items. Then the student's response pattern will not be mapped into a point that corresponds to consistent use of a single rule. Rather, the point will fall in the vicinity of the point for the rule that has been used for solving most of the items. These may be called "perturbations" from the "pure" rule point. Or, if we want to be more specific about the extent of perturbation--i.e., the number of items on which there was a departure from the modal rule--we may call the points "one-slip points," "two-slip points," and so on. Tatsuoka and Tatsuoka (submitted to Psychometrika) have shown that the probability of there being no more than some number s ($<n$) of slips in a test with n items is given by a compound binomial distribution. Hence, asymptotically, the perturbed points will be distributed normally around the pure rule point from which they are perturbations. Moreover, under reasonable assumptions, the variance of this distribution will--at least for rule points that are not too far from the $\hat{\theta}$ axis--be equal to the variance that was previously displayed for $f_\theta(\underset{\sim}{x})$, namely, $\Sigma_{j=1}^{n}P_jQ_j[P_j(\theta) - T(\theta)]^2$. This, along with the facts that $\hat{\theta}$ is normally distributed with mean θ and variance $1/I(\theta)$ and that ζ and $\hat{\theta}$ are uncorrelated, allows us to conclude that the perturbed points will follow a bivariate normal distribution with a known centroid and a known, diagonal covariance matrix. Consequently, the points corresponding to response patterns with no more than a certain number of slips will lie inside ellipses whose minor and major axes are parallel to the reference axes of rule space. These ellipses will constitute various iso-density ellipses of the bivariate normal distributions around the pure rule point. The upshot is that, if all the rule points and the observed response points that are perturbations from them were to be plotted in rule space, the result would be something like this: There would be several swarms of points, each of which would be most densely concentrated around one of the pure rule points, and would become sparser and sparser as we go farther from the center (i.e., the rule point) in any direction.

We can now explicate just what is meant by the title of this paper, "a model for identifying erroneous rules of operation," as follows: It is a technique whereby, given a student's point $(\hat{\theta}_i, \zeta_i)$ in rule space, we are enabled to decide to which one of the ellipses this point most likely belongs. This is because we have, through this model, translated our original problem into a typical problem of statistical decision theory: the problem of classifying a given point into one of several bivariate normal populations--or, more generally (as we shall soon see) into one of several 2m-variate normal populations.

One of the ways for solving such a problem is to invoke the "minimum-D^2 rule," where D^2 is the (squared) Mahalanobis generalized distance. Without loss of generality, we may assume that we have eliminated all but two rule points $\underset{\sim}{R}_1$ and $\underset{\sim}{R}_2$ as candidates for the rule point from which the given student's response point is a perturbation. If we denote the common covariance matrix of the two bivariate normal distributions with $\underset{\sim}{R}_1$ and $\underset{\sim}{R}_2$ as their centers by $\underset{\sim}{\Sigma}$, then the D^2 of the given response point from the centroid $\underset{\sim}{R}_k$ of the Rule k ellipse is

$$D^2_{\underset{\sim}{x}k} = [(\hat{\theta}_x, \zeta_x) - \underset{\sim}{R}_k]' \underset{\sim}{\Sigma}^{-1} [[(\hat{\theta}_{\underset{\sim}{x}}, \zeta_{\underset{\sim}{x}}) - \underset{\sim}{R}_k] \qquad (k = 1,2)$$

Upon computing the two D^2s, we would classify $\underset{\sim}{x}$ as a perturbation from $\underset{\sim}{R}_1$ if $D^2_{\underset{\sim}{x}1} < D^2_{\underset{\sim}{x}2}$ and otherwise as a perturbation from $\underset{\sim}{R}_2$.

Once we have decided what particular rule it is that the student's response pattern is most likely to be a perturbation from (i.e., what rule he/she most likely used almost consistently except for a few "slips"), it is a short step to diagnosing the particular misconception that was most likely held by the student for him/her to have adopted that rule in the first place. This is because the rules were originally inferred from a careful error analysis of actual test papers by experienced subject-matter specialists of the material covered by the test.

Application to a Test on Subtraction of Signed Numbers. Suppose that a test consisting of the ten items listed in the first column of Table 1 (without the answers, of course) was given to a group of seventh-graders, and that four of the students each used one of the four erroneous rules described at the bottom of the table. The four pairs of columns—each pair headed by a rule number—show the answers obtained by using the four rules, respectively, and the binary score (1 or 0) for each item.

Table 1. The Binary Response Vectors for a Set of Ten Items Responded to by Four Incorrect Rules.

	Rule 1		Rule 2		Rule 3		Rule 4	
Items	Response	x_j*	Response	x_j	Response	x_j	Response	x_j
-3 - (-7) = +4	-4	0	+4	1	+4	1	+4	1
-2 - 8 = -10	+6	0	+6	0	+6	0	-10	1
5 - (-12) = +17	-7	0	+7	0	+17	1	-7	0
-11 - +8 = -19	-3	0	+3	0	-19	1	-19	1
9 - 4 = +5	+5	1	+5	1	+5	1	+5	1
-15 - (-9) = -6	-6	1	+6	0	-6	1	-6	1
-13 - 5 = -18	-8	0	+8	0	-8	0	-18	1
8 - (-6) = +14	+2	0	+2	0	+14	1	+2	0
-5 - +11 = -16	+6	0	+6	0	-16	1	-16	1
1 - 10 = -9	+9	0	+9	0	-9	1	-9	1

Rule 1: The student subtracts the smaller absolute values from the larger absolute value and takes the sign of the number with the larger absolute value in his/her answer.

Rule 2: The two numbers are always subtracted as seen in Rule 1 but the + sign is always taken in the answers

Rule 3: The student converts -2 - 8 and -13 - 5 into -2 + 8 and -13 + 5, respectively, but the other eight items items are converted to addition correctly. Then the right addition rule is used to answer them.

Rule 4: The student has a strange idea about the parentheses. Converts operation sign, -, to +, first. Then he/she follows the rule: if the signs of the two numbers are minus, then change the sign of the second number to a +; if the signs of the two numbers are not alike, then the sign of the second number becomes a minus.

* x_j is the score for the jth item in the binary response vector $\underset{\sim}{x}$.

We see from Table 1 that Rules 1 and 2 both yield correct answers in two of the ten problems, but not the same two. Similarly, if a student consistently uses either Rule 3 or 4, he/she would get eight problems right (again, not the same eight). Thus, consistent use of either of the first two rules or either of the second two would—assuming the discrimination parameters a_j to be equal for all of the items—respectively yield the same estimated θ values. However, examining the descriptions of the four rules at the bottom of the table, we see Rule 2 represents greater ignorance than Rule 1, and Rule 3 might be the result of careless slips while Rule 4 is quite a "weird" one. Therefore, consistent users of Rule 1 and those of Rule 2 should not be treated alike. This is probably true for the sheer purpose of guaging their achievement in signed-number subtraction, and is certainly true for the purpose of diagnosing their misconceptions and giving them remedial instruction; similar remarks hold for users of Rules 3 and 4. This is where the ECI ζ comes into play. Students who get the same θ by virtue of consistently using Rule 1 and Rule 2, respectively, would be pulled apart by their different values--those using Rule 2 being plotted higher up in rule space (assuming the group to which they belong is reasonably competent). Likewise, those using Rule 4 would come higher up than Rule-3 users, because Rule 4 is a much more unusual rule. Readers who are interested in further details on how ζ works are referred to K. Tatsuoka (1983, 1985).

Sometimes, even the two-dimensional rule space will not suffice to pull the different rule points apart enough. In fact, tests of signed-number subtraction are a case in point. In such cases it may be possible to achieve better resolution, as it were, by introducing subscores for each item; e.g., a subscore for the absolute value of the answer being right or wrong, and another subscore for the sign of the answer. Then there will be one θ for each "component" and likewise one ζ for each, thus generating a rule space of four dimensions, or more generally 2m dimensions, when there are m components.

Application to Adaptive Testing. The ideas of rule space can also be used in speeding up the "convergence" in computerized adaptive tests--i.e., for getting a stable estimate of a student's θ value by administering, on the average, a much smaller number of items than in the traditional approach in which only the information value $I(\theta)$ is considered in choosing the next item. (See Lord, 1980.) In highly simplified terms, the idea is to have at hand--i.e., stored in the computer--a curve of $P_j(\theta) - T(\theta)$ plotted against $T(\theta)$ for each item in the pool. These curves will be called "Item Disparity Curves" (IDC) for subsequent reference. Then, in selecting the next item to be given to a student, we take into consideration--in addition to the $I(\theta)$--the IDC of each item remaining in the pool, and choose an item whose IDC is as far removed as possible from the IDC of the item just taken, at the current estimate $\hat{\theta}_{(k)}$ of the student's θ value.

Space does not permit indicating the proof that the method outlined above will indeed speed up the convergence to the "correct" θ value, but it is soundly based in functional analysis--which is a branch of mathematics that has many important applications in statistics and psychometrics but which, to quote Ramsay (1980), "...we in North America are handicapped by the fact that a course in [it] is seldom a part of the preparation of an applied statistician" (1982, p. 394). A proof has been made by K. Tatsuoka, and it will soon be available in a CERL Research Report of the ONR-sponsored Computerized Adaptive Testing and Measurement (CATM) Project.

References

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Ramsay, J. O. (1982). When the data are functions. Psychometrika, 47, 379-396.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20, 345-354.

Tatsuoka, K. K. (1984). Caution idices based on item response theory. Psychometrika, 49, 95-110.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. Journal of Educational Statistics, 10, 55-73.

Tatsuoka, K. K., and Linn, R. L. (1983). Indices for detecting unusual response patterns: Links between two general approaches and potential applications. Applied Psychological Measurement, 7, 81-96.

Tatsuoka, K. K., and Tatsuoka, M. M. Bug distribution and pattern classification. (Research Report 85-3-ONR). Urbana: University of Illinois, Computer-Based Education Research Laboratory. (Submitted to Psychometrika)

Appendix

Proof that $f(\underset{\sim}{x})$ Increases with Unusualness of Response Pattern $\underset{\sim}{x}$. It was stated without proof, on the third page of this paper, that the standardized version, ζ , of $f(\underset{\sim}{x})$ becomes larger as the response pattern $\underset{\sim}{x}$ becomes more "unusual" for the group of which the responder is a member. We now prove this. Expanding the second expression for $f(\underset{\sim}{x})$ given on page two, we get

$$f(\underset{\sim}{x}) = \Sigma_{j=1}^{n}P_j(\theta)[P_j(\theta) - T(\theta)] - \Sigma_{j=1}^{n}x_j[P_j(\theta) - T(\theta)] .$$

We may assume, without loss of generality, that the items have been numbered in ascending order of difficulty so that, for each θ, we have

$$P_1(\theta) > P_2(\theta) > \ldots P_n(\theta) .$$

(The order may differ for different values of θ.) Then, since $T(\theta)$ is the mean of the n $P_j(\theta)$ values, there must be some k such that

$$P_j(\theta) - T(\theta) > 0 \qquad \text{for all} \quad j \leqslant k$$

and

$$P_j(\theta) - T(\theta) < 0 \qquad \text{for all} \quad j > k .$$

This, together with the fact that the first sum in the above expression for $f(\underset{\sim}{x})$ is a constant for any fixed θ, implies that for any response pattern that has a small number of $x_j = 1$ for $j \leqslant k$ (i.e., the easier items) and a large number of $x_j = 1$ for $j > k$ (the harder items), the quantity $f(\underset{\sim}{x})$ will have a larger numerical value than for a response pattern for which the opposite is true. But the first-mentioned type of response pattern (with a small number of corrects among the easier items and a large number among the harder items) is clearly an "unusual" or atypical one. This shows that, for a fixed θ , the larger $f(\underset{\sim}{x})$ is, the more unusual is the response pattern $\underset{\sim}{x}$--for the group in which the items were calibrated. The standardization to get ζ makes this property hold across different values of θ.

Acknowledgement

This research was sponsored by the Personnel and Training Research Program, Psychological Sciences Division, Office of Naval Research, under Contract No. N00014-82-K-0604. Contract Authority Identification Number NR 150-495.

Implementation of a Computer System to Support Diagnostic Testing

C. David Vale
Assessment Systems Corporation
St. Paul, Minnesota

The MicroCAT^{tm} Testing System, developed by Assessment Systems Corporation, is a complete system of computer programs to support computerized adaptive testing (CAT). MicroCAT runs on an IBM Personal Computer and includes facilities for performing all of the functions necessary to implement adaptive tests. These functions include authoring test items, calibrating items according to item response theory (IRT) models, authoring tests using any of a variety of CAT strategies, and administering tests. The details of the MicroCAT system have been described elsewhere (Assessment Systems Corporation, 1984; Val^, in preparation, a, b) and will not be detailed here.

MicroCAT served as the basis for the testing system implemented in the Basic Electricity and Electronics (BE&E) School of the Naval Training Center (NTC) in San Diego. The purpose of this implementation was twofold: to provide a delivery system for the diagnostic testing techniques developed by the University of Illinois, and to provide an environment in which to evaluate the MicroCAT system, which was not then in commercial release.

There were two general requirements that the system for NTC had to meet. First, it had to administer the current NTC tests in a form that was comparable to their original mode of administration. (This was to allow the collection of data on the current forms in an operational testing environment.) Second, the system had to be capable of administering tests according to the prescribed diagnostic strategies.

In the standard mode of administration at NTC, a student is assigned a microfiche card which contains a test. He or she then reports to a carrel with a microfiche reader and responds to the test questions on an optically scannable answer sheet. After completing the test, the student puts the answer sheet into an optical scanner that is connected to a computer terminal which, in turn, is connected to a mainframe computer in Memphis, Tennessee. A computer-managed instruction program running on that computer scores the results, updates the student's records in the database, and reports the score to the student. It also tells the student which test to take next.

Our goal in designing the MicroCAT interface to this mode of testing was to be able to install the computerized testing system in such a manner that it would be perceived by examinees as comparable to the microfiche version of the test. Its operation also had to be completely transparent to MIISA, the computer-managed instruction program, so that the results of the computerized tests could be used operationally in the computerized instructional management system. MIISA is a very complex program that manages all of the instructional assignments and record keeping at NTC. It was written several years ago, runs on a mainframe computer, and is very resistant to change of any kind. Any approach to implementation that required reprogramming of MIISA was not viable.

Although the MicroCAT system was originally designed for stand-alone testing, a general networking capability was added to allow a proctor to assign and monitor tests at several stations from a single proctor's station. This standard MicroCAT networking system was still unable to communicate with MIISA, but the proctoring station provided a good starting point for the connection. To achieve the transparency desired, the proctoring program was adapted so that it could interact with MIISA as well as with a proctor. It was changed to make the proctoring terminal emulate one of the terminals that normally communicated with the MIISA mainframe. To accomplish this, we made the proctoring station's serial port look like a GE Terminet printing terminal connected to an optical scanner. As far as MIISA was concerned, it was communicating with an optical scanner connected to a terminal. When an examinee reported to the testing room, the proctor assigned him or her to a testing station, where the examinee entered his or her social security number. The testing station communicated this to the proctoring station, which queried MIISA for the appropriate test to administer and assigned it to the examinee's station. After the examinee completed the test, the proctoring station passed the responses to MIISA, obtained the scores, and printed them out for the examinee on the system printer. From MIISA's perspective, the computerized system was capable of performing all of the functions usually done on paper.

Figure 1 diagrams the testing system as implemented at NTC. It contained eighteen terminals. Fifteen of these were testing stations, two were network servers (which contained all of the tests and response data), and one was a proctoring station. One network server would have been sufficient; the second one was for backup in case the first one failed. One of the testing stations also had the hardware necessary to convert it to a proctoring station if the first one failed.

There was a fear among the Navy Chiefs in charge of testing that it might be difficult to convince the examinees that the computerized mode was equivalent to the paper-and-pencil mode. Their concern was that if tests were administered in two forms, superstitions would develop among the students regarding which mode was easier. Obviously, this problem would be exacerbated if there was any substance to the claim.

Specifically, three factors were initially assumed to be related to the examinees' acceptance of the computerized mode as an equivalent mode of testing: (1) the system had to respond quickly with the next item after the examinee answered the previous one, (2) it could not lose an examinee's work if any portion of the system failed, and (3) it had to support standard test-taking strategies such as skipping difficult items and coming back to them later.

MicroCAT was fast enough to satisfy the first requirement, but originally it did not support the other two features. Therefore, a recovery feature was added to save the examinee's responses on a diskette as soon as they were made. If a testing station failed, the examinee could simply remove the diskette in his or her station, put it in another station, enter his or her social security number, and continue the test with no loss of data.

The ability to skip items and later return to them was also added to the MicroCAT system. At the end of the test, the examinee is asked if he or she would like to review all of the items, some specific ones, or just those he or she skipped. A

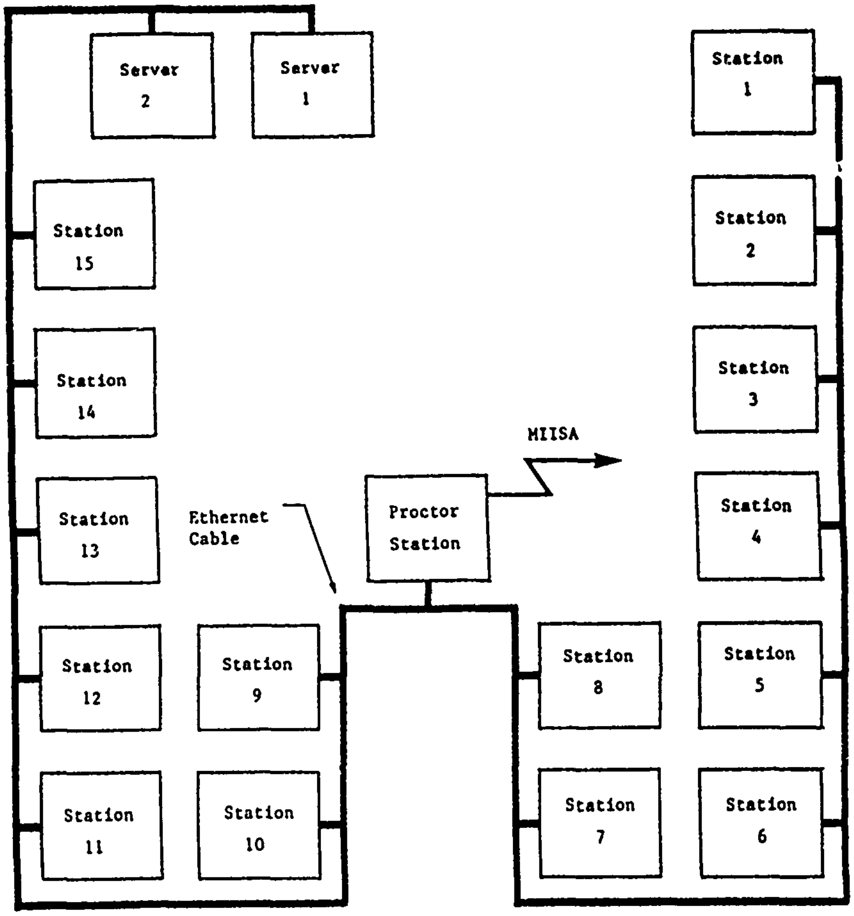
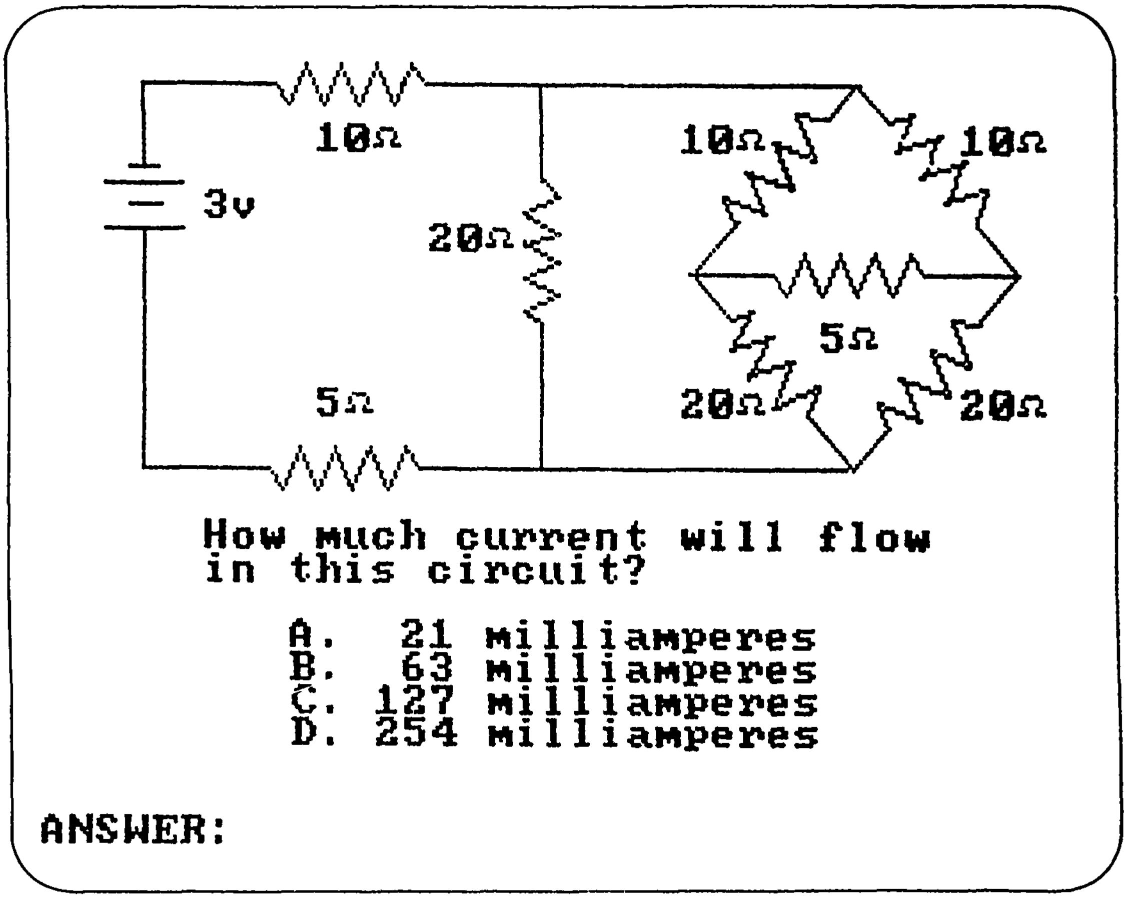


Figure 1. Structure of the NTC Implementation

review is granted as desired, and the specified items are re-presented to the examinee with his or her previous response shown. The examinee can then change the response or leave it as it was.

After these features were added to the MicroCAT system, it was installed for a small-scale evaluation. In general, the initial system ran without any serious errors. Several small problems did require attention, however. First, as the system was originally set up, it was possible for students to take two tests simultaneously. Since they received no feedback regarding the correctness of responses, there was really no advantage to be gained by doing this. Nevertheless, the problem was corrected by modifying the proctoring program to keep track of who was on the system and to allow examinees to take only one test at a time. The second potential problem was that hitting some combinations of keys could abort the testing system. This was solved by disabling all of these combinations except one that required the examinee to simultaneously push three keys. It is highly unlikely that anyone would accidentally hit all three keys at once; anyone who did hit them would probably be deliberately trying to abort the system. It was not possible or worthwhile to subvert such an attempt, because a determined examinee could always reset the station simply by unplugging it.

The system was implemented operationally in August of 1984. The test items presented were similar in format to the one shown in Figure 2 (which is not an active test item). Most examinees had no trouble with the new system. The major surprise on the first day of testing was that, on the average, examinees took ten minutes to respond to an item. This did not begin to tax the capacity of the network, which was designed to handle a response from each examinee every ten seconds.



How much current will flow in this circuit?

A. 21 milliamperes
B. 63 milliamperes
C. 127 milliamperes
D. 254 milliamperes

ANSWER:

Figure 2. Sample Electronics Item

Since the data collected by computer administration were to be analyzed by the University of Illinois, a data transfer scheme was needed. The MIISA link is a real-time link in that testing waits for communication. Transferring the data to the University of Illinois, on the other hand, had to be done only when the data were needed or when the disks on the NTC network were full. A system was developed whereby the test proctor periodically dumped the data from the system disks to two sets of diskettes, one for the University of Illinois and one for backup. After dumping the data, the proctor's instructions were to mail one set to the University of Illinois and to keep the backup set until receipt was confirmed. The data on the system disk were erased after the diskettes were made. Except for the difficulty of getting the proctor to make the data diskettes on a regular basis, this scheme worked well.

Testing has not been interrupted because of any system problems. It was interrupted for several weeks, however, by the implementation of new versions of the tests. The frequent changes in tests, which had not been anticipated when the system was installed, required frequent communication with the University of Illinois. It had been intended that the University of Illinois would do the test development and then either manually install the tests in the San Diego system or mail complete test files with installation programs to be run by the proctor. However, as the test changes became more frequent, it became apparent that it would be more efficient for NTC personnel to make the changes themselves and install the tests.

Test development in the MicroCAT system is a three-stage process. First, the items are authored using the system's Graphics Item Banker. Then the test is specified using an authoring language. Finally, the authoring language is compiled, a process that reformats the items and processes the instructions in a manner that allows items to be presented rapidly. Implementing a test in the NTC system required the further step of copying the compiled test onto the appropriate disk volume.

NTC test administration personnel mastered the process with relative ease. However, several problems arose that had bothered us in the early part of this effort but which we had forgotten until the NTC personnel began to use the system. One such problem was that if diskettes were swapped while the item banker was running, a bank would be destroyed. We originally circumvented this problem by not swapping diskettes, but this solution was obviously not optimal. We therefore wrote a utility program that could recover a bank destroyed in this manner. The other problem that we re-discovered was that, using the Ethernet network from 3Com, two people sharing a disk volume can, under certain circumstances, destroy each other's work. For example, NTC personnel recently destroyed an item bank by writing portions of a memo over it. Fortunately, the new program was able to restore most of what was lost.

In general, the implementation at NTC has been successful. The MicroCAT system has flawlessly performed most of the tasks required of it. More than 2,400 items have been banked for this application, approximately 50 different tests have been implemented, and over 1,500 tests have been administered. Informal evidence from the BE&E School suggests that the system is fast enough for administering all of the tests and that the tests are perceived as psychologically parallel to the microfiche form (although we have heard that the computer display is easier to read than the microfiche). The system has not yet been used for diagnostic testing, but custom interfaces (portions of the program that allow programmers to augment the MicroCAT system) have been provided to allow diagnostic testing routines to be implemented within the MicroCAT system.

The MicroCAT Testing System is now a commercial product and the contract that supported its development and implementation at NTC is near an end. Its use at NTC will continue and the University of Illinois will implement the diagnostic strategies in the near future. It has been a successful implementation that has demonstrated the relative ease with which the transition can be made from paper-and-pencil testing methods to computerized testing, even when the new system must be integrated into a complex instructional management system that is already in place.

References

Assessment Systems Corporation. (1984). *User's manual for the MicroCAT testing system.* St. Paul: Author.

Vale, C. D. (in preparation, a). *Implementation of a microcomputer-based testing system in a Navy training environment* (Research Report ONR-85-XX). St. Paul: Assessment Systems Corporation.

Vale, C. D. (in preparation, b). *MCATL: A language for authoring computerized adaptive tests* (Research Report ONR-85-XX). St. Paul: Assessment Systems Corporation.

Acknowledgments

The preparation of this paper and part of the development of the MicroCAT system was supported by the Personnel Research and Training Programs, Psychological Sciences Division, Office of Naval Research, under contract N00014-83-C-0634 (NR 507-002).

MicroCAT is a trademark of Assessment Systems Corporation.

APPLICATION OF RULE SPACE IN A NAVY TESTING ENVIRONMENT

by John M. Eddins
University of Illinois at Urbana-Champaign

Introduction

The MicroCAT testing system was installed in the BE&E School at the San Diego Naval Training Center in order to extend to an actual training environment the theoretical work in diagnostic adaptive testing described by Dr. Tatsuoka, and reported in detail in Tatsuoka (1985, in press), Tatsuoka & Tatsuoka (1985) and Tatsuoka, Tatsuoka & Baillie (1985). I will present first a summary of our overall plans for the project, next our progress to date, including some of the problems we have encountered, and finally what we see as the next steps.

Summary of Project Plans

The principal goal of the project is to develop one or more diagnostic adaptive tests for the BE&E curriculum using the rule space model, and to evaluate the effectiveness of these tests with data collected from pilot groups of Navy trainees. The overall plan can be outlined in five phases.

1. Collect and analyze data from current Navy tests.
2. Link rule space procedures to the MicroCAT system.
3. Add experimental items to current tests.
4. Create experimental tests.
5. Give experimental tests to pilot groups.
6. Report the results.

Substantial progress has been made on the first two phases, and we are currently working on the third.

Progress to Date

1. Data collection and analysis

The MicroCAT system was installed and tested during the summer of 1984, but ongoing revisions of the instructional materials and tests delayed full implementation until spring, 1985.

BE&E trainees at the San Diego base began testing with the system in late March, 1985, and data was collected for Modules 1, 2, 4, 5, 6 and 7 during April through August, 1985. As it turned out, one more revision of the tests was necessary. The revised tests are now on line and the computer testing lab is back in operation.

As each BE&E trainee takes a test, the MicroCAT system stores a record which includes, for each item, the item number, answer key, student's response, and time to respond. These data are added to a file which eventually is read to a diskette and mailed to us at the University of Illinois. Our plan is to gather enough data to analyze the items statistically and estimate item parameters, on the assumption that some existing items can serve as a starting point for developing items applicable to rule space. This requires a minimum of 200 to 300 subjects, with a significant percentage giving wrong answers. Between April and August we collected data for about 1500 subjects; however, these were divided among five versions each of six tests, leaving only around fifty for each test version. The different test versions for Mod 4 proved to be identical except for the ordering of the answer foils, so we re-ordered the data appropriately and merged it into a composite set, giving us 235 subjects. An analysis of variance across the different forms confirmed their equivalence.

A summary of responses for the items in this dataset is shown in Table 1. A high percentage of the responses on most items are correct, and several items show essentially no discrimination. These data are based on the first pass through the test, and they include partial tests taken for remediation, hence the substantial number of skipped items.

Table 1.

BE&E Test, Mod. 4 -- Summary of Responses.

All test forms equivalent to form 1.

Item#	Key	Rsp1	Rsp2	Rsp3	Rsp4	Skip
1	1	196	8	1	1	29
2	4	9	25	0	165	36
3	3	2	40	125	24	44
4	1	169	7	19	6	34
5	1	207	0	1	0	27
6	3	1	12	189	0	33
7	3	0	0	202	1	32
8	1	196	2	2	0	35
9	4	1	0	2	199	33
10	3	9	2	188	0	36
11	1	204	0	9	0	22
12	2	16	178	6	1	34
13	3	19	41	109	27	39
14	3	0	0	212	1	22
15	1	152	24	11	20	28
16	4	36	14	0	154	31
17	4	10	18	45	101	61
18	4	2	13	2	168	50
19	3	0	6	139	8	82
20	3	3	14	168	2	48
21	3	8	1	177	0	49
22	2	2	197	1	0	35
23	3	2	2	187	1	43
24	2	17	120	0	12	86
25	4	19	5	1	166	44
26	2	4	197	15	*	19
27	1	151	21	42	*	21
28	3	2	26	186	*	21
29	1	172	16	20	*	27
30	2	8	202	6	*	19
31	1	205	1	6	*	23
32	1	152	56	5	*	22
33	1	70	120	23	*	22
34	3	24	14	175	*	22
35	2	17	187	12	*	19

*Only three choices for these items.

Our attempts to estimate item parameters were frustrated at first because of the limited number of both items and subjects, and because of the large number of skipped items resulting from partial tests taken as remedials. We eliminated the partial tests, leaving 193 subjects, selected 22 of the most promising items and succeeded in estimating the A and B parameters for these items. We used a computer program created by Yamamoto, Baillie and Tatsuoka which implements an EM algorithm developed by Bock and Aitkin (1981). The EM method has the advantage of being able to estimate item parameters with relatively few items. Results are shown in Table 2.

Table 2.

BESE Test, Mod. 4 -- Item Parameters for Selected Items.

	A	B
Item#	(discrimination)	(difficulty)
10	.92832	-.49033
11	.80779	-.76712
12	.48945	-.84937
13	.80148	1.00827
15	.29399	-.72174
16	.55109	.09081
18	2.42278	.65190
19	2.33766	.91749
20	2.35894	.63324
21	3.77397	.64296
23	2.87624	.39554
25	2.93106	.71003
26	.72511	-.66316
27	.86399	.52368
28	1.07095	.07809
29	.77854	.10053
30	.38718	-2.75522
31	.98740	-.53114
32	.38085	-.18672
33	1.16947	1.68272
34	.88895	.15271
35	.62432	-.52315

2. Linking of Rule Space Procedures with MicroCAT

To set up a test to be administered on the MicroCAT system, a bank of test items is provided, together with a file which lists the items to be administered and specifies the logical basis on which each successive item is to be chosen. The choice can range from a simple top down sequence of all items on the list to a decision computed by a program external to the MicroCAT system and passed back to it. In our case, administering a diagnostic adaptive test using rule space is in the latter category.

At the beginning of a test, item parameters stored in the item bank are read into an array in memory (Figure 1). As each test item is administered the student's response is stored, and program control is passed to the rule space programs along with student response information.

The student's response is scored as right or wrong (1 or 0), the weighted distances to all remaining items are calculated (Tatsuoka & Tatsuoka, 1985), and the next item is selected for optimal speed of convergence. If sufficient convergence has been achieved, then the error probabilities are estimated between the current point and the centroids of the ellipses stored in the bug information bank. If the minimum value of these probabilities satisfies a specified criterion, next item is set for stop; if not, next item is not changed. Next item and program control are then returned to the MicroCAT test driver. The computer programs for the rule space procedures were developed by Robert Baillie. Details of the theoretical basis for these procedures will be described in a future report.

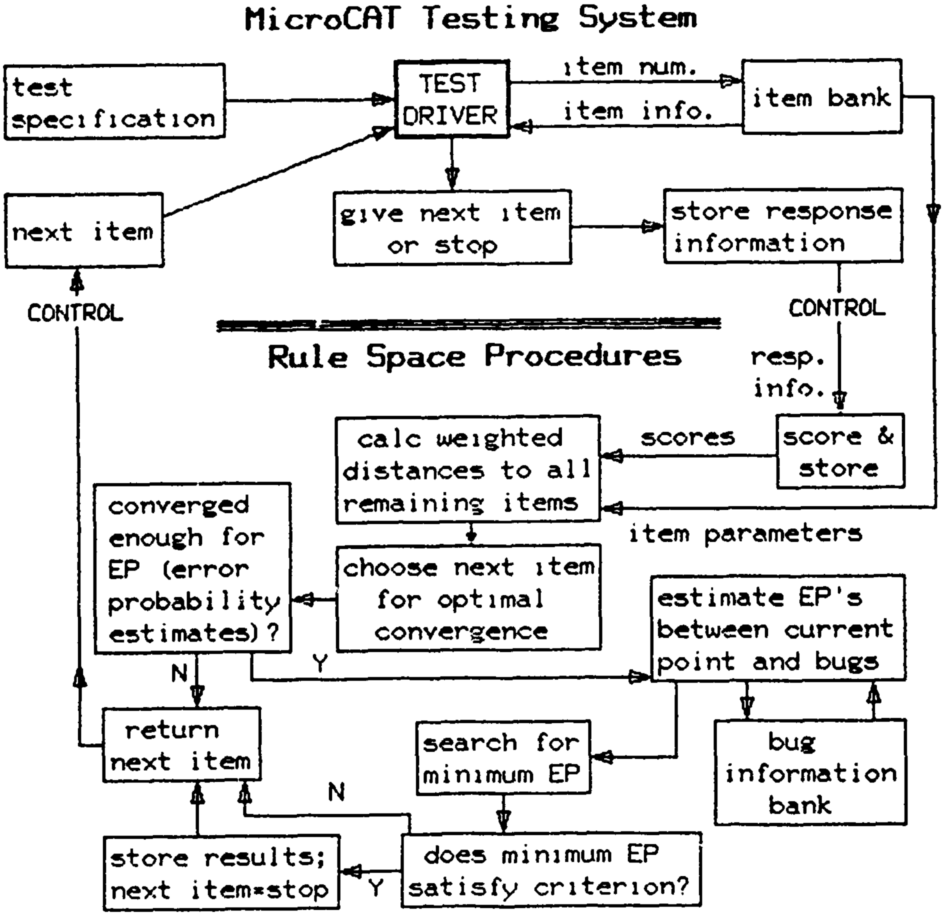

MicroCAT Testing System
test specification
TEST DRIVER
item num.
item info.
item bank
next item
give next item or stop
store response information
CONTROL
CONTROL
Rule Space Procedures
resp. info.
calc weighted distances to all remaining items
scores
score & store
item parameters
converged enough for EP (error probability estimates)?
choose next item for optimal convergence
estimate EP's between current point and bugs
N
Y
return next item
search for minimum EP
bug information bank
N
store results; next item=stop
Y
does minimum EP satisfy criterion?

Figure 1. Linking of Rule Space procedures with MicroCAT.

Future Plans

The principle hurdle remaining before we can actually create and install experimental test items is a detailed analysis and verification of specific learning tasks to be addressed, together with hypotheses as to causes and results of various types of errors. Designing a diagnostic test for use with the rule space procedures requires that the test items be constructed with specific characteristics, and these item characteristics must be analyzed and verified with extensive data collection and computer analysis. Parallel versions of the items also are needed.

At present we are examining in detail the concepts, operations and errors represented by the Navy's BE&E tests. Three items selected from the Mod 4 test will illustrate something of the nature and complexity of the task. In the first item (A), a simple series circuit containing five resistors is shown and the student is asked the effect on voltage drops at an open resistor -- increase (1), decrease (2) or no change (3). The second item (B) shows the same circuit and answer choices, but asks the effect on current at a shorted resistor. The third item (C) shows a series circuit with values given for applied voltage and two of the resistors, one of which is open. The student is asked what a voltmeter across the open will indicate.

For item A, 56% of the students chose (2) -- decrease in voltage drop at an open; for item B, 26% chose (2) -- decrease in current at a short; for item C, 18% chose (1) -- zero volts read across an open. Out of 52 students choosing (2) for item B, 34 also chose (2) for item A; out of 36 choosing (1) for item C, 27 also chose (2) for item A; but out of the total group only 5 chose all three. Apparently, items B(2) and C(1) represent different errors, while item A(2) subsumes both errors. This is also confirmed intuitively. Response (2) to item B probably results from confusing shorts with opens (the <u>result</u> of a short often is an open). Response (1) to item C probably represents confusion about the nature of voltage drop (nothing "happens" across a break, so voltage drop = 0). These are different misconceptions, yet response (2) to item A can result from either one.

The task before us is to pursue this type of inquiry until we have isolated and charted a sufficient number of such relationships among items to enable us to describe several common errors and specify test items accordingly. We also are investigating ways to utilize the computer to speed up and simplify this otherwise tedious and complex task. What we are looking at here is the relationship among just one foil for each of three items. To systematically consider such relationships among all foils for all items obviously is a task for a computer. The critical human task is defining the relevant concepts and errors.

With a multiple choice test it is important that each foil be designed to provide specific information, so that the choice of the next item will be based on more than simply a right-wrong score. Our first experimental test items will likely be items taken from the current tests with some changes in the foils or the circuit values. We have the facility to insert these items into the existing tests without disturbing the regular scoring. We could, for instance, insert five extra items at the end of the Mod 4 tests which never would appear in the Navy's computer records but would provide us with response data.

As a simple illustration, assume that an item such as A is found to suggest four possible errors. We present this item first, followed by several others which confirm some errors and eliminate others. This can be diagrammed in a process network where each node represents a decision branch. By tracking these decisions the computer can identify the student's errors, provided the student is entirely consistent and is following a predicted route. Unfortunately, the realities of human behavior do not conform to such a deterministic model. By using pattern classification techniques, the rule space approach adds an element of probability to the model, with the promise of much better rates of error detection.

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46(4), 443-459.

Tatsuoka, K. K. (in press). Diagnosing cognitive errors: Statistical pattern classification based on item response theory. Behaviormetrika.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classifcation approach. Journal of Educational Statistics, 10, 1, 55-73.

Tatsuoka, K. K., & Tatsuoka, M. M. (1985). Bug distribution and pattern classification (Technical Report 85-3-ONR). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Tatsuoka, K. K., Tatsuoka, M. M., & Baillie, R. (1984). Application of adaptive testing to a fraction test (Technical Report 84-3-NIE). Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

An Overview of the Accelerated CAT-ASVAB Program

by

W. A. Sands

Computerized Testing Systems Department
Manpower and Personnel Laboratory
Navy Personnel Research and Development Center
San Diego, California 92152-6800

INTRODUCTION

Group-Administered Aptitude Tests

Group-administered aptitude tests can be described as conventional, paper-and-pencil tests and Computerized Adaptive Tests (CAT). As the names indicate, these differ in administration mode. Less obviously, but perhaps more importantly, they also differ in the way in which items are selected for administration. In the usual paper-and-pencil test, all examinees are administered the same items in the same sequence. In contrast, a CAT instrument is dynamically tailored to the measured ability level of the individual examinee, during the course of the test administration. This means that, at least potentially, every individual receives a different test.

Typically, in a CAT administration, the first item selected for administration is one of medium difficulty, since we know nothing about the examinee's ability level. If the examinee responds correctly, the ability estimate is raised to above average, and a more difficult itam is selected for administration. If the examinee answers this second item incorrectly, the ability estimate is lowered somewhat through the updating procedure. As a result, an easier item is selected as the third question. This process of selecting an item, scoring the examinee's response, updating the ability estimate, and choosing the next item for administration continues until some stopping rule is reached. This test termination criterion may be either the administration of a prespecified number of items (fixed length testing), or the administration of items until the ability estimate meets a prespecified level of precision (variable length testing).

Armed Services Vocational Aptitude Battery (ASVAB)

The ASVAB is a conventionally-administered, paper-and-pencil aptitude test battery used by all the U.S. military services for both enlistment eligibility screening and for subsequent classification and placement into entry-level training. The paper-and-pencil version of the battery (P&P-ASVAB) includes eight power tests and two speeded tests. Administration time for P&P-ASVAB takes about three and one-half hours.

The P&P-ASVAB is administered under two large-scale testing programs. The Production Testing Program involves the administration of the battery in the 68

Military Entrance Processing Stations (MEPS) and in about 900 Mobile Examining Team (MET) sites located across the country. The Student Testing Program is administered in about 14,000 high schools. These two testing programs are quite large, each involving the administration of the battery to between 800,000 and 1,000,000 persons annually.

COMPUTERIZED ADAPTIVE TESTING VERSION OF ASVAB (CAT-ASVAB)

Objectives

The Computerized Adaptive Testing (CAT-ASVAB) Program has two broad objectives. The first involves the development of a system that automates the administration, test scoring, and computation of the Armed Forces Qualification Test (AFQT) score and various other composite scores derived from ASVAB used by the individual military services. Such a system must be capable of use in both the fixed-base MEPS and in the portable testing environment of the MET sites, while interfacing with the existing score reporting system. The second objective of the CAT-ASVAB program is to evaluate the suitability of CAT-ASVAB as replacement for the P&P-ASVAB in the Production Testing Program.

Approach

The original approach to the development of the CAT-ASVAB System was a three-stage competitive "flyoff" between three contractors from private industry. During the first stage, the three contractors developed system design concepts and supporting analyses. In the second stage, limited production models were to be developed, field-tested, and evaluated. The final stage would have involved one of the three original contractors going into full-scale production, deployment, and implementation.

The approach has changed as a result of three factors. First, the timelines submitted by the contractors for Stage 2 were considerably longer than we had planned. Secondly, some remarkable advances have been made in microcomputer technology during the past few years. Finally, LTGEN E. A. Chavarrie, Deputy Assistant Secretary for Military, Manpower and Personnel Policy, in his keynote address at the last MTA convention in Munich, provided strong encouragement for reducing the long timelines, commensurate with meeting the performance objectives of the program. As a result of these influences, we have adopted a markedly different approach. With a focus on early implementation, we have initiated work on the Accelerated CAT-ASVAB Program (ACAP).

ACCELERATED CAT-ASVAB PROGRAM (ACAP)

Objective

The objective of ACAP is to field-test CAT-ASVAB as soon as possible. In pursuit of this objective, we will procure off-the-shelf, commercially-available microcomputer equipment. Software will be designed and developed in-house at NPRDC.

Evaluation Criteria

ACAP will be designed to meet the nine evaluation criteria originally established for the full-scale version of CAT-ASVAB: (1) performance, (2) suitability, (3) reliability, (4) maintainability, (5) ease of use, (6) security, (7) affordability, (8) flexibility/expandability, and (9) psychometric acceptability. Each of these nine major criteria includes numerous subcriteria.

The performance criterion includes both general and specific requirements. The system must automate the current P&P-ASVAB functions and anticipated additional functions of CAT-ASVAB. System response time cannot exceed a maximum of two seconds, and this response time must be independent of the number of examinees taking the test (system load). The display must have a resolution of 400x300 pixels, and the test must be displayed in 7x9 characters. The system must support an interface with the existing MEPS Reporting System, and the CAT-ASVAB Maintenance and Psychometric (CAMP) facility. The computer software should employ a "top-down", structured design, use a high-level language, and be adequately documented.

Suitability requires the system to operate in a normal office environment in terms of temperature and humidity. No significant modification to existing facilities should be required (e.g., electrical power). There should be no necessity for specially skilled operators and no significant staffing changes should be required. Finally, the system must be portable to support testing in the MET site environment.

Reliability is an important concern for the system. It is imperative that the system be available and operate reliably for scheduled testing sessions. The system must be capable of restarting from the point of failure and recovering from failure without loss of data.

Maintainability is a major consideration for any large scale computer system. No skilled technicians are to be required. The hardware/software must incorporate self-diagnostic capabilities which can be readily understood by test administrators. An adequate integrated logistics support system must be established and maintained for the life of the system.

The system must be easy to use, both for the test administrator and the examinee. No computer experience or expertise should be required. Set-up procedures should be clear, unambiguous, and adequately documented. The display legibility and resolution should support both text and graphics material. Introduction of experimental test items should be transparent to both the examinee and the test administrator.

Security is an important system consideration. The item sequence should be unpredictable. Measures must be taken to prevent printout or inspection of the item files. Use of the system must be limited to authorized personnel. A multi-level password access procedure should be implemented. System access must produce an audit trail which can be inspected by system managers. Finally, the system should be designed to minimize equipment theft, by reducing or eliminating pilferable components.

The system must be affordable, i.e., the life-cycle cost of CAT-ASVAB must be comparable to that of P&P-ASVAB. At present, an economic analysis of the system is being conducted under contract.

Flexibility/expandability is an important dimension of the system. In addition to supporting the delivery of a CAT version of ASVAB, the system must allow for future, add-on peripheral devices. While present plans call for the examinee to use a specially-designed keypad input device, the system must support standard keyboard input. A programmable, high-precision clock is essential, as future testing will almost certainly involve the measurement of response latency. Other future testing possibilities include provision to measure a person's ability to identify and track a moving target. This will require a system capable of internal and/or external expansion and provisions for additional interfaces (e.g., a joystick).

Finally, the system must be psychometrically acceptable. CAT-ASVAB must measure the same aptitudes measured by P&P-ASVAB. The CAT-ASVAB and P&P-ASVAB versions of the battery must be equated to insure that the scores are interchangeable. The CAT-ASVAB system must meet stringent professional test standards.

Progress/Plans

We have already achieved several goals. We have developed the functional requirements for the system. An equating plan has been developed, and is currently under review by policy and technical representatives from each of the services and the U.S. Military Entrance Processing Command (USMEPCOM). We have procured a small number of development systems to begin software design/development. A procurement action to obtain equating systems has been initiated and is in progress.

Plans include completion of the design and development of the ACAP software in 1986. Data collection and extensive statistical analyses will take about a year. We intend to begin the initial operational test and evaluation in selected MEPS (and their satellite MET sites) in 1987.

CONCLUSION

There have been a number of important changes in the development of the CAT-ASVAB system since the last Military Testing Association meeting in November 1984. Emphasizing the importance of demonstrating the extensive capabilities of an adaptive testing system for ASVAB, we are concentrating on the Accelerated CAT-ASVAB Program (ACAP). However, ACAP is viewed as an interim system, not as a replacement for the full-scale development of CAT-ASVAB. We plan to use "lessons learned" from the limited deployment of ACAP to strengthen our functional requirements specifications for the development of the full-scale CAT-ASVAB system.

Ronald B. Tiggle, Ph.D.
and
Mr. Bernard A. Rafacz
Navy Personnel Research and Development Center
San Diego, California, 92152-6800

NAVPERSRANDCEN is currently involved in a major system development effort that is concerned with the research, development, and eventual implementation of a Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery(CAT-ASVAB). The goal of this effort is the implementation of CAT-ASVAB on a nationwide distributed computer network. This network will permit the United States Military Entrance Processing Command (USMEPCOM) to adaptively administer the ASVAB to civilian applicants for military service. The CAT-ASVAB System is intended to replace the operational paper-and-pencil battery (P&P-ASVAB) currently used for selection and classification of enlisted personnel.

P&P-ASVAB testing currently occurs at 68 Military Entrance Processing Stations (MEPS), two substations and approximately 900 field locations identified as Mobile Examining Team (MET) sites. The MEPS/MET sites are under the administrative responsibility of USMEPCOM. CAT- ASVAB System components will be used in both of the MEPS and MET testing sites. The system must provide an automated, on-line system for test delivery and score reporting using adaptive, conventional, and timed psychometric tests. Item response theory (Lord, 1980) constitutes the theoretical foundation for CAT-ASVAB adaptive testing.

CAT-ASVAB System Concept

In order to propose a system concept/design for a Local CAT- ASVAB Network (LCN), it is necessary to have available a set of specifications and standards for the performance of the system once implemented. Fortunately a government-specified set of performance standards exists, and it has been documented in the CAT-ASVAB Stage 2 Full Scale Development (FSD) Request for Proposal. This document outlines the functional requirements for the development of computer hardware specific to CAT-ASVAB functions. Before the FSD is implemented, an early, smaller-scale development known as the Accelerated CAT-ASVAB Project(ACAP) will be completed to provide pilot data in support of the CAT Stage 2 FSD. The scope of the ACAP effort will conform as much as possible to the CAT-ASVAB Stage 2 FSD. However, the CAT- ASVAB functions to be addressed under ACAP will be dependent on the computer hardware design that is selected. Unlike the FSD, it is not a goal of ACAP to develop computer hardware specific to CAT- ASVAB functions. Rather a system design will be selected from a set of candidate designs, and then commercially-available computer systems will be surveyed in order to identify the most appropriate system to meet the functional requirements of CAT-ASVAB.

CAT-ASVAB Operational Requirement

The primary requirement of the CAT-ASVAB System is that it be capable of administering a battery of instruments, equivalent to the present components of the P&P-ASVAB. The current production battery (P&P- ASVAB) includes 10 tests; eight of these are cognitive power tests; two are speeded tests. However, once it is implemented, CAT-ASVAB will be capable of administering other cognitive and non-cognitive operational and experimental instruments, as determined by DoD policy.

At present, 20 percent of the production P&P-ASVAB testing occurs at MEPS; the remaining 80 percent occurs at MET test sites.

MEPS P&P-ASVAB testing is currently but one part of the processing of applicants for enlistment, and it occurs at fixed-site locations in relatively controlled environments. Experienced and well-trained examiners conduct the testing sessions. In contrast to the MEPS testing environment, the MET site testing is, in a large number of cases, administered by an Office of Personnel Management (OPM) employee working under a service agreement with DoD. For the most part, MET site testing is conducted in borrowed facilities on an ad-hoc basis. USMEPCOM has no permanent control over the MET facilities, and no authority to modify them. Thus, the CAT-ASVAB System will be required to be used by non-USMEPCOM examiners who must set-up and take- down, possibly even transport, the CAT-ASVAB testing equipment to and from non-USMEPCOM facilities as required to support examining schedules.

Functional Requirements

Based on the functional specifications stated in the CAT Stage 2 RFP, it is intended that a Local CAT Network(LCN) be developed that would permit the administration of CAT-ASVAB to civilian applicants at any of the MEPS or MET sites within CONUS. An LCN would consists of (up to 24) Examinee Test (ET) Stations linked to a single Test Administration (TA) Station via a hard-wired electronic telecommunications line. In addition, a Data Handling Computer (DHC) would reside at each MEPS to support the telecommunications function among LCN units located at the MEPS and the MET sites for that MEPS, and with USMEPCOM MEPS computer systems to be used for archiving CAT-ASVAB data.

MEPS Sites. The functional capability that is required at the MEPS, in terms of psychometric testing, is identical to that required at the METS. Specifically, MEPS equipment is stationary, but identical to MET site equipment. Identical LCN components at MEPS and MET sites are necessary to accommodate the equating of CAT-ASVAB, commonality of equipment for software and hardware maintenance purposes, and to permit the cost effective sharing of equipment across both types of sites. In contrast to most MET sites, each examiner at a MEPS testing site must be able to monitor up to 24 ET Stations in any LCN.

The MEPS site implementation of CAT-ASVAB must also include a DHC unit. The main function of the DHC is to collect data daily from each LCN within the associated MEPS administrative segment, including LCN's located at MET sites. Data is transmitted to the DHC either over a hard-wired connection (in the case of MEPS LCN's) or modems in the case of MET LCN's.

MET Sites. At the MET sites, transportable computer systems will be used to administer CAT-ASVAB. The hardware configuration is to be based on the concept of a "generic" LCN. This generic LCN will consist of six ET Stations being monitored by a single TA Station, including any peripheral equipment. Note that many more (up to 24) ET Stations must be monitored by a single TA station and still maintain CAT-ASVAB performance requirements. In addition, it is important that the selected equipment support the CAT- ASVAB Stage 2 portability requirements; i.e., number of packages and weight requirements for a generic LCN (no more than eight components weighing a total of no more than 120 lbs., each component weighing 23 lbs). Environmental preformance requirements such as temperature, humidity, etc. must also be met.

The computer hardware configuration for an LCN may be described as follows: Each ET Station would consist of a response device, a screen display, and include access to sufficient RAM and/or data storage to permit the administration of any CAT-ASVAB test; the amount of RAM required depends on the particular application software and networking design being used to implement the functions. Each ET Station would be tied into a TA Station by networking cables; the TA Station being essentially an ET Station with a mass storage device and full-sized keyboard attached. Finally, a single(very portable) printer and modem for the TA Station would complete the complement of equipment composing the LCN.

The operational requirement for an LCN is to administer CAT-ASVAB to those military applicants scheduled for testing at the MET site. Initially, an Office of Personnel Management (OPM) examiner may be required to pick- up the LCN equipment at a staging area; if it were not secured at the testing site itself. The equipment would be transported to the test site, carried from the vehicle to the test site, and configured, ready for testing by the examiner. Once all examinees have completed testing, the examiner will attempt to telecommunicate examinee personal and test item response data to the DHC unit at the associated MEPS. This will be done using a modem and dial-up telephone line, if available at the test site. If this is not possible, the data will be transferred once the equipment is returned to the staging area.

Finally, if the equipment were not secured at the testing site, the examiner would package the equipment into transportable packages, carry these packages to a vehicle, and return the equipment to the staging area.

Design Considerations

It can be concluded from the survey of the requirements for developing a CAT-ASVAB System, that design efforts must be focused on the requirements for a generic LCN. This mainly includes the portability and functional capability of an LCN.

Portability. With respect to the portability aspect of an LCN, it should be clear that the response of the ACAP system to this requirement will be contingent on the capabilities of currently available commercial computer hardware. The ACAP System will not include the development and/or the building of computer systems to meet CAT-ASVAB needs, but rather will use commercially available computer hardware to support the accelerated field- test effort.

On-Line Data Storage. The on-line data storage requirement is the factor which will most influence the selection of computer hardware to support the ACAP System development. On-line data storage requirements will also significantly impact upon the design of the software for the ET, TA, and DHC units. The TA and ET Stations must have access to the on-line storage of two Forms of the CAT-ASVAB. Based on an estimate of 100 items per test, each Form will require approximately 850 Kilobytes (KB) of storage; assuming that the data is stored as a sequential file or equivalent. Actual data storage requirements would be 30-50% higher if the system design required that the data be stored as a random access file. Only one Form must be available to an ET Station during a test session. Additional storage for use by the TA and ET Station will have to be allocated for two experimental item pools, of 170 KB each (one pool for each form); an experimental item set derived from the experimental item pool, (10-125 KB); application software (approximately 250 KB per unit); two survey questionnaire item banks, 80 KB for both questionnaires; and examinee personal data such as information required on the USMEPCOM 714-A Form; and examinee personal and test item response data, 15 KB per examinee.

Test Administration Time. Depending on the network that is proposed to support ACAP, the time required to administer a CAT-ASVAB test could also affect the minimum response time required to have test data for the next item available at ET Stations.

Candidate Local CAT-ASVAB Network Designs

Using the preceding design considerations as a guide, the current commercial market in portable (or, at least, transportable) computer systems indicate that there are three basic designs upon which to develop a workable ACAP System. For the purpose and scope of the ACAP effort, the discussion will be focused upon the storage (and retrieval) capability of the candidates with respect to the 850 KB test item bank. It should be remembered that the important consideration is that the examinees have available (within the response time specified in the Stage 2 RFP), the correct test item as dictated by the underlying item selection strategy being used at the time.

The basic designs to be discussed are generic designs and, as such, may not be entirely represented by an actual example on the commercial market. The examples provided are those system designs that come very close to representing the generic design being discussed.

Design # 1 - CAT-ASVAB Items Stored on Removable Media

In this type of design, each ET Station would consist of sufficient internal storage on removable media (e.g., 3 1/2 in. micro-floppy diskettes) to accommodate the storage of the entire test item bank. Internal RAM would be about 512 KB. Necessarily, the test item bank files would have to be encrypted or "scrambled" on the removable media. The ET Station would be very portable and weigh from 11-17 lbs., and include a flat panel Liquid Crystal Display (LCD), electroluminescent, or equivalent low-weight display. The computer system that may currently best illustrate this hardware configuration is the Data General/One.

The main advantage to this type of configuration is that it is basically a very portable system. The total weight for a generic LCN could be as low as 80 lbs., including seven very transportable components (assuming that the LCN Networking requirement was suppressed).

There are many disadvantages to this design.

1. Security. The entire test item bank must reside on the removable media, necessarily jeopardizing the security of the test item bank files.

2. Media Updating. Each ET Station will require two removable diskettes installed in the disk drives to accommodate CAT-ASVAB testing. If this design were going to be installed at the ACAP field-test MEPS, and MET sites, approximately 400 ET Stations would be involved. This would require 800 micro-floppy diskettes to be inventoried and secured; a very large media creation, distribution, and security problem each time the test item bank is updated.

3. Ease of Use. Use of a removable storage media will require a significant amount of operator intervention to insert/remove diskettes. For each ET Station, two movements are required to "boot" the computer and to receive testing software. Eight diskette movements are required to transfer the examinee's response data to the examiner's work diskette; two movements can be avoided for subsequent examinees. A MET site LCN testing 10 individuals with six ET Stations available would require at least 100 movements.

4. Maintenance. Since the micro-floppy drives are in constant use during the testing process, system maintenance may be higher than some other configuration in terms of disk drive maintenance and diskette replacement. Each test item being displayed will require at least one disk access.

Design # 2 CAT-ASVAB Items Stored on a Central File Server

This type of LCN design is configured around a central file server (e.g., a hard disk) which acts as the repository for the CAT-ASVAB item banks and supporting data files. In this type of design, the capabilities of the network supporting the movement of data from the file server to each examinee station is of paramount importance. A minimal amount of RAM is available at the ET Station (less than 512 KB), with perhaps one internal floppy disk drive available. A central file server is required because each ET Station cannot support the entire requirement of data storage without adding significantly more components to the overall network. Typically, the ET Station is bulkier and may not support the latest flat screen technology; e.g., LCD or electroluminescent displays. The Macintosh Computer by Apple Corp. is an example of this type of design. In general, any configuration of equipment requiring a central file server would be an example of this design.

Perhaps the main advantage to this type of network is that a very sophisticated networking capability must be installed in order to "make it work". Because this capability is available, one could, (theoretically) also install a TA Station monitoring capability. The monitoring capability would have to be installed in such a fashion so as not to compromise the response time requirements for the display of test items at the ET Stations in the LCN. Admittedly, the monitoring function could be potentially very useful at certain large MEPs, in which many examinees are being examined simultaneously. Unfortunately, this is also the situation in which the ET Station response time requirement would be most compromised. However, this system could be the least expensive of the networks being investigated. The cost of the file server could be distributed over each ET Station which could function without any removable media being available. Another advantage (as opposed to Design #1) is that the movement of test item bank data and/or examinee response data to/from the ET Station is automatic and does not require examiner intervention. Once the network is set-up and working properly, the examiner's tasks are minimized with respect to data movement requirements.

The main disadvantage to this design is concerned with the reliability of the file server itself; when the file server fails, the entire network is inoperable. To meet the CAT-ASVAB Stage 2 reliability requirements, it will be necessary to include two identical such in the LCN. This means that the system is heavy, environmentally intolerant, and requires a large number of components to be transported. In addition, for MET sites in which the equipment has to be assembled, and disassembled, during each testing session, a heavy requirement will be placed on the examiner to serve as a computer operator. This could conceivably result in a reclassification of the OPM examiner position.

Another disadvantage is that the system response time and monitoring requirements are functionally related in Design#2. It is very difficult to imagine a network operating system that can simultaneously accommodate these requirements in a fully loaded LCN; 24 ET Stations attached to a single TA Station. What is the maximum response time at any ET Station when all examinees are requesting items (simultaneously) from the same source (i.e., file server)? Note that sufficient time must also be allowed for de-encryption of the items before display, as well as decompression of graphics items, as necessary. In summary, another disadvantage of Design # 2 is that the maximum system response time is relatively large compared to other system designs, and therefore is a potentially compromising consideration relative to hardware selection for purposes of ACAP.

Design # 3 - Test Item Banks Stored In Examinee Test Station Random Access Memory
The TA and ET Stations in this design would consist of a large amount of internal RAM, on the order of at least 1.5 MB. The ET Station would be supported by one micro-floppy drive and probably include the latest in electroluminescent or LCD display. Therefore, for purposes of recovery due to network failure, the ET Station would be very responsive as it is capable of operating independently. In addition, as a networking capability will be available, the TA Station could perform the functions of an "electronic" file server. Total RAM available on the TA Station could be 1.5 to 3.5 MB (preferably higher); allowing for great flexibility in the total number of alternate forms available during any one test session.

Note that several removable media are required in order to "boot" the systems. However, a total of no more than TWO micro-floppy diskettes are required to store the test item banks (per Form) and supporting data files; each ET Station would also require one micro-floppy to be installed as a "working" diskette for failure recovery purposes. Normally (after initial "boot-up" at the beginning of a test session), no micro-floppy diskette movements are required by the examiner; i.e., the network would accomplish all data movements.

The main advantage of this design is that it offers a large degree of flexibility with respect to design options. The ET Stations are capable of operating as stand-alone devices and, as such, it is virtually impossible for an examinee's test session to fail to be completed; each ET

Station backs-up every other ET Station. For this reason, and as it minimizes accesses to a mechanical device, Design # 3 should be the most reliable of the network designs discussed. In addition, this design offers a very high level of security. Once power is removed from the computer, the volatile random access memory (RAM) is erased. This provides dependable security for the test i'em banks. Furthermore, as noted above, only two removable micro- floppy diskettes are required per form; regardless of the number of stations in the LCN.

Another important consideration, in comparison to Design # 2, is that the LCN monitoring and the system response time requirements are not functionally related. In addition, it is possible to configure a collection of computer hardware for Design # 3 that permits the entire 850 KB test item bank (for a form) to reside at the ET Station. Therefore, the response time required to display test items will be network independent. The item display process will be accomplished at RAM speed, resulting in a maximum response time that is on the order of millisecconds, as opposed to seconds.

Generally speaking, up to failure recovery, during actual test administration, the advantages of Design # 3 include: a) no removable media are requires, b) minimization of the need to use mechanical devices, c) a high level of security for test item banks, d) excellent system response time characteristics, and e) de-encryption of test item bank need occur but once; when this information is initially transferred to the "electronic" file server. In addition, some of the better features of Design # 2 are also characteristic of this design: a) examiner intervention to move data in an LCN is not required, and b) TA Station monitoring capability can be automated.

The primary disadvantage of this design is that it does tend to cost more than some alternatives. However, it is certainly true that the cost of computer hardware is decreasing. Another disadvantage is that there is only one viable candidate system on the market that would come close to exemplifying this design; that being the Hewlett-Packard Integral Computer. This system is currently capable of 1.5 MB of internal RAM (to include a networking interface card), with a single 3 1/2 in., 710 KB capacity, micro- floppy. Finally, in the case of the Integral, the ET Station would be somewhat heavier than some alternatives, weighing approximately 23 lbs. (assuming that the printer is removed).

Recommendations.

Of the three Designs discussed, the authors recommend Design # 3. They believe that it will give the Government the greatest amount of flexibility as the ACAP system is field-tested and the LCN configuration needs to be adjusted to accommodate future requirements. In addition, this design offers a system with the greatest amount of reliability and test item security. No mechanical devices are required to maintain normal CAT-ASVAB testing (up to recording examinee data on the micro-floppy associated with an ET Station for back-up purposes in the event that station fails during a test session). Test item bank data will be stored in volatile RAM instead of removable media, insuring the erasure of this sensitive information immediately when power is removed.

REFERENCES

(6 June 1984). Computerized Adaptive Testing Stage 2 Full Scale Development Request for Proposal issued by the Navy Regional Contracting Office, Long Beach.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hilldale, NJ:Erlbaum.

A Validity Study of the Computerized Adaptive Testing Version of the Armed Services Vocational Aptitude Battery

Kathleen Moreno & Daniel O. Segall
Navy Personnel Research & Development Center

William F. Kieckhaefer
RGI, Incorporated

PURPOSE

The purpose of this study was to assess the construct and predictive validity of a Computerized Adaptive Testing (CAT) version of the Armed Services Vocational Aptitude Battery (ASVAB). This study was conducted as part of an effort to evaluate CAT-ASVAB as a replacement for the paper-and-pencil ASVAB (P&P-ASVAB).

BACKGROUND

Over the past decade numerous empirical studies have been conducted to evaluate the construct and predictive validity of adaptive aptitude testing (Sympson and Moreno, 1985). Overall, the results of these studies indicate that adaptive testing is as valid as conventional, paper-and-pencil testing. However, the majority of these studies were conducted using verbal ability or arithmetic reasoning tests. Very little research has been conducted using aptitude tests measuring other types of ability.

The battery of interest in this study, the ASVAB, consists of tests measuring ten types of ability. The ASVAB is used by all military services for selection and classification of military applicants. This study examined the construct and predictive validity of all ASVAB tests.

APPROACH

Examinees

Examinees were military recruits scheduled for training in one of the military service specialties selected for inclusion in this study. Over all services, 7,515 examinees were tested. Sample sizes for each specialty are shown in Table 1.

Tests

P&P-ASVAB. The P&P-ASVAB is a group-administered, conventional battery in which all examinees answer all items in the same sequence. The P&P-ASVAB consists of eight power tests and two speeded tests: General Science (25),[1] Arithmetic Reasoning (30), Word Knowledge (35), Paragraph

[1] Values in parentheses are test lengths.

Comprehension (15), Auto and Shop Information (25), Mathematics Knowledge (25), Mechanical Comprehension (25), Electronics Information (20), Numerical Operations (50), and Coding Speed (84). There are six parallel forms of the P&P-ASVAB. This study used forms 8A and 9A. Number correct scores served as estimates of ability.

CAT-ASVAB. The CAT-ASVAB used in this study is an experimental version designed to measure the same abilities as those measured by the P&P-ASVAB. There are nine power tests and two speeded tests: General Science (197),[2] Arithmetic Reasoning (166), Word Knowledge (194), Paragraph Comprehension (95), Auto Information (179), Shop Information (189), Mathematics Knowledge (135), Mechanical Comprehension (70), Electronics Information (168), Numerical Operations (50), and Coding Speed (84). The nine power tests were administered adaptively using maximum information item selection and bayesian scoring. All power tests were terminated at a fixed length of 15 items, expect for Paragraph Comprehension, which was terminated after 10 items. The two speeded tests were administered in a conventional manner, with the test terminating after a fixed time. For the speeded tests, number correct scores were used to estimate ability.

Criterion Variables

Since each service has numerous training schools, prediction of performance for only a selected number could be assessed in this study. Schools were selected so that a wide variety of specialties would be represented. In addition, since military services use composites of test scores for selection and classification, schools were selected so that school composite scores would span all ASVAB tests. Table 1 lists the selected specialties and the criteria used for each specialty. For the majority of specialties, final school grade (FSG) or time to completion (TC) was used. However, for some specialties these measures were not available. In these cases, analyses were performed to determine which measure should be used.

Procedure

Examinees were tested approximately two weeks after arrival at a recruit training center, prior to entrance into training schools. Examinees were group-administered the experimental CAT-ASVAB and those P&P-ASVAB tests that were used in computing a recruit's school selection composite score. The tests were counter-balanced so that half the examinees took CAT-ASVAB first and half took P&P-ASVAB first. The CAT-ASVAB was administered using Sanyo monitors and Apple III computers networked with a Corvus hard-disk drive.

Pre-enlistment ASVAB scores were collected on all examinees from DD forms 1966. School performance data were collected after examinees had completed training.

[2] Values in parentheses are item pool sizes.

Data Analyses

Predictive Validity. For each of the selected specialties, composite scores were computed from CAT-ASVAB standardized scores and from P&P-ASVAB standardized scores. Validity coefficients were obtained for each test version by correlating school composite scores with school performance data. In order to test for significant difference between test versions, *t* values were computed. Since examinees in this study were a selected sample from the military applicant population, validity coefficients were corrected for range restriction using a multivariate approach (Lawley, 1943). No significance testing was performed using corrected validity coefficients.

Construct Validity. For each service, the intercorrelation matrix of CAT-ASVAB and P&P-ASVAB test scores was factor analyzed using the principal axes method, followed by a varimax rotation to simplify the factor structure.

RESULTS

Predictive Validity

Table 1 shows the validity coefficients obtained using CAT-ASVAB and P&P-ASVAB. Significance tests revealed no differences between the validity coefficients for the two test versions, even though CAT-ASVAB tests are much shorter than P&P-ASVAB tests.

Construct Validity

Table 2 shows the results of the factor analysis using data from the Air Force sample. Four factors were extracted, based on an eigenvalue of 1.0 or greater. These factors have been labeled as technical, verbal, mathematical, and speeded factors. As shown, the CAT-ASVAB tests had similar loadings to those of the corresponding P&P-ASVAB tests. Findings were similar for the other three services.

CONCLUSIONS

These results suggest that CAT-ASVAB is a viable alternate to P&P-ASVAB. In this study, CAT-ASVAB tests seem to be measuring the same abilities as the P&P-ASVAB tests and predict school performance as well as P&P-ASVAB tests, even though CAT-ASVAB test lengths are much shorter. However, before replacing the P&P-ASVAB with CAT-ASVAB, the two versions should be compared in terms of differential prediction by test version and subgroup membership. Such analyses are currently being performed at NPRDC.

Table 1

Validity Coefficients for CAT-ASVAB and P&P-ASVAB

Specialty	Criterion	Sample Size	Validity Coefficients CAT-ASVAB	 P&P-ASVAB
NAVY				
Radioman	TC	186	-.41 (-.69)	-.40 (-.68)
Mess Management	FSG	170	.45 (.74)	.40 (.71)
Hospital Corpsman	FSG	192	.56 (.77)	.60 (.80)
Electronics Tech.	TC	143	-.41 (-.80)	-.46 (-.83)
Hull Maint. Tech.	FSG	170	.35 (.75)	.35 (.75)
Sonar Tech.	FSG	205	.46 (.79)	.46 (.80)
AIR FORCE				
Avionics	FSG	147	.56 (.80)	.56 (.80)
Administration	FSG	208	.22 (.57)	.15 (.53)
Aircraft Maint.	FSG	245	.54 (.81)	.49 (.79)
Medical	FSG	95	.64 (.87)	.62 (.86)
Security Police	FSG	456	.49 (.78)	.45 (.77)
MARINES				
Avionics	FSG	228	.49 (.81)	.48 (.81)
	TC	228	-.58 (-.84)	-.54 (-.84)
Administration				
Lejeune	FSG	72	.14 (.32)	-.05 (.13)
Pendelton	FSG	39	.22 (.49)	.29 (.56)
Aircraft Mech.				
School 1	FSG	181	.34 (.54)	.30 (.53)
School 2	FSG	69	.50 (.70)	.47 (.70)
Motor Transport	FSG	151	.29 (.47)	.30 (.48)
Combat Engineer	Sum(All)	123	.69 (.82)	.66 (.80)
Field Radio Opr.	Sum(1-4)	128	.33 (.43)	.21 (.37)
	Sum(5-8)	128	.06 (.46)	.09 (.47)
ARMY				
Infantry	Sum(All)	329	-.24 (-.34)	-.28 (-.36)
Mechanic				
Fort Dix	Average	198	.57 (.74)	.59 (.76)
Fort Jackson	PC	186	.35 (.52)	.38 (.54)
Motor Transport	Sum(All)	277	-.47 (-.63)	-.44 (-.63)
Administration	Wtd Sum	145	-.35 (-.64)	-.44 (-.72)
Telecom. Opr.	Sum(All)	169	.15 (.28)	.22 (.32)
Medical	FSG	225	.63 (.85)	.59 (.83)

Note. TC is the time for course completion; FSG is a final school grade; Sum(All) is a sum of training module scores; Sum(1-4) is a sum of scores on training modules one through four; Sum(5-8) is a sum of scores on training modules five through eight; Average is an average of the scores on all training school modules; PC is a percent correct score on the end-of-course test; and Wtd Sum is a sum of module scores minus a weighted typing score.

Numbers in parentheses are validity coefficients corrected for restriction in range.

Table 2

Results of a Factor Analysis of CAT-ASVAB and P&P ASVAB Scores from an Air Force Sample

	Varimax Rotated Factor Matrix			
	Factor 1 (Tech)	Factor 2 (Verbal)	Factor 3 (Math)	Factor 4 (Speeded)
P&P-ASVAB				
AR	.28	.15	.66*	.31
WK	.17	.82*	.08	.05
PC	.16	.50*	.13	.16
NO	-.07	.00	.24	.70*
GS	.40*	.63*	.22	-.01
CS	.00	.01	.06	.71*
AS	.82*	.14	-.01	.00
MK	.17	.25	.80*	.23
MC	.65*	.21	.36*	-.01
EI	.61*	.32*	.19	-.03
CAT-ASVAB				
AR	.31*	.27	.71*	.22
WK	.15	.85*	.16	.03
PC	.17	.68*	.23	.13
NO	-.03	.13	.26	.65*
GS	.33*	.73*	.28	.00
CS	-.03	.10	.08	.71*
AS	.90*	.15	.09	-.02
MK	.08	.29	.74*	.23
MC	.66*	.22	.30*	-.03
EI	.64*	.42*	.26	-.08

*factor loading > .30

REFERENCES

Lawley, D. (1943). A note on Karl Pearson's selection formulae. *Royal Society of Edinburgh Proceedings, Section A, 62,* 28-30.

Sympson, J. B., & Moreno, K. E. (1985, August). *Validity of adaptive testing: A summary of research results.* Paper presented at the American Psychological Association Convention, Los Angeles, CA.

Assessment of the Unidimensionality of CAT-ASVAB Subtests

Mary K. Schratz
Navy Personnel Research and Development Center
San Diego, California 92152-6800

Background

A joint-service project is underway to develop a computerized adaptive testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB). CAT-ASVAB is intended to replace the paper-and-pencil ASVAB (P&P-ASVAB), used by the four military services to select and classify applicants. The present P&P-ASVAB battery consists of ten aptitude subtests. These subtests, and the number of items included in each for a given form of P&P-ASVAB, are listed in Table 1 below.

Table 1. P&P-ASVAB Subtests and the Number of Items Included per Form

Subtest	Number of Items
General Science (GS)	25
Arithmetic Reasoning (AR)	30
Word Knowledge (WK)	35
Paragraph Comprehension (PC)	15
Numerical Operations (NO)	50
Coding Speed (CS)	84
Auto and Shop Information (AS)	25
Mathematics Knowledge (MK)	25
Mechanical Comprehension (MC)	25
Electronics Information (EI)	20

The Numerical Operations and Coding Speed subtests are speeded measures; all other subtests are power measures. All examinees taking a given P&P-ASVAB form are administered the same set of test items in all subtest areas.

While CAT-ASVAB is intended to measure the same subtest areas as P&P-ASVAB, the power subtests of CAT-ASVAB will be administered adaptively. Within each power subtest, different items will be selected for computer administration to examinees, depending on their performance on previously administered items. The testing process is thus individualized for examinees.

The Problem

The theoretical framework supporting the adaptive testing process to be implemented in CAT-ASVAB is item response theory (Lord, 1980). The item response theory methods to be used assume that each subtest is unidimensional. The outcome of this assumption is that each item included in a given subtest should measure the same unitary construct, in addition to having specific and error variance components associated with it. The Committee for an Evaluation Plan for the Computerized Adaptive Vocational Aptitude Battery (Green, Bock, Humphreys, Linn, & Reckase, 1982) states that unidimensionality is always advisable for tests of ability, but is more important for adaptive tests. Though some may be of the opinion that the IRT model to be applied in CAT-ASVAB is strong enough to counteract potential problems with respect to multidimensionality in CAT-ASVAB subtests, further study of this problem is necessary. Unidimensionality may be considered an important problem for an adaptive test because different items are administered to examinees. For a traditional test such as the P&P-ASVAB, all examinees are presented the same test items, and are given an equal opportunity to respond to them irrespective of their dimensionality. The individualized nature of adaptive tests may raise questions related to test fairness. The present paper describes the approach to be taken, progress, and plans for dimensionality analyses of CAT-ASVAB items.

Approach

A recent advance in the development of factor-analytic approaches to exploring test dimensionality is "full-information item factor analysis." Bock, Gibbons, and Muracki (1985) contend that of the various methods that have been proposed for investigating dimensionality of item sets, item factor analysis is the most sensitive and informative. This method of item factor analysis is based on item response theory; it uses all data as distinct item response vectors. Thurstone's multiple factor model is used. The procedure is implemented by marginal maximum likelihood estimation and the EM algorithm. Statistical significance of the addition of successive factors to the model is tested by a likelihood ratio criterion. Provisions for the effects of guessing on multiple choice items, and for omitted and not reached items, are included.

One of the applications of this methodology to real data, presented by the authors as evidence for the accuracy and practical utility of the method, is an analysis of the power subtests of P&P-ASVAB. The analysis was conducted in a ten-percent random sample of data from the Profile of American Youth Study. The number of cases used in the analysis was 1,178, drawn from a total sample of 11,817 subjects. The details of the item factor analyses are presented in the Bock et al. (1985) report. For the purposes of this presentation, the results obtained for P&P-ASVAB are relevant to the selection of and application of an approach for studying test dimensionality in the intended replacement battery, CAT-ASVAB. Thus they will be described briefly.

For the P&P-ASVAB General Science test, Bock et al. (1985) found two significant factors; one factor was interpreted as a physical science factor and the other factor was interpreted as a biological (or health science) factor. Two factors were also found for the Arithmetic Reasoning subtest. While the second factor found was a minor one, the authors have interpreted it as a business arithmetic factor. For the Word Knowledge subtest, clear evidence for a second factor was found, though the factor has no apparent relationship to item content. For the Auto and Shop Information subtest, the authors found clear evidence for two factors separating the two types of items. For the Mathematics Knowledge subtest, two significant factors were also found; one factor involved items requiring knowledge of formal algebra and the second factor involved numerical calculation and mathematical reasoning. Only one factor was found for the Paragraph Comprehension, Mechanical Comprehension, and Electronics Information subtests.

Bock et al. (1985) conclude that the applications of the procedure reported in their paper show that, for moderately large samples, minor factors can be detected. The procedure is recommended as an exploratory technique in searching for item features that are responsible for individual differences in cognitive test performance.

Given the reported adequacy and practical utility of the full-information factor analysis approach to the detection of multidimensionality in item sets, and the indications of multiple factors found by Bock et al. (1985) in the subtests of the P&P-ASVAB battery which CAT-ASVAB is intended to replace, NPRDC plans to make use of this procedure in study ng the dimensionality of the CAT-ASVAB items.

Progress and Plans

One of the requirements of the adaptive testing process is the development and calibration of a large bank of test questions, covering a wide range of difficulty for the intended test-taking population, for use in the item selection process. In a study carried out by the Air Force Human Resources Laboratory, and conducted under contract by Assessment Systems Corporation, a bank of 2,118 test items intended for operational use in CAT-ASVAB was developed and calibrated in 1983. This bank of test questions covers all of the P&P-ASVAB power subtest areas, with more than 200 items developed for each subtest. On the basis of statistical and judgemental criteria, items will be selected from the total number available for inclusion in the final CAT-ASVAB battery. One of the criteria for accepting an item for inclusion in the final battery is measurement of a one-dimensional universe represented by a pool of items. This pool may represent a subtest, or, alternatively, a subset of items within a subtest. From unidimensional pools of items, individual test items will, of course, be selected for administration to examinees in the adaptive testing process.

Preliminary full-information item factor analyses of one of the more suspect subtest areas of the CAT-ASVAB item bank, given the results of Bock et al. (1985), have been conducted. Analyses of the General Science items indicate that indeed the item factor analysis procedure is sensitive to the presence of minor factors in the data. The CAT-ASVAB General Science items were developed to measure three main subject areas: (1) Life Science; (2) Physical Science; and (3) Earth Science (Prestwood, Vale, Massey, & Welsh, 1984). Items were randomly assigned to four test booklets for calibration purposes and approximately 2,500 examinees were tested with each booklet. Inspection of the individual items, the content categories which they were written to represent, and the preliminary item factor analysis results suggests some clustering of items in terms of subject matter. Where a preponderance of items of one type have been assigned to a booklet, the procedure appears to be sensitive to the detection of minor factors.

It is NPRDC's intention to conduct full-information item factor analyses for each power subtest to be included in the CAT-ASVAB battery. This will be done in a joint calibration analysis of both the CAT-ASVAB items and P&P-ASVAB items for each subtest. While CAT-ASVAB is intended to replace P&P-ASVAB, both batteries will concurrently be administered in an operational setting. CAT-ASVAB subtest scores must be scaled to those of P&P-ASVAB and, as is the case for the present P&P-ASVAB, a single CAT-ASVAB score will be generated for each subtest. The joint item factor analysis of CAT-ASVAB items and P&P-ASVAB items for each subtest is expected to result in as many or more factors than those determined for the corresponding P&P-ASVAB subtes' 'lone. Where more than one dimension is present, the joint calibration will allow for transformation of the resulting item parameters to congruence with those parameter estimates obtained from the analysis of P&P-ASVAB items alone. Dr. Bruce Bloxom of Vanderbilt University is presently working on a procedure for combining *m* multidimensional ability scores into a single score compaiable to that obtained on P&P-ASVAB.

Recommendations

The individualized nature of adaptive tests raises interesting test development issues related to test fairness. The adaptive testing process to be implemented in CAT-ASVAB is supported by item response theory methods which assume that a single underlying trait is measured within a subtest. Bock et al. (1985) have provided a promising procedure for investigating conformance to this assumption. It is recommended that the full-information item factor analysis approach, currently being used as an exploratory technique in the development of CAT-ASVAB item pools, be considered for use in other test development applications involving item response theory methods.

REFERENCES

Bock, R.D., Gibbons, R., & Muracki, E. (1985, August). Full-information item factor analysis (MRC Report 85-1). Chicago, ILL: Methodology Research Center/NORC.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.B., & Reckase, M.D. (1982, May). Evaluation plan for the Computerized Adaptive Vocational Aptitude Battery (Research Report 82-1). Baltimore, MD: The Johns Hopkins University, Department of Psychology.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Prestwood, J.S., Vale, C.D., Massey, R.H., & Welsh, J.R. (1985, September). Development of an adaptive item pool for the ASVAB (AFHRL-TR-85-19). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

SEX DIFFERENCES IN IRT TRUE-SCORE EQUATING

Daniel O. Segall
Navy Personnel Research and Development Center

William F. Kieckhaefer
RGI, Incorporated

Kathleen E. Moreno
Navy Personnel Research and Development Center

Introduction

The current Armed Services Vocational Aptitude Battery (P&P-ASVAB) is a paper-and-pencil test with a fixed sequence of test items. The Navy Personnel Research and Development Center is developing a computerized adaptive version (CAT-ASVAB) as a possible replacement for that test. The CAT-ASVAB will tailor the difficulty of the test items administered to the individual from responses to earlier items. This testing method is expected to increase the efficiency of selecting and classifying new accessions.

If the CAT-ASVAB becomes operational, it will be implemented gradually so that some examinees will be administered the CAT-ASVAB while others will receive the P&P-ASVAB. The two versions use different estimators of ability. The P&P-ASVAB uses a number correct score for each examinee, while the CAT-ASVAB computes a bayesian estimate of ability. Scores from the two versions must be equated so that personnel selection and classification decisions do not vary between test versions.

Braun and Holland (1982) give one definition of test equating. They adopt the definition Form-X and Form-Y are equated on population P if the distribution of the transformed y scores in population P is the same as the distribution of the untransformed x scores. Applying this definition to the current problem, the CAT-ASVAB and P&P-ASVAB are equated on population P if the distribution of the transformed CAT-ASVAB scores in this population is the same as the distribution of the P&P-ASVAB scores. This definition has the desirable quality of assuring equal flow rates for the two versions of the ASVAB.

Unfortunately two tests that are equated on population P may not be equated for various subpopulations that are included in P. Test scores that are equated for the military applicant population may not be equated for either the population of female applicants, or the population of male applicants.

This paper investigates the application of IRT true-score equating to the experimental CAT-ASVAB. An effort is made to determine whether CAT-ASVAB scores can be transformed to a paper-and-pencil scale without placing either males or females at a disadvantage relative to their P&P-ASVAB scores.

Method

SUBJECTS. During April of 1984, 200 male and 200 female Army recruits at Fort Jackson, South Carolina participated in this study. Each subject took both the CAT-ASVAB and the P&P-ASVAB.

SUBTESTS. The five subtests of the P&P-ASVAB selected for this study were taken from Form 8a: Arithmetic Reasoning (AR), Word Knowledge (WK), General Science (GS), Paragraph Comprehension (PC), and Numerical Operations (NO).

There were five CAT-ASVAB subtests designed to measure the same aptitude as the P&P-ASVAB subtests mentioned above: AR, WK, PC, NO, and GS.

PROCEDURES. Proctors seated all subjects in the testing area for the ASVAB and instructed them to complete the Privacy Act Statement. Then, half the subjects took seats in the adjacent CAT-ASVAB testing area. Subjects in this condition completed the CAT-ASVAB first and the ASVAB second. Subjects in the other condition (i.e., the remaining half of the subjects) took the P&P-ASVAB first and the CAT-ASVAB second.

Equating Transformations

Although each CAT-ASVAB subtest is designed to measure the same cognitive ability as its P&P-ASVAB counterpart, the two versions are not on comparable scales. The CAT-ASVAB power subtests produce an ability estimate ($\hat{\theta}$) while the P&P-ASVAB produces a number correct score (x). Although the CAT-ASVAB subtest of Numerical Operations does produce a number-correct score similar to the P&P-ASVAB, the method of responding has been shown to effect the CAT-NO score distribution. Thus for each content area, some method of equating the two versions is necessary.

We used two different procedures to equate the five subtests, depending on whether the subtest was adaptive or speeded. We used a procedure similar to one recommended by Green, Bock, Linn, and Recakase (1985) to equate the CAT-ASVAB power subtests. This procedure transforms the thetas into expanded expected number correct (EENC) scores. We used an equipercentile method to equate the CAT-NO speeded subtest. Both procedures are described in detail below.

POWER-SUBTESTS. For the power-subtests scores, we performed a two-stage equating. First we calculated expected number correct scores for the four CAT-ASVAB power subtests. Equation (1) transformed the estimated CAT ability, $\hat{\theta}_a$, for each person a, to an expected number correct score ξ_a.

$$\hat{\xi}_a = \frac{1}{6} \sum_{k=1}^{6} \sum_{i=1}^{n} P(a_{ik}, b_{ik}, c_{ik}; \hat{\theta}_a), \tag{1}$$

where n equals the number of items in the P&P subtest, and $P(a_{ik}, b_{ik}, c_{ik}; \hat{\theta}_a)$ represents the item characteristic curve defined by the three parameter logistic model evaluated at $\hat{\theta}_a$, for item i of P&P-ASVAB form k (where k = 1,2,...,5,6)

We substituted item parameter estimates $\hat{a}_{ik}$, $\hat{b}_{ik}$, $\hat{c}_{ik}$ (Sympson & Hartmann, 1985) for the values a_{ik}, b_{ik}, c_{ik} in equation (1). Then for each estimated ability, $\hat{\theta}_a$, an expected number correct score, ξ_a, was obtained from (1).

Second, we applied a linear transformation to the scores computed from equation (1). This produced expanded expected number correct (EENC) scores that possessed the same mean and variance as the observed scores of the corresponding P&P-ASVAB. Finally, we rounded these EENC scores to the nearest integer value to produce the CAT-EENC scores. We repeated all the above procedures for each of the CAT-ASVAB power subtests (AR, WK, PC, and GS).

SPEEDED-SUBTEST. Numerical Operations (NO) was the only speeded subtest included in this study. This subtest is not adaptive and produces a number correct score.

We obtained NO-equating data from 1,364 Army recruits. Each recruit received both versions of the subtest: (a) the CAT-NO subtest, and (b) forms 8 or 9 of the P&P-NO subtest.

We used an equipercentile method to equate CA↓-NO to P&P-NO. First, we obtained CAT-NO and P&P-NO number correct scores at 99 different cumulative-percentile points. These cumulative-percentile points were obtained at unit intervals ranging from 1 to 99, inclusive. Next, we used least-squares-polynomial regression to smooth the equipercentile-equating function. We calculated polynomial regressions of several different orders and judged the quintic regression to provide the best fit based on a root-mean-squared-error criterion. We equated scores below the second percentile (of the CAT-NO score distribution) using linear interpolation from the (0,0) point to the point corresponding to the second percentile. Then we obtained smoothed-equipercentile estimates for each number-correct score using the estimated-polynomial-regression equation (or by linear extrapolation). Finally, these values were rounded to the nearest integer.

Results

Kolmogorov-Smirnov two-sample (KS) tests were used to test the difference between the P&P-ASVAB and equated CAT-ASVAB distribution functions. KS tests were run for all five subtests.

The total sample was first randomly divided into two groups, with the restriction that total group size was approximately equal and the number of females and males did not differ by more than one across the two groups. The next step computed expected number correct (ENC) scores from CAT-ASVAB thetas for each group. The linear transformation which transforms ENC scores to EENC scores was estimated for each group separately. Each transformation was then used to compute the CAT-EENC scores for that group.

Two comparisons were made: (1) a comparison of the CAT-EENC score distribution of Group A to the P&P-ASVAB score distributions of Group B and, (2) a comparison of the CAT-EENC score distribution of Group B to the P&P-ASVAB score distributions of Group A. Differences were tested using the KS statistic. The above procedure was repeated separately for the male, female, and combined samples.

Table 1 presents the results of the KS tests. Each comparison examines the difference between the CAT-EENC (or number correct for the NO subtest) distribution function and the corresponding P&P-ASVAB distribution function. Only two comparisons were significant at the .05 level: (1) the comparison of scores on NO for females, and (2) the comparison involving the AFQT composite for females.

Discussion

The results of the KS analysis indicate that the IRT true-score equating procedure provides similar distributions for the two versions of the ASVAB. Neither males or females appear to be placed at a disadvantage relative to their P&P-ASVAB scores.

References

Braun, H. I., and Holland, P. W. (1982). Observed-score test equating: a mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin

(Ed.), *Test Equating*. New York: Academic Press, 9-50.

Green, B. F., Bock, R. D., Linn, R. L., Lord, F. M., & Reckase, M. D. (1985). *A plan for scaling the Computerized Adaptive Armed Services Vocational Aptitude Battery (ASVAB)* (MPL TN 85-2). San Diego, CA: Manpower and Personnel Laboratory, Navy Personnel Research and Development Center.

Sympson, J. B. & Hartmann, L. (April, 1985). Item calibrations for Computerized Adaptive Testing (CAT) experimental item pools. In D. J. Weiss (Ed). *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Table 1
Kolmogorov-Smirnov Two-Sample Tests

		Comparison One		Comparison Two	
Sex	Subtest	Z	Probability	Z	Probability
Males	AR	.927	.36	.480	.98
	WK	.758	.61	.478	.98
n_1=72	PC	.677	.75	.541	.93
n_2=73	NO	.883	.42	.500	.96
	GS	.816	.52	.715	.69
	AFQT	.926	.36	.500	.96
Females	AR	1.178	.13	.478	.98
	WK	.970	.30	.904	.39
n_1=99	PC	.460	.98	.547	.93
n_2=100	NO	1.679	.01	.676	.75
	GS	1.078	.20	.901	.39
	AFQT	1.444	.03	.600	.86
Combined	AR	.737	.65	.517	.95
	WK	.884	.42	.948	.33
n_1=171	PC	.472	.98	.642	.81
n_2=173	NO	1.223	.10	.684	.74
	GS	.476	.98	.845	.47
	AFQT	.802	.54	.474	.98

Reducing the Predictability of Adaptive Item Sequences

C. Douglas Wetzel

Navy Personnel
Research & Development Center
San Diego, CA 92152

James R. McBride

The Psychological Corporation
San Diego, CA 92101

Previous research into the psychometric properties of Computerized Adaptive Testing (CAT) has shown that adaptive tests are much more efficient than conventional tests (e g , Weiss, 1974, 1982). Different methods of adaptive testing vary widely in efficiency. The most efficient are those in which test items are chosen one at a time, in a manner which optimizes some function of the difference between item difficulty and the current estimate of the examinee's ability However, many optimization CAT strategies yield a predictable sequence of test items early in testing that could lead to over exposure of items and possible compromise. This is because the possible sequences of test items form a binary tree The same item will always be chosen first: only two items can be chosen second, and so on. As a result, all examinees who answer the first several items the same way will encounter identical sequences of test items, making compromise easy and almost inevitable.

This repetition of predictable item sequences is illustrated in Figure 1 with plots of individual examinee ability estimates as a function of test length. The course of testing for each examinee is traced up the page, with all examinees starting at an ability estimate of zero prior to being given their first test question (i.e. item zero). It can be seen that after the first test item is administered, there are only two possible ability estimates from right or wrong responses, and after the second item there are just four possible ability estimates, and so forth. As the number of items increases, the common paths shown early in testing fan out into a number of unique ability estimates. Smaller changes in the ability estimates are found later in testing (items 10-15) where they "home in" on a region of the ability continuum and become more reliable.

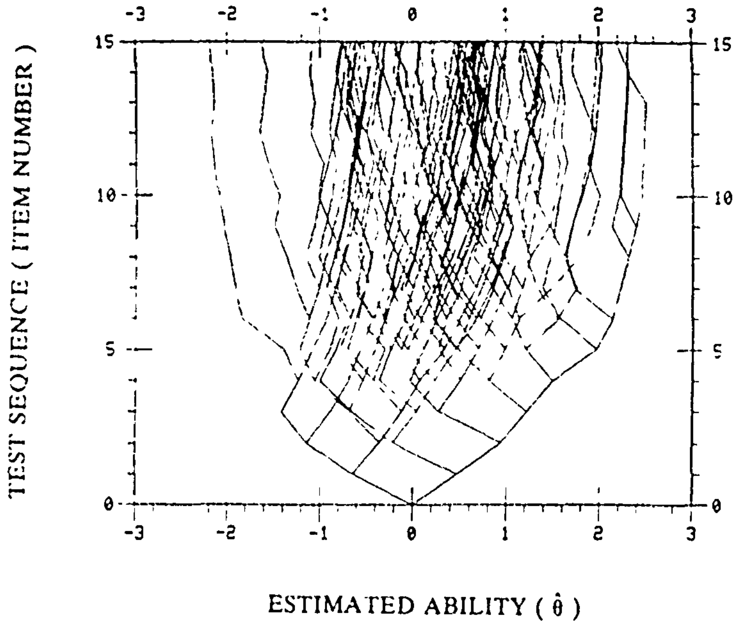


Figure 1. Item-by-item ability estimates ($\hat{\theta}$) for 75 Marine Corps recruits given a 15 item adaptive test using Owen's Bayesian strategy (Owen, 1969, 1975)

This situation suggests the need for a method which retains the efficiency of the mathematically optimal adaptive strategies, but one which eliminates the occurrence of predictable sequences of test items One method of avoiding predictable item sequences is to choose an item at random from among a set of nearly optimal items In the present work, a stratified maximum information (STMI) strategy was selected for extensive study of random selection from several good or informative items taken at different points in the sequence of items A question to address is whether this technique to reduce the repeated exposure of the very best items in the bank would reduce the psychometric quality of the resultant adaptive test, relative to several other adaptive and conventional test strategies

APPROACH

A two-stage computer simulation was used to investigate the effect of various item selection strategies on the psychometric characteristics of the resultant tests The first stage simulation generated item parameter estimates with typical error characteristics. This item bank was used for the second stage in simulated administrations of adaptive and conventional tests for normal and rectangular examinee true ability distributions.

Generating Fallible Item Parameter Estimates: A simulated item bank was created based on real item parameters representative of a "live" testing situation The item bank consisted of two parameters sets. the "true" parameters $\{ a , b , c \}$, and simulated estimates $\{ \hat{a} , \hat{b} , \hat{c} \}$ Unlike many synthetic item banks (e.g. Wetzel & McBride, 1983), real items written for live examinees often yield unique item banks without uniform distributions when broken down by each item parameter, and may have a distinct positive correlation between the $\hat{a}$-parameter and $\hat{b}$-parameter $r_{\hat{a}\hat{b}}$ (Sympson, Weiss and Ree: 1982). To achieve a test information curve that approximated that of real test items, a simulated bank was based on item parameters from real test items. Estimated item parameters from real items were used as true parameters for a simulation of the item calibration phase. The 200 real item parameters were obtained from J B. Sympson of NPRDC from his calibration of five ASVAB content areas word knowledge, arithmetic reasoning, paragraph comprehension, general science and, mathematical knowledge. These banks had each been calibrated with LOGIST (Wood, *et al*, 197C) on approximately 1500-2000 live examinees for entrance into the armed forces. In the present study, a stratified random sample from these five banks (970 total items, from individual banks of 180-210) was taken by randomly selecting 40 items from each of the five banks to yield a new combined total of 200 items These 3-parameter logistic estimates were then used *as if* they were 'true' item parameters in a simulation in which examinees were administered all 200 items Examinee item responses were simulated by using the 3-parameter logistic model (Birnbaum, 1968) to generate simulated binary responses to the test items using a probability sampling technique often employed for this purpose (Vale and Weiss, 1975). If a random number drawn from a uniform distribution on the interval (0,1) was less than the 3-parameter logistic model probability of a correct response $P(\theta)$, then the examinee was credited with a correct answer, otherwise an incorrect item response was specified The simulated correct and incorrect item responses were created for 1500 normally distributed simulated examinees for these 200 items and then calibrated for the present study with LOGIST These parameter estimates were then joined with the generating "true" parameters for simulations of the various test strategies studied here.

Examinee True Ability (θ) Distributions: Each test strategy simulation run was conducted twice with 1900 simulated examinees (1) once with a *Rectangular θ Distribution* of 19 groups (100 examinees each) 25 θ units apart over the the interval

-2.25 to +2.25 ($\bar{X}$ = 0.0, SD = 1.369), and (2) once with one group of 1900 from a standard *Normal θ Distribution* ($\bar{X}$ = 0.0001, SD = 0.999).

Simulation of Testing Strategies: A fixed test length of 15 items was selected as a representative number of items in which an adaptive test could possibly be implemented. For each test strategy, 1900 examinees were simulated from the two ability distributions. Examinees always responded to the true item parameters { a, b, c } on the basis of true ability (θ) according the 3-parameter logistic model. An item-response was scored correct by the probability sampling technique described above. Item selection was based on the estimated parameters { $\hat{a}$, $\hat{b}$, $\hat{c}$ }. Ability estimation in the two Bayesian tests (see below) used the same estimated item parameters as used for item selection. All adaptive tests began with an initial ability estimate ($\hat{\theta}$) of 0.0 and the two Bayesian tests assumed an initial normal prior distribution of ability with variance 1.0. Either Bayesian or maximum likelihood ability estimates were limited to the range -3.00 to +3.00. Each testing strategy is described below.

Owen's Sequential Bayesian Test. This adaptive strategy chooses an item to minimize the expected value of the posterior variance of the ability distribution (Owen, 1969, 1975). The ability estimate (distribution) is updated after each item and the parameters of the Bayes posterior distribution are used as parameters of the prior distribution for the next item.

Stratified Maximum Information Test - Bayesian Scoring (STMI-B). This adaptive test selects the item with approximately the greatest item information ($I(\hat{\theta})$) at the current Bayesian ability estimate $\hat{\theta}$. Items are selected from a prearranged "information table" consisting of stratified lists of information values calculated for fixed θ levels. Each list of the information table contains information values arranged in descending order of the values of their information functions at the midpoints of a series of narrow intervals of ability. In this study, 36 lists in .125 wide θ increments spanned the ability range from -2.25 to +2.25. The ability estimate was updated after each item with the same Bayesian ability estimation procedure employed in the Owen's test above. This STMI-B strategy is a "hybrid" (Wetzel & McBride, 1983) between previous strategies in that the same information table method is employed, but Bayesian ability scoring is used instead of maximum likelihood scoring (Sympson, Weiss & Ree, 1982).

Seven versions of the STMI-B test were simulated to investigate probabilistic item sequences produced by randomizing the choice among items. The top k consecutive items with greatest information in an information table's list were selected and held in a temporary vector. One of these k items was then selected randomly, with each of the items having equal $1/k$ selection probabilities. This randomization among the most informative items within a given list of an information table occurred in the list closest to the current ability estimate. Two types of randomization conditions were studied, which differed in whether the $1/k$ probability was constant for all 15 test items or varied as a function of the order of item administration.

Constant Selection Ratios for All Items Administered

(a) 1:4 1:4 (b) 1:5 1:5 (c) 1:10 1:10 (d) 1:20 1:20 (e) 1:40 1:40

Selection Ratios Adjusted According to Test Length

(f) 1:5 1:4 1:3 1:2 1:1 1:1 (g) 1:10 1:8 1:6 1:4 1:2 1:2

The constant selection ratios remained the same for each of 15 items administered (a-e). The first adjusting selection ratio strategy (f) selected the first item in the test from the best five available items in the current information table list (1:5), the second item from four (1:4), the third from three (1:3), the fourth from only two (1:2), and then the fifth through the fifteenth items were each the single most informative item available (1 1) The second adjusting ratio strategy (g) used the first four ratios shown for the first four items and then used a ratio of 1 2 for the fifth through the fifteenth items These adjusting ratio strategies randomize more for early items since they are most subject to compromise Randomization decreases as the test proceeds so the strategy will select more appropriate items when the ability estimate is closest to it's terminal value

Maximum Information Full Search - Bayesian Scoring (MI-B). Each item is selected by a search of the entire bank for the item with the greatest item information at the exact value of the current estimated ability ($\hat{\theta}$) This exhaustive item-by-item search is made by actually calculating item information for $\hat{\theta}$ "on-line" each time an item is to be selected during test administration This strategy was created to assess any effect of granularity in the STMI-B test, where the θ continuum was divided into discrete increments in rounding $\hat{\theta}$ to the nearest .125 midpoint. Previous studies using the MI strategy have employed maximum likelihood scoring rather than the present Bayesian scoring (e g , McKinley & Reckase, 1981, Weiss, 1982).

Stratified Maximum Information Test - Maximum Likelihood Scoring (STMI-ML). This strategy employs Bayesian ability estimation until at least one correct and one incorrect item response have been obtained and, then uses maximum likelihood estimation for the remainder of the items in the test. Items were selected from an information table calculated over the same .125 wide θ increments used for the STMI-B strategy above This strategy was first used by Sympson, et. al. (1982).

Weiss's Stradaptive Test. This mechanical adaptive strategy uses a pre-sorted item bank divided into 'strata' on the basis of the $\hat{b}$-parameter and then arranges items within each stratum in descending order of the values of their $\hat{a}$-parameters (cf. Weiss, 1974). There were nine strata in this study, each 0.5 ability units wide, over the range -2.25 to +2.25. Items were selected from the top of the stack in each stratum. Branching to another stratum occurred after each item, branching up one stratum after a correct response, and down one stratum after an incorrect response.

Peaked Conventional Test This conventional test was designed by selecting the 15 items with the greatest values of information at the central ability value of 0.0. All simulated examinees were administered this same set of 15 items.

Flat Conventional Test: This test was created by selecting the item with the greatest item information value at each of 15 equally distributed ability points over the interval -2.0 to +2.0 All simulated examinees were administered this same set of 15 items, with item difficulty increasing as the test proceeded.

RESULTS

Fidelity: The correlation of fidelity between true and estimated ability ($r_{\theta\hat{\theta}}$) for the simulated test strategies was based on the normal θ distribution consisting of a single group of 1900 examinees All coefficients were based on Bayesian ability estimates, except the STMI-ML strategy which used maximum likelihood ability estimation. All the optimization strategies yield similar fidelities (Owen's Bayesian {.951}, STMI-ML {.950}, STMI-B {.953}, MI-B {.956}), with the Stradaptive test {.935} performing slightly better than the conventional tests The peaked conventional test { 893} is lower than the flat conventional { 922} , since the peaked test does not span the extremes of ability As the denominator of the fixed selection ratio increases, the STMI-B test shows a small decrease in fidelity {.953, 955, 952, 951, .936}, but at 1/40 it never falls below the conventional tests The adjusting ratio STMI-B strategies yield fidelities { both 957

} comparable to the 1/1 and 1/5 fixed ratios and to the MI-B full search strategy.

Average Test Information: Test information was employed as an index of precision, or of how well a set of items discriminates an ability level from nearby ability levels, the reciprocal of the square root of information is inversely related to the standard error of an ability estimate (Lord, 1980) It was used here as a measure of the appropriateness of the set of items administered for a given true ability level. "Test information" $\{\sum I(\theta)\}$ is the sum of the individual "item information" $\{I(\theta)\}$ values for the items administered in an individual examinee's test. The true item parameters and true examinee ability were used to compute these test information values, which were then averaged over the 100 examinees at each of the 19 levels of true ability. Figure 2 shows the obtained test information (Lord, 1980) averaged over the 100 examinees in each of the 19 rectangular θ distribution groups for each test. A peak is shown around a $\theta \approx$ 1.25 which is owed to the correlation of the $\hat{a}$- and $\hat{b}$-parameters (r_{ab} = .577, r_{ac} = .059, r_{bc} = -.131) The magnitude of the $r_{\hat{a}b}$ correlation was very similar to that obtained by Sympson, et. al (1982).

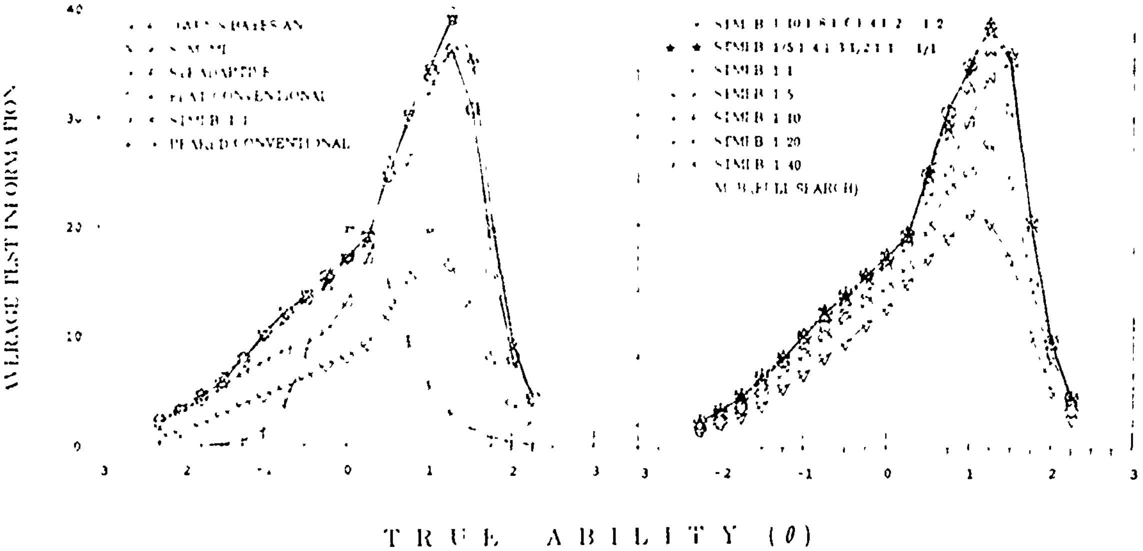


Figure 2. Average test information for test strategies over 19 true ability (θ) groups.

The left-hand panel of Figure 7 shows the adaptive tests to yield more test information than the two conventional tests over a wide range of θ, excepting the peaked conventional test at the narrow region around $\theta \approx 0$. The conventional tests represent one extreme of mismatch, where the same fixed set of items are presented regardless of the location of the examinee on the ability continuum. The Stradaptive test generally yielded less test information than the other adaptive tests at lower ability levels because it used a mechanical strategy that did not correct for guessing. The 1/1 STMI-B and STMI-ML tests yielded the best test information overall and were practically equivalent to the Owen's Bayesian test

The right-hand panel of Figure 2 shows eight versions of the maximum information strategy using Bayesian scoring For the five constant ratio STMI-B strategies { 1/1, 1/5, 1/10, 1/20 & 1/40 }, test information is degraded monotonically as the denominator of the ratio increases, i.e., the selection set included items farther down in the information table which had somewhat lower discriminations, higher guessing and more inappropriate difficulties As the ratio changes from 1/1 to 1/10, the amount of obtained change becomes larger for each doubling of the denominator. Substantial reductions were produced with the extreme 1/20 and 1/40 ratios, and a small but acceptable degradation in test precision resulted when 1/5 available items was randomly selected throughout the test.

The remaining tests shown in the right-hand figure panel all yielded test information that was approximately the same as the STMI-B 1/1 condition and in excess of the STMI-B 1/5 test. First, the MI-B condition achieved no more test information than any other condition, indicating that the .125 wide increments on which the STMI information table tests were based were small enough to closely approximate this full search condition. Second, the two STMI-B adjusting selection ratio strategies yielded test information approximately equivalent to the STMI-B 1/1 and MI-B conditions which used no randomization at all.

CONCLUSIONS

This work suggested the following conclusions: (1) The STMI strategy, with Bayesian ability estimation seems to work about as well as the best adaptive testing strategies; (2) Predictable sequences of test items can be avoided by modifying STMI so that items are selected at random from a nearly optimal set of items, (3) As long as that set is small in number, the adaptive test will not lose an appreciable amount of efficiency, (4) If the set is small to begin with, and gets progressively even smaller (through specification of a shrinking set size) the adaptive test is virtually as efficient as the strategy which chooses the optimal item every time.

REFERENCES

Birnbaum, A. (1968) Some latent-trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley.

Lord, F.M. (1980) *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, New Jersey: Erlbaum.

McKinley, R.L. and Reckase, M.D. (1981) *A comparison of a Bayesian and a maximum likelihood tailored testing procedure* (Research Report 81-2). Columbia, MO. University of Missouri, Educational Psychology Department.

Owen, R.J. (1969) *A Bayesian approach to tailored testing* (Research Bulletin 69-92). Princeton, NJ: Educational Testing Service.

Owen, R.J. (1975) A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351-356.

Sympson, J.B., Weiss, D.J. & Ree, M.J. (1982) Predictive validity of conventional and adaptive tests in an Air Force training environment. (Technical Report 81-40) Brooks Air Force Base, Texas: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Vale, C.D. and Weiss, D.J. (1975) *A simulation study of stradaptive ability testing* (Research Report 75-6). Minneapolis: University of Minnesota, Department of Psychology.

Weiss, D.J. (1974) *Strategies of adaptive ability measurement* (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology.

Weiss, D.J. (1982) Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

Wetzel, C.D. and McBride, J.R. (1983) *Influence of fallible item parameters on test information during adaptive testing*. (NPRDC TR 83-15). San Diego, CA. Navy Personnel Research and Development Center.

Wood, R.L., Wingersky, M.S., & Lord, F.M. (1976) *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters* (RM-76-6). Princeton, NJ: Educational Testing Service.

Speeded Tcsts - Can Computers Improve Measurement?

by
John H. Wolfe
Navy Personnel Research and Development Center
San Diego, California 92152-6800

INTRODUCTION

The Armed Services Vocational Aptitude Battery (ASVAB) contains two speeded tests, Coding Speed(CS) and Numerical Operations (NO). In paper-and pencil mode, these tests are administered with a fixed time limit and scored by counting the number of correct responses. Computerized administration of the same items offers several interesting alternative methods of scoring. For example, one can administer the tests with no time limit, so that everyone finishes the test, and then measure the time each examinee used. Greaud and Green (1985) showed that scoring such a test with a "rate" measure (equal to the number of correct responses divided by the test time) increased overall test reliability. Computer scoring by rates has two advantages over ordinary administration: (1) there is no "ceiling" effect for the fastest examinees, and (2) the scores for the slowest examinees are based on the same number of items as the fastest examinees, and therefore have improved reliability.

Further improvements in reliability can be expected from measuring the examinee's response times for each item, and combining these times into an appropriate total score or scores. As a starting point for proposing alternative scoring functions, consider the Greaud and Green "rate":

$$rate = \frac{number\ correct}{Total\ Time}. \qquad (1)$$

By dividing the numerator and denominator by the total number of items, N, the formula is seen to be equivalent to:

$$rate = \frac{P_c}{\frac{1}{N}\sum_{i=1}^{N} T_i} = \frac{P_c}{\bar{T}}, \qquad (2)$$

where P_c is the proportion of correct responses and $\bar{T}$ is the sample mean of the item response times, taken over all of the items.

The first method of improving on the formula might be to compute the mean time for only the correct responses. If two examinees take the test, and one of them is able to answer an item wrong twice as fast as the other examinee answers it wrong, it is not clear that the first examinee should be scored higher on the test. It seems plausible that incorrect responses should be eliminated from the scoring.

The second modification to the formula would eliminate "outliers" from the computation of the mean item response time. It is not uncommon for an examinee taking a speeded test to be distracted or pause to ask a question of the proctor. Extraordinarily long response times should be identified and omitted from the scoring.

One potential problem with computing the sample mean of the item times is that the distribution of times is highly skewed. In general, the sample mean of a skewed distribution is not necessarily a good estimate of the population mean. Thus, a third approach to improving reliability of scoring would be to seek some transformation of the times that makes their distribution more nearly normal. Some possibilities that have been suggested in the literature are $\log T_i$, $\sqrt{T_i}$, and $\frac{1}{T_i}$. To these can be added $\frac{1}{\sqrt{T_i}}$. For each transformation, one can construct a corresponding "rate" measure by computing the mean of the transformed times and then transforming the mean back onto the time scale to get a $\hat{T}$ to replace $\bar{T}$ in equation (2). Thus,

$$rate_{\log T} = \frac{P_c}{\exp\left[\frac{1}{N}\sum_{i=1}^{N}\log T_i\right]}. \tag{3}$$

$$rate_{\frac{1}{T}} = P_c\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\right] \tag{4}$$

$$rate_{\sqrt{T}} = \frac{P_c}{\left[\frac{1}{N}\sum_{i=1}^{N}\sqrt{T_i}\right]^2}. \tag{5}$$

$$rate_{\frac{1}{\sqrt{T}}} = P_c\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sqrt{T_i}}\right]^2. \tag{6}$$

METHOD

Eighty-five randomly chosen recruits at the Recruit Training Center, San Diego, were administered two alternate forms of Coding Speed and two alternate forms of Numerical Operations presented on Apple /// personal computers. The order of testing was randomized. The CS tests contained 12 sets of seven items per screen, and the NO tests contained 17 sets of three items per screen. The onset of a new screen was synchronized to the computer so that stimulus timing began when the electron beam was in the upper left corner of the monitor. Timing of key presses was measured with a hardware clock with millisecond accuracy. Pascal software that was used to detect key presses introduced constant errors on the order of a tenth of a second. Each test was administered with a 16-minute time limit, which is more than twice the usual time. This was sufficient for most, but not all subjects to complete the tests.

Outliers were defined as times whose logarithms were more than three standard deviations above the mean log time for that examinee. Reliabilities were computed for all responses, correct responses only, and correct responses trimmed for outliers.

Skewness and kurtosis statistics were computed for individual item times in Form A of the Coding Speed test, along with four transformations of the times. For all tests, means of the times and their transformations were computed, and five different "rate" scores were obtained. Skewness and alternate form reliabilities were computed for the five means and the five rate scores.

RESULTS

Table 1 shows the alternate form reliabilities of the Greaud and Green rate scores when all item times are scored, when only correct responses are scored, and when correct responses are trimmed at the upper tail for outliers. Restricting the scoring to correct responses appeared to have no effect on reliability, but trimming outliers raised reliability somewhat.

Table 1
Effect of Trimming Incorrect Responses and Outliers On Reliability of Rate Scores

	Coding Speed	Numerical Operations
All Responses	.81	.72
Correct Responses	.80	.73
Trimmed Correct Responses	.83	.75

Table 2 summarizes the skewness and kurtosis characteristics of 84 Form A Coding Speed item times and their transformations. As expected, the times were quite skewed, and in addition, had considerable kurtosis. Taking logs of the times eliminated the skewness, and also reduced kurtosis. The reciprocal transformation made matters much worse. The square root transformation reduced skewness, but not as much as the logarithms. Taking the reciprocal of the square roots made skewness and kurtosis worse. From these data, it appears that the log transformation is best, and that the reciprocal should not be used to normalize the data.

Table 2
Mean Skewness and Kurtosis of 84 Coding Speed Item Times and Their Transformations

	Skewness	Kurtosis
T_i	0.99	1.47
$\log T_i$	-0.03	0.84
$1/T_i$	1.27	5.50
$\sqrt{T_i}$	0.51	0.62
$1/\sqrt{T_i}$	0.64	2.55

Table 3 shows the skewness of five different means and five "rates" based on these means. Although the central limit theorem merely implies that the means of the times should be normally distributed within each individual, it is still somewhat surprising that the mean times, reciprocals, and square roots are significantly skewed across individuals. Again, the log transformation showed the least skewness. None of the rate measures were skewed in the Coding Speed tests, and all of the rates were significantly skewed in the Numerical Operations tests.

Table 3
Skewness of Alternative Scoring Formulas

	Coding Speed		Numerical Operations	
	Form A	Form B	Form A	Form B
Mean T_i	0.84*	1.46*	0.93*	0.52*
Mean $\log T_i$	0.14	0.31	0.16	-0.15
Mean $1/T_i$	0.64*	0.41	0.56*	0.77*
Mean $\sqrt{T_i}$	0.48	0.81*	0.54*	0.17
Mean $1/\sqrt{T_i}$	0.21	0.08	0.20	0.46
$P_c/(Mean\ T_i)$	-0.18	0.07	0.52*	0.61*
$P_c/Exp(Mean\ \log T_i)$	-0.13	0.05	0.56*	0.66*
$P_c(Mean\ 1/T_i)$	-0.00	0.03	0.60*	0.72*
$P_c/(Mean\ \sqrt{T_i})^2$	-0.17	0.06	0.54*	0.64*
$P_c(Mean\ 1/\sqrt{T_i})^2$	-0.08	0.04	0.58*	0.69*

* Significant at $p < .05$ by 2-tailed t-test.

Table 4 is an expanded version of Table 1, in which mean times, mean log times, and rates based on mean log times are also shown. For all measures, it appears that restricting scoring to correct responses did not improve reliability. Eliminating outliers did improve reliability. The log transformation improved reliability if outliers were not trimmed, but not otherwise. Trimming outliers improved reliability if untransformed times were used, but not otherwise. Rate scores were more reliable than average times for Coding Speed, but not for Numerical Operations.

Table 4
Reliabilities of Alternative Scoring Functions

	Coding Speed	Numerical Operations
	All Responses	
Mean Time	.76	.74
Mean Log Time	.79	.75
P_c/Mean Time	.81	.72
P_c/Exp(Mean Log Time)	.82	.75
	Untrimmed Correct Responses	
Mean Time	.74	.74
Mean Log Time	.78	.76
P_c/Mean Time	.80	.73
P_c/Exp(Mean Log Time)	.81	.75
	Trimmed Correct Responses	
Mean Time	.77	.76
Mean Log Time	.78	.76
P_c/Mean Time	.83	.75
P_c/Exp(Mean Log Time)	.82	.76

DISCUSSION

The results presented in this paper are only preliminary: more subjects remain to be tested, and additional analyses need to be performed, particularly on the Numerical Operations items. Nevertheless, certain conclusions and directions for future work stand out:

One successful method for improving reliability is to eliminate outliers. Future work should explore this method in more detail. The optimal cutting point for trimming the data is a question that needs to be answered empirically.

Another method that improved reliability as much as trimming outliers was the log transformation of times. So far, there is no indication that both methods combined are better than one of them. However, combining both methods does not decrease reliability, and may inspire greater confidence. In the end, considerations of computational speed and program complexity may be the determining factors in deciding which method(s) to use.

In all of the work described here, it has been implicitly assumed that the items are homogeneous in difficulty, and only individual differences have been examined. The next step in the research should be to examine items, and to develop a model that encompasses both item characteristics and individual differences in ability. This approach should be especially useful in Numerical Operations, where addition, subtraction, multiplication, and division have quite different average response times.

REFERENCE NOTES

Greaud, V.A. & Green, B.G. (1985) *Equivalence of conventional and computer presentation of speed tests.* Unpublished manuscript.

MEDIUM OF ADMINISTRATION EFFECTS ON ATTITUDES TOWARD ASVAB TESTING

William F. Kieckhaefer
RGI, Incorporated

Daniel O. Segall
Kathleen E. Moreno
Navy Personnel R&D Center

BACKGROUND

Computerized adaptive testing is being considered for use in military selection and classification. One research question under investigation concerns whether medium of administration affects examinee attitudes toward the current Armed Services Vocational Aptitude Battery (P&P-ASVAB). Recognizing the importance of this, the Defense Advisory Committee on Milit?·y Personnel Testing recommended that the reactions of examinees to the computerized adaptive version of that test (CAT-ASVAB) should be systematically collected and analyzed (Linn, Bond, Britell, Campbell, Jaeger, Novick, and Uhlaner, 1983).

Early researchers reported on particular aspects of the relation between attitudes and computerized testing. For example, Hedl, O'Neil, and Hansen (1973) showed that less favorable reactions of the subjects to a computerized test were due to a lack of clarity in the instructions and unfamiliarity with computer terminals. In a related study, Walther and O'Neil (1974) found that subjects with greater test anxiety or negative attitudes toward computers performed more slowly and made more errors on a test than subjects with lower levels of test anxiety. More recently, Nilles, Carlson, Gray, Hayes, Holmen, and White (1980) supported the view that computer usage generates anxiety which negatively impacts on user attitudes and performance.

Other researchers present a different view. Schmidt, Urry, and Gugel (1978) found that examinees completing a computerized adaptive test had positive attitudes toward it. In a study on computer-managed instruction, Robinson, Tomblin, and Houston (1982) also reported positive user attitudes.

More specifically for ASVAB testing, Mitchell, Hardwicke, Segall, and Vicino (1983) reported generally positive attitudes of male Navy recruits toward taking the CAT-ASVAB. Some attitudes corresponded with subjects' level of experience with computers or keyboards. For example, those with "little to none" computer experience were more likely to indicate that computerized testing was more impersonal than paper-and-pencil testing. Subjects with "little to none" keyboard experience were generally more likely to express uneasiness about taking a test on a computer and to indicate that the computerized test was more difficult. This research showed that while computerized testing may be considered impersonal by some, this perception does not imply a negative attitude toward computerized tests. These findings were supported in the service-wide study of Hardwicke and Yoes (1984).

PURPOSE

Investigations of user attitudes toward computerized testing can benefit from clearer attitudinal items and improved experimental designs. Furthermore, little is known about the relation between attitudes toward ASVAB testing and examinee ability or such background variables as race and sex. The purpose of this study was to assess the effects of the medium of administration on attitudes toward ASVAB testing. While some research does exist in this area, this research is still exploratory. Therefore, the research objectives were to:

(1) Examine the effect of medium of administration on examinee attitudes toward testing; and

(2) Determine if this relation differs as a function of sex, race, or ability.

METHOD

This is part of a larger study investigating several aspects of subgroup differences in ASVAB testing. While a total of 3,094 Navy recruits were tested under the larger study, only the data for 619 recruits tested at the Recruit Training Center in Orlando, Florida were available at the time of this presentation. Therefore, findings presented here are preliminary. The sample was 60.9% male and 39.1% female. Also, 49.4% were White, 18.7% were Hispanic, and 31.7% were Black.

Test proctors selected all recruit companies of men and women which began training between April and June of 1985. Then proctors randomly selected recruits from a company for testing and randomly assigned them to complete either the P&P-ASVAB or the CAT-ASVAB first These were the Medium of Administration conditions. Test proctors collected the background information regarding examinee sex and race from each subject's enlistment form (DD form 1966) They also obtained Pre-enlistment scores on the Armed Forces Qualification Test (AFQT) from that form.

After completing the first version of the ASVAB (i.e., P&P or CAT), subjects responded on a seven-point scale to nine questions presented in a paper-and-pencil format. These were the dependent variables in the present study. While recruits did complete the other ASVAB version also, data from the second version are not relevant to this attitude study and are not reported here.

RESULTS

Responses to each of the nine attitude questions served as the dependent variables in a three-factor analysis of variance: Medium by Sex by Race. The significance of the AFQT covariate was obtained for each dependent variable. An analysis of covariance was performed for items with significant AFQT covariates.

For each attitude question, Table 1 shows the number of respondents (N), the overall mean, and the cell means for significant main effects in the analysis of variance. Figure 1 depicts the significant interaction effects involving medium. Table 2 shows each attitude question and the distribution of responses. For significant medium effects, the distribution is shown for each group.

Only question 4 had a significant-AFQT covariate. When entered into an analysis of covariance, two of the previously significant effects were no longer significant: the main effect of Race and the interaction effect of Sex by Race.

DISCUSSION

While these are preliminary findings, six of the nine attitude questions had significant main effects of Medium--all favoring the CAT-ASVAB. Those taking the CAT-ASVAB felt better about taking the test battery. In addition, they felt less tired, they experienced less eye strain, and they thought the test battery was shorter. These results are probably due to the fewer number of items required by the CAT ASVAB. In general, the CAT-ASVAB measures the same content areas as the P&P-ASVAB with about half the number of items.

Furthermore, those taking the CAT-ASVAB felt more relaxed during the test. Perhaps this is because they proceeded at their own pace while those taking the P&P-ASVAB had specified time limits. Also, those taking the CAT-ASVAB reported that the instructions were clearer. This may be attributable to the interactive nature of the instructions, which included immediate feedback during a keyboard familiarization sequence and sample questions.

Three questions showed no main effects of medium. These indicated no differences between administration conditions on anxiety, perceived difficulty of the questions, and perceived fairness.

Table 1

Cell Means for Main Effects

Item	Attitude	N	Overall	Medium[a]	Sex[b]	Race[c]
1	Overall feelings	597	2.7	2.4 3.0		
2	Fatigue	596	3.5	4.0 3.1		
3	Anxiety	595	3.5			
4	Question Difficulty	594	3.9		3.8 4.0	4.0[d] 3.8 3.8
5	Fairness	595	2.2			
6	Test Length	593	4.0	3.4 4.5		
7	Pressure	593	3.0	2.7 3.2		
8	Eye Fatigue	593	3.3	3.4 3.2		3.5 3.2 3.2
9	Instruction Clarity	593	1.5	1.4 1.6	1.6 1.4	

[a] The order of cell means is CAT-ASVAB, P&P-ASVAB.
[b] The order of cell means is Male, Female.
[c] The order of cell means is Black, Hispanic, White.
[d] This effect was not significant when AFQT was a covariate.

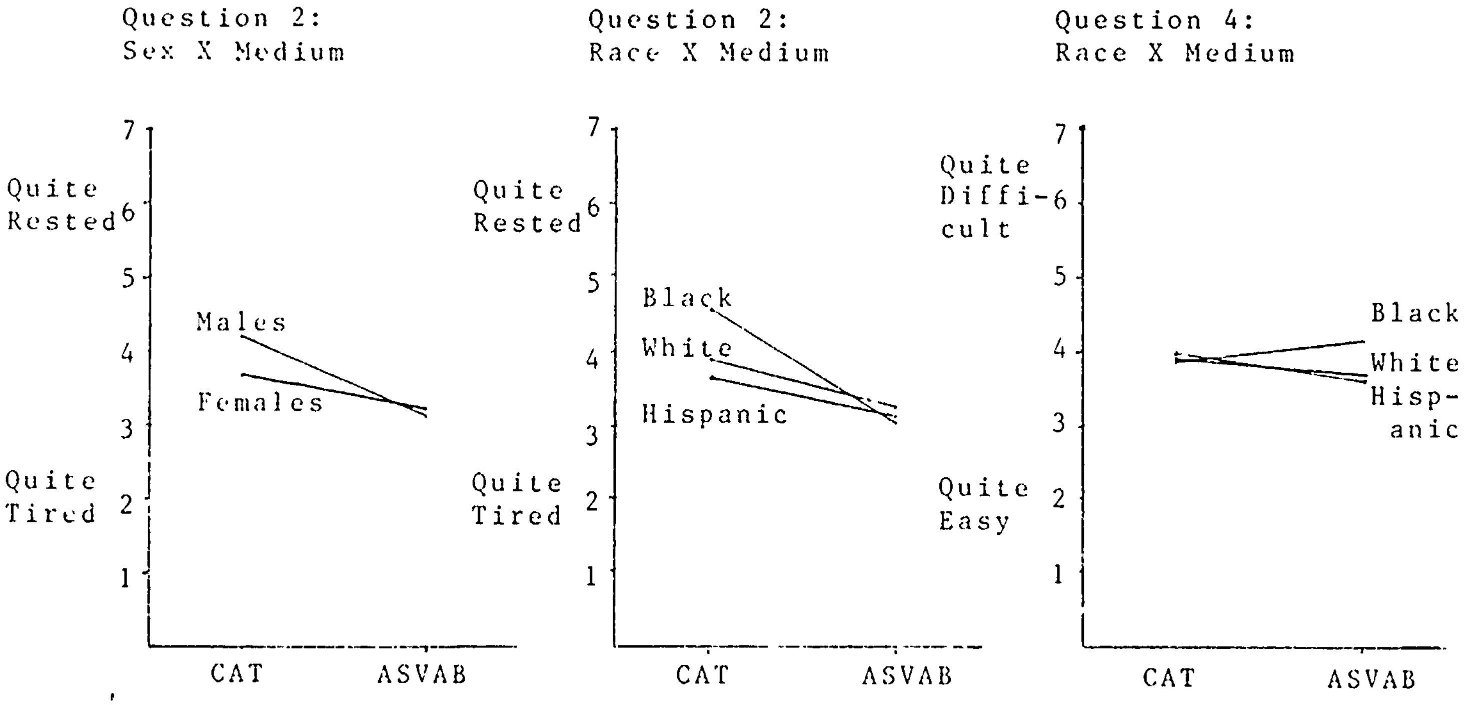


Figure 1. Interaction effects with Medium.

Table 2
Distribution of Responses to Attitude Questions

1. Overall, how did you feel about taking the test battery?

	1	2	3	4	5	6	7
	extremely	*quite*	*slightly*	*neither*	*slightly*	*quite*	*extremely*
	good	*good*	*good*		*bad*	*bad*	*bad*
CAT-ASVAB	16.4%	42.7%	26.3%	9.5%	4.7%	0.4%	0.0%
P&P-ASVAB:	5.6%	35.6%	26.9%	24.1%	5.3%	2.2%	0.3%

2. Overall, how tired did you feel at the end of the test?

	1	2	3	4	5	6	7
	extremely	*quite*	*slightly*	*neither*	*slightly*	*quite*	*extremely*
	tired	*tired*	*tired*		*rested*	*rested*	*rested*
CAT-ASVAB:	2.2%	8.4%	35.0%	22.6%	10.2%	16.8%	4.7%
P&P ASVAB.	5.3%	19.9%	47.2%	18.0%	4.3%	4.3%	0.9%

3. During the test, how anxious did you feel?

	1	2	3	4	5	6	7
	extremely	*quite*	*slightly*	*neither*	*slightly*	*quite*	*extremely*
	calm	*calm*	*calm*		*anxious*	*anxious*	*anxious*
Overall:[a]	8.4%	29.6%	11.6%	17.5%	25.0%	6.2%	1.7%

4. What is your opinion of the difficulty of the questions?

	1	2	3	4	5	6	7
	extremely	*quite*	*slightly*	*neither*	*slightly*	*quite*	*extremely*
	easy	*easy*	*easy*		*difficult*	*difficult*	*difficult*
Overall:[a]	1.0%	14.5%	20.4%	28.8%	31.8%	3.5%	0.0%

5. How fair do you feel the test was?

	1	2	3	4	5	6	7
	extremely	*quite*	*slightly*	*neither*	*slightly*	*quite*	*extremely*
	fair	*fair*	*fair*		*unfair*	*unfair*	*unfair*
Overall:[a]	23.9%	55.1%	7.7%	9.6%	2.4%	0.8%	0.5%

[a] For comparisons with non-significant medium effects, the distribution of responses is shown for the combined CAT-ASVAB and P&P-ASVAB groups

Table 2

(continued)

6. What is your opinion of the length of the test battery?

	1	2	3	4	5	6	7
	extremely short	*quite short*	*slightly short*	*neither*	*slightly long*	*quite long*	*extremely long*
CAT-ASVAB:	6.6%	17.9%	21.2%	38.3%	13.9%	1.8%	0.4%
P&P-ASVAB:	0.9%	5.0%	10.0%	33.8%	30.0%	13.8%	6.6%

7. During the test, how relaxed or pressured did you feel?

	1	2	3	4	5	6	7
	extremely relaxed	*quite relaxed*	*slightly relaxed*	*neither*	*slightly pressured*	*quite pressured*	*extremely pressured*
CAT-ASVAB:	21.2%	35.0%	17.2%	12.0%	12.8%	1.8%	0.0%
P&P-ASVAB:	10.3%	29.8%	17.6%	16.9%	20.7%	3.8%	0.9%

8. During the test, how strained or tired did your eyes feel?

	1	2	3	4	5	6	7
	extremely tired	*quite tired*	*slightly tired*	*neither*	*slightly rested*	*quite rested*	*extremely rested*
CAT-ASVAB:	7.7%	17.2%	37.6%	17.5%	5.8%	9.1%	5.1%
P&P-ASVAB:	8.2%	22.9%	33.9%	21.0%	5.3%	6.9%	1.9%

9. How clear do you feel the instructions were?

	1	2	3	4	5	6	7
	extremely clear	*quite clear*	*slightly clear*	*neither*	*slightly confusing*	*quite confusing*	*extremely confusing*
CAT-ASVAB:	68.2%	29.2%	0.7%	0.7%	1.1%	0.0%	0.0%
P&P-ASVAB:	53.0%	41.7%	1.6%	1.6%	0.9%	0.6%	0.6%

The other main and interaction effects are not so readily interpreted. A cultural bias hypothesis would favor White males over all other subgroups. Since this result occurred for only four of the remaining nine main and interaction effects, these results do not support a cultural bias hypothesis.

When all the data are ready, analyses will include measures of computer knowledge and test performance. Then, studies will investigate the relations between attitudes, computer knowledge, biographical characteristics, ability, and test performance.

References

Hardwicke, S. B., & Yoes, M. E. (1984). Attitudes and performance on computerized vs. paper-and-pencil tests. Paper presented at the Second Annual Air Force Conference on Technology in Training and Education, Wichita Falls, Texas.

Hedl, J. J., O'Neil, H. F., & Hansen, D. N. (1973). Affective reactions toward computer-based intelligence testing. *Journal of Consulting and Clinical Psychology, 40,* 217-222.

Linn, R. L., Bond, L., Britell, J. K., Campbell, J. P., Jaeger, R. M, Novick, M. R., & Uhlaner, J. E. (1983). *Report of the March 3 & 4, 1983 Meeting of the Defense Advisory Committee on Military Personnel Testing.* San Diego, CA.

[illegible] A., Hardwicke, S. B., Segall, D. O., & Vicino, F. L. (1983). Computerized adaptive testing: A preliminary study of user acceptability. *Proceedings of the 25th Annual Conference of the Military Testing Association.*

Nilles, J., Carlson, F. R., Gray, P., Hayes. J., Holmen, M., & White, M. J. (1980). *Technology assessment of personal computers.* Los Angeles, CA: University of Southern California, Center for Future Research.

Robinson, C. A., Tomblin, E. A., & Houston, A. (1981). *Computer-managed instruction in Navy technical training. An attitudinal survey* (Report No. 82-19). San Diego, CA: Navy Personnel Research and Development Center.

Schmidt, F. L., Urry, V. W., & Gugel, J. F. (1978). Computer-assisted tailored testing: Examinees reactions and evaluations. *Educational and Psychological Measurement, 38,* 265-273.

Walther, G. H., & O'Neil, H. F. Jr. (1974). On-line user-computer interface--The effects of interface flexibility, terminal type, and experience on performance. *AFIPS Conference Proceedings, 43,* 379-384.

STATISTICAL PROCESS CONTROLS AS AN ENHANCEMENT TO JOB DESIGN

Steven L. Dockstader
Navy Personnel Research and Development Center

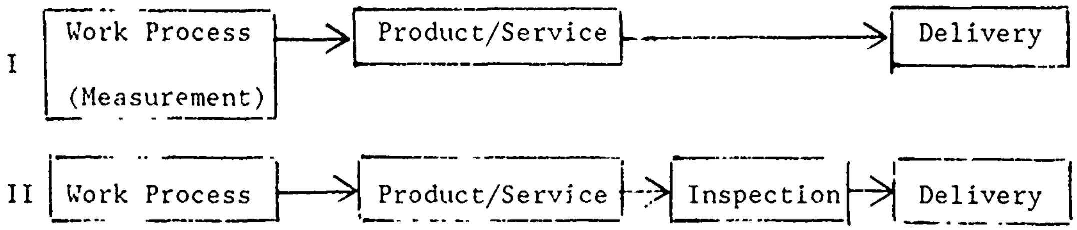
Abstract

The paper presents an argument for adopting statistical process control as an approach for improving productivity in organizations. The fundamental problems of resitance to change are discussed and job changes resulting from SPC are logically analyzed in terms of job characteristics theory. It is concluded that performance increases can be expected from the approach, but whether they are on the ability or motivational dimension is unclear and should be the subject of future research

Introduction

The purpose of this paper is to attempt to provide an explanation for the varying degrees of success found for projects seeking to improve productivity through changes in quality controls. It is the contention of this author that productivity derived from quality improvement is the result of (a) leadership initiatives which support a change in the methods of quality control and (b) factors intrinsic to the methods which contribute to both their acceptance and incentive value to the individual.

There are two basic approaches used to achieve product or service quality. The methods vary on the extent to which the responsibility for quality is vested in the producer of the product or service or with an external agent, such as a quality checker. The methods can be described schematically as follows:



Approach I will be referred to as the process control approach, while approach II will be referred to as the product inspection approach. Proponents of process control contend that it leads to higher levels of quality and lower costs than the product inspection approach. Sherkenbach (1984) has argued that greater efficiencies are achieved by the process control approach because the methods, over time, eliminate the problems in the (production) system which result in defects. This, in turn eliminates waste, rework, and the need for inspectors. Juran (1974), in a massive review of process control work, has pointed out that systems problems account for the vast majority of quality defects--as high as 85%--and that the systems problems can best be dealt with by the process control approach.

Problem

While the arguments of Juran and Sherkenbach have both logical and empirical support--most dramatically in Japanese industry--the question remains as to why such a successful approach has not achieved wide acceptance within the United States where the methods were initially developed. Tribus (1984) has argued that the management philosophy of most American business and industry is such that there is resistance to turn over control systems to the workforce, i.e., that quality is properly the responsibility of management and can not and should not be deployed to the workforce. He then goes on to point out that there has been no economic reason to change from the current system because U.S. manufactured goods have dominated world markets since WWII and costs associated with the product inspection approach to quality control were of no consequence in a noncompetitive market.

Recent changes in world market conditions have forced many companies to reexamine their approach to quality (Houston et al., 1985). Being less driven by profits and competition, public sector organizations can be expected to resist the change to the more effective process control approach. From this, it would appear that a substantial amount of education of management and concerted action by leadership must occur. This argument has been made most strongly by Deming (1985), although it applies to almost all types of organizational change efforts of this magnitude (see Metz, 1984 for an excellent discussion). Part of the education of management should concern an understanding of the resistances to change in control systems and the likelihood that they can be successfully overcome. The balance of this paper will address these issues.

Factors Influencing Acceptance of Control Systems

Lawler (1976) has analyzed the nature of control systems in organizations and the kinds of resistances to their development. In his analysis, he points out that resistance to control systems is most likely when:

1. The control system measures performance in a new area.
2. The control system replaces a system that people have a high investment in maintaining.
3. The standards for control are set without participation.
4. Control system feedback does not go to those who are measured.
5. Control system feedback goes to higher levels in the organization and is used in the reward system.
6. People affected by the existing system are relatively satisfied and see themselves committed to the organization.
7. The people affected by the system are low in self-esteem and authoritarianism.

Using Lawler's analysis, we can examine the resistance to change from the product inspection approach to the process control approach. We can also use the information to suggest ways in which leaders can approach the change process. Following that, an argument will be made on the basis of contemporary work motivation theory as to why the process control approach should be very successful, once implemented.

In the following discussion, Lawler's seven sources of resistance will be considered in turn for the operations of a large aircraft overhaul facility in the Navy. Although the observations made here are not based upon empirical data, they have been corroborated by senior managers in that organization and by on-site research personnel.

1. The process control approach does measure performance in new areas. In fact, the essence of the approach involves measuring several significant features about the production process prior to completion of a product or service.

2. Ultimately many of the personnel who are currently used in the quality control department will be deployed to other parts of the organization, or be conducting quality control activities not currently being performed (e.g., incoming supplies, customer services, etc.). This displacement and/or retraining of personnel is viewed as a threat by those in the current quality control function.

3. Quality control standards are usually established by engineers or quality technicians. In the case of process control, however, a fixed standard has no meaning. Control is defined by taking actions to keep the process within variability limits which are determined by the process itself. Because the limits change as a function of improvements in the system, no fixed standard can be applied.

4. Using the process control approach, the basic data is collected by the performer. In this sense, feedback is immediate. Furthermore, because the information gathered is typically a historical record with relational information on the record (e.g., a control chart), the worker can evaluate the data and determine what action, if any, need be taken.

5. Whether the data is fed to higher levels and used within the reward system depends upon a number of factors. The most significant is the degree to which the worker has discretion to make decisions concerning corrections to the system. This, in turn, is usually based upon the extent of the system changes and their costs, but could also be a reflection of the management philosophy of the organization. This will be considered in greater detail in a subsequent discussion.

6. This is the "status quo" factor, and it can be said that a change in the intertial state of the organization will be determined by whether or not a "critical mass" (Deming, 1985b) can be developed to overcome the status quo. The state of inertia in most bureaucracies, such as those in most large bureaucratic organizations, is at steady state and resistant to change under normal workload conditions.

7. It is difficult to assess this factor. The people most affected by the quality control system are those in the "production" area. As a group, they are the largest in number and exert the greatest influence on achieving the mission of the organization. However, under the current product inspection approach, they receive the most censure when product quality does not meet specifications/test. Managers have been of the opinion that this has led the workers to lose identity with the quality of their products because someone else has been responsible for detecting it.

Lawler has indicated that, to the extent that these factors hold for workers, they will engage in non-productive or even counter productive behaviors. Using his analysis and the previous discussion, it appears that the process control approach should meet less resistance in terms of factors 3, 4, 6, and 7. That is to say that it (a) does not deal with standards per se, (b) provides feedback information to the performer, (c) is of greater benefit to most of the workforce than the existing system and (d) can enhance the self esteem of the worker as he begins to take charge of the quality of his work.

Of the other factors, only the second appears to be of significant concern in terms of resistance to change. In the organization under study, quality control is vested in a functional department. While performing inspections or audits of the work conducted in the production area is not their only function, it does define their central raison d'etre. In addition, this organization is one of several which reports to a headquarters. Both the headquarters and the sister organizations contain quality control functions based upon inspection and audit. Resistance here would have to be overcome.

Factors 1 and 5 are potentially areas of resistance because of the new measures and added work required (1) and because the information could be used to evaluate performance of the workers (5). Neither of these is necessarily negative, but the workforce is often wary concerning the use of performance measures. If the management philosophy and the culture of the organization is one that rewards improvement then there will be little resistance.

Process Control and Worker Motivation

Our discussion thus far has focussed on the desirability of changing from product inspection to process control and the nature of resistances in making such a change. While it appears obvious that such a change is both desireable and feasible from a management standpoint, what is in it for the worker? After all, with the exception of some of the existing quality control personnel, the major job changes will be that of the worker and perhaps his immediate supervisor. If this change is not seen by the worker as having incentive value, then it will very likely be resisted.

Job Characteristics Theory (Hackman & Lawler, 1971; Hackman & Oldham, 1976) provides a conceptual framework to evaluate the design of a workers job to include the process control approach to quality control. The theory is based upon a plethora of research which has revealed that there are three psychological states which contribute to worker motivation. These are feelings of meaningfulness, responsibility, and knowledge of results. The theory goes on to describe what job characteristics will result in these feelings. The theoretical relationships can be schematicized as follows:

Five Core Characteristics	Psychological States	Outcomes
1. Skill Variety 2. Task Identity 3. Task Significance	Feeling of meaningfulness	High intrinsic motivation High quality work
4. Autonomy	Feeling of responsibility	High satisfaction
5. Feedback	Knowledge of results	Low absenteeism and turnover

Considering each of these characteristics in turn, we can determine the motivating potential, or incentive, of the workers job when process control becomes a part of the job. Skill variety is obviously increased because the job will now involve collection of data, charting of data, and reporting process aberrations. Task identity should increase because attention will be focussed on aspects of the process which were previously receiving less formal, e.g., measurement, attention. The perceived significance of the task may also be enhanced because taking process control actions should occasion interaction with supervisors, staff, and managers which would not ordinarily occur. Autonomy will be increased because quality control actions and responsibilities will now be formally placed in the hands of the worker. Finally, feedback will be immediate in terms of the things being measured. Feedback as a result of process changes will, in most cases, be immediate as well.

From this logical analysis and the model displayed in Table 1, we can predict the outcomes displayed there. During the course of the forthcoming year, these hypothetical relationships will be tested in the Navy maintenance environment. The use of process control as a method to enrich jobs has not received attention in the empirical literature, but the aforementioned analysis suggests that it should be an effective way to motivate workers as well as increase the quality of their efforts.

Locke (1980) has indicated that job enrichment has not been effective in motivating employee performance, when the effects of goal setting have been controlled. Such a confounding of variables is not expected in the process control situation, because the demand characteristics of process control are not upon employee effort, but upon removal of system-generated variation. Performance increases, then, would probably result from enhanced ability to perform the job, as opposed to the motivation to increase effort.

This latter point begs the question of whether employee motivation could be at all affected by adopting the process control approach. Lawler's (1976) discussion of expectancy models of worker motivation would suggest that, in the absence of extrinsic rewards for quality improvement, the motivational impact can only be derived from (a) the $E \rightarrow P$ probability or (b) or the valence associated with outcomes other than extrinsic rewards. From our logical analysis above, it is not clear which of the two, or some combination of both could account for changes in the motivating potential of a job enriched by inclusion of process controls for quality. These questions will be addressed in future research.

References

Deming, W. E. Transformation of western style of management. Interfaces 15: 3 May-June 1985 (pp. 6-11)

Hackman, J. R. & Lawler, E. E. Employee reactions to job characteristics. Journal of Applied Psychology, 55, 1971, 259-286.

Hackman, J. R. & Oldham, G. R. Devrlopment of the Job Diagnostic Survey. Journal of Applied Psychology, 60, 1975, 159-170.

Houston, A., Shettel-Neuber, J., & Sheposh, J. Management methods for quality improvement based on process analysis and control. NPRDC Technical Report, 1985 (in press).

Juran, J. Quality Control Handbook. New York: McGraw Hill, 1974.

Lawler, E. E. Control systems in organizations. In Dunnette, M. D. (Ed.) Handbook of Industrial and Organizational Psychology, Chicago: Rand McNally, 1976.

Locke, E. A., Feren, D. B., McCaleb, V. M., Shaw, K. N. & Denny, A. T. The relative effectiveness of four methods of motivating employee performance. In Duncan (Ed.) Changes In Working Life, Wiley. 1980.

Metz, E. J. Managing change: Implementing productivity and quality improvements. National Productivity Review, Summer 1984, 303-314.

Scherkenbach, W. W. The process of continuous improvement. Presented at the symposium "Managing Systems of People and Machines for Quality and Productivity", San Diego, 1984.

Tribus, M. Creating the quality company. Presented at the symposium "Managing Systems of People and Machines for Quality and Productivity", San Diego, 1984.

A GROUP WAGE INCENTIVE SYSTEM: DESIGN AND IMPLEMENTATION ISSUES

Deborah A. Mohr
Navy Personnel Research and Development Center
San Diego, California 92152-6800

INTRODUCTION

In an effort to improve performance and reduce costs in a naval shipyard, a group wage incentive system for production workers was developed, implemented, and evaluated. The system was designed to improve performance efficiency without negatively affecting schedule adherence, product quality, or workers' job attitudes.

This project was part of a continuing research program to investigate the effects of wage incentive systems in Navy industrial facilities. Previous projects evaluated the effects of performance contingent reward systems (PCRSs) with a variety of civil service employees: key entry operators, small purchase buyers, and aircraft engine mechanics. Under a PCRS, employees earn cash bonuses (incentive awards) for work performed above established standards. The more performance exceeds the standard, the larger the bonus. PCRS rewards are paid through existing award programs, are recurrent (being accrued as often as performance exceeds standard), and are in addition to employees' base salary.

The present effort differed from previous projects in that awards were based on measures of group performance. Shipyard production workers typically work together in teams (called work gangs) of 10 to 20 employees supervised by one foreman. Thus, a PCRS based on measures of group performance was more appropriate than one based on individual performance measures.

INCENTIVE SYSTEM OVERVIEW

The performance measure used for this system was one of performance efficiency. It was calculated by dividing the manhours allowed to complete a work gang's jobs by the manhours actually expended to complete the work. Thus, when work is completed in exactly the time allowed for that work, the work gang's performance efficiency, called a performance factor (PF), is 100 percent. When work is completed in less time than the allowance, the gang's PF will be greater than 100 percent and manhours will be saved. Both inputs to this measure (manhours allowed and expended) were routinely collected by the shipyard's management information system (MIS). Prior to implementation, the shipyard MIS was further enhanced to provide more accurate performance measures and to provide monthly automated incentive award calculations and continual award tracking.

Under the shipyard's PCRS, work gangs were eligible for awards whenever they saved manhours by completing their jobs in less time than the manhours allowed for those jobs. The value of these saved hours was shared with employees in the form of incentive awards. The work gang's saved hours were distributed to members based on each worker's contribution to the workgang (his or her share of the gang's total work hours). Based on the 50 percent sharing rate used during the system test, half of the cost savings associated with a work gang's manhour savings were paid out to gang members as incentive awards. The remaining 50 percent was retained by the shipyard. The actual value of each saved hour was based on the employee's accelerated hourly wage rate.

A similar incentive system was established for shop foremen in which all foremen comprised one group eligible for awards whenever performance of the entire shop resulted in manhour savings. In addition, to encourage foremen to work together to help the shop improve, each foreman received a one-time bonus of $125 the first time the shop's PF exceeded 100 percent.

Because the shop selected for test of the incentive system historically spent many more manhours to complete its work than were allowed, few work gangs would save manhours and earn incentives at typical performance levels. Since incentive systems do not motivate employees to improve performance unless they believe it's possible to earn awards, shipyard managers decided to adjust all performance measures upward by 10 percent for the purposes of subsequent award calculations. Thus, work gangs actually accrued manhour savings whenever their PF exceeded 90 percent and foreman earned awards whenever the shop's PF exceeded 90 percent. Despite this adjustment, the incentive system rewarded employees for performance improvement.

IMPLEMENTATION

Prior to implementation of the test system, a shipyard instruction was issued documenting the incentive system and specifying responsibilities during the test period. A senior military officer was assigned as project manager and a general foremen within the test shop served as system coordinator. An agreement was negotiated with the local union and approval was obtained from the appropriate headquarters commands. Finally, employees and supervisors in the test shop were given training to assure their understanding of the enhanced performance measurement system and the group incentive system. The PCRS was then implemented for test in Shop 31, the shipyard's inside machine shop. Shop 31 is one of 17 shops at the shipyard and employs approximately 480 wage grade employees and 23 foremen assigned to 18 work gangs.

RESULTS

During the 19 months of the system test a total of $177,000 was earned by employees. Sixteen of the eighteen work gangs (comprising 89 percent of shop employees) earned awards. Of those employees earning awards, the average total earnings during the test period was $419. (Total earnings ranged from $1 to $2488.) Foremen earnings averaged $237 for the 2 months that the shop's performance factor exceeded 90 percent.

Evaluation of the incentive system test revealed that the program produced a significant increase in the shop's performance efficiency (see Figure 1). For analysis purposes the 19 month test was divided into two phases: Incentive Phase 1 consisted of the first 8 4-week incentive periods and Incentive Phase 2 consisted of the remaining 11 4-week incentive periods. The shop showed a 7.5 percent improvement over average baseline performance during Incentive Phase 2 (an improvement from 91.4% to 97.6%). During the first incentive phase, the shop maintained its baseline performance despite a severe workload reduction (performance averaged 89.0%). Two comparison shops (see Table 1) showed substantial performance decreases during the same time, although their workload reductions were less severe than that of the test shop. Since the end of the 19-month test, the test shop has maintained its improved performance. As expected, implementation of the system did not cause any negative effects on the shop's schedule adherence or product quality.

Participants' job attitudes and evaluations of the incentive system were assessed during the test. Although recognizing certain problems related to system operation (particularly the effects of the workload reduction), 80% of those expressing an opinion favored continuing the incentive system. No positive or negative effects on workers' job attitudes (e.g., job satisfaction and job stress) were found.

A cost savings analysis revealed that the net cost savings due to improvements over baseline performance during the system test exceeded $600,000. If similar results occurred following expansion of the system to all other production shops, the shipyard could realize net cost savings of approximately $6,794,000 annually.

The shipyard realized a number of concurrent positive outcomes from the system test, including improvements in shop practices and initiation of management actions directed toward resolving productivity impediments. Foremen began taking greater care in preparing employee time cards, correcting labor mischarges, and reviewing work documents before beginning jobs. As a result of the interest in improvement engendered by the incentive system, a number of productivity impediments were highlighted during the test. These impediments were not new. Rather, they were long-standing shipyard problems that became more salient when money was tied to performance. The incentive system provided the impetus to attack these problems and as a result a shipyard-wide problem-solving team was established and successfully resolved a number of these issues.

IMPLEMENTATION AND MAINTENANCE ISSUES

Throughout this effort, various issues surfaced that revealed the complexity of designing and implementing productivity improvement systems in real organizations. While the success of the test system indicates these issues can be effectively resolved, nonetheless, some important conclusions can be drawn from this project.

Several issues arose during the design of the incentive system. The decision to develop a group system was a logical result of analysis of the shipyard's work settings. Implementation of a typical incentive system (most likely based on measures on individual performance) would have been inappropriate and quite possibly ineffective. Managers considering a PCRS should realize that no standard system exists. The PCRS must be designed to fit the organization and its priorities.

A number of incentive system parameters had to be specified in the design phase, as well. These included: the incentive level (the performance level at which employees were eligible to earn awards), the sharing rate (the proportion of cost savings shared with employees), and the savings distribution method (the way savings were shared among work gang members). As discussed, the incentive level was dropped to 90 percent in the belief that this level would be seen as attainable by workers and that it would improve the motivating potential of the system. The 50 percent sharing rate used in this test was selected because it was the maximum allowable by federal regulations and because it was likely to be perceived as fair to both employees and the organization. Distribution of savings based on worker inputs was used to further strengthen participants' perceptions of fairness. These design parameters can be assumed to have been effective based on the favorable results of the test period. However, there is no way to determine if different parameters might have been substantially more effective. Little research has been done to investigate the effectiveness of different levels of these parameters (e.g., a 20% vs. a 50% sharing rate) or the situational variables that require parameter changes.

During the implementation and maintenance phases, additional issues arose. Primary among these was management's relationship to the incentive system. The success of organizational change efforts such as incentive systems is at least in part contingent on active support from management. A high degree of commitment to the program is necessary before and after implementation, commitment involving more than just verbal support. It is difficult to implement effective changes when either top management or those expected to implement change are unsupportive.

During the test of the incentive system, the shipyard experienced a rather significant workload reduction. This appeared to have precluded performance improvement until the shop's staffing level was brought into balance. Workers are unlikely to improve their productivity if, by so doing, they believe they will risk their jobs. In implementing performance improvement programs, managers must continually address the balance between workload and staffing within the organization and must develop means to capitalize on the effects of resulting improvement.

Tying money to performance highlighted a number of long-standing shipyard problems that were subsequently addressed by a problem-solving team. The importance of tapping this increased interest in performance improvement cannot be overestimated. Many shipyard managers believed that the incentive system's major benefits were in encouraging supervisors to do their jobs and in focusing efforts on resolving productivity impediments. Such auxiliary benefits of incentive systems should not be overlooked.

Finally, the issue of incentive system expansion surfaced. To continue to run a test system in only one shop is not feasible. With proven cost savings resulting from performance improvement the next logical step is to expand to other shops. Managers must carefully consider how and how far to expand a successful incentive system. During expansion, care must be taken to adapt the system to other sites and to continue to monitor its effectiveness. Managers should also consider means to include production support and other indirect workers in such incentive systems.

Although there are a substantial number of complex issues that must be faced in developing and implementing wage incentive systems, their proven effectiveness indicates that such efforts are worthwhile. Further research to identify effective design parameters, the increased use of automation to support wage incentive systems, and the benefits that can be derived from additional experience with these systems will help to limit the effort required to design and implement wage incentive systems in the future.

The opinions expressed in this paper are those of the author and should not be construed as official or as reflecting the views of the Department of the Navy.

Figure 1

Trends in Performance Factors for Shop 31

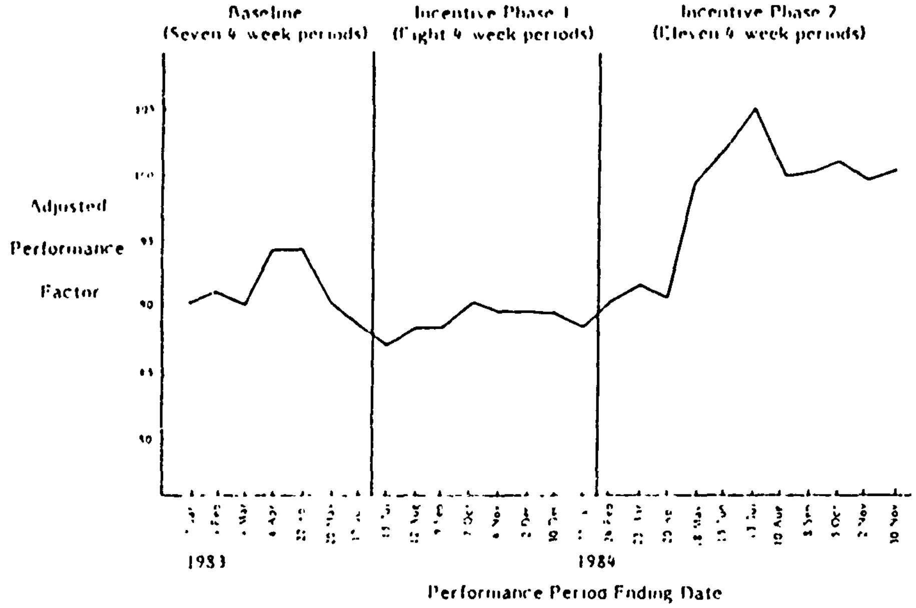


Table 1

Performance and Workload Trends For Three Key P. oduction Shops During Baseline and Two Incentive Phases

Performance Factor (PF)					
	Baseline[a]	Incentive Phase 1[b]		Incentive Phase 2[c]	
	Average	Average	% Change	Average	% Change
Shop	PF[d]	PF	From Baseline	PF	From Baseline
31	.867	.853	-1.6	.926	+6.8
38	.696	.682	-2.0	.716	+2.9
56	.780	.774	-.8	.792	+1.6

Workload: Average Man-day Allowances per 4-week Perlod					
	Baseline	Incentive Phase 1		Incentive Phase 2	
	Average	Average	% Change	Average	% Change
Shop	Man-days	Man-days	From Baseline	Man-days	From Baseline
31	8281	5559	-32.9	5952	-28.1
38	7877	7250	- 8.0	7408	- 6.0
56	7913	7816	- 1.2	7935	+ .3

[a] Baseline: 10 January 1983 - 14 July 1983.

[b] Incentive Phase 1: 15 July 1983 - 27 January 1984.

[c] Incentive Phase 2: 28 January 1984 - 30 November 1984.

[d] Figures represent the average PF within each time frame.

GOAL SETTING WITH NAVY PRODUCTION WORKERS: PROBLEMS AND POTENTIAL

Kent S. Crawford

Navy Personnel Research and Development Center
San Diego, California 92152-6800

Paper presented at the Military Testing Association
San Diego, California, October 1985

ABSTRACT

A goal setting program was implemented in an Navy industrial organization that used engineered performance standards. Results indicated that workers whose baseline performance was below standard set more difficult goals and improved their performance more than high performing workers. Discussion centered on the role of context in influencing goal setting effectiveness in Navy organizations.

BACKGROUND

There is growing concern in the United States with what has been labeled the "U.S. Productivity Crisis" (Newsweek, 1980). This crisis is manifested in the declining rate of growth in the output per hour of labor. The United States finished well behind six other industrial nations in productivity increases from 1968 to 1978 (Bureau of Labor Statistics, 1979). Within the Navy, concern over worker productivity has created increasing interest in productivity improvement at all levels of the organization.

Traditionally, productivity programs in both the military and civilian sectors have centered on technological improvements and capital investments. While the importance of these hardware-oriented approaches is obvious, there is a growing body of organizational literature that suggests that significant productivity improvements can be realized through improved worker motivation (Greiner, Hatry, Koss, Millar, & Woodward, 1981). Several different techniques have been investigated, including autonomous work groups, job restructuring, participative management, and monetary incentive systems. Each of the above approaches has been shown to have merit under differing circumstances (Cummings & Molloy, 1977; Patten, 1977).

Goal Setting

Goal setting is an area of organizational research that seems to be especially promising in terms of enhancing worker motivation and performance. Research has shown that goals are a major source of work motivation (Mitchell, 1979). Likewise, a recent review of field studies using goal setting techniques found a 16% median improvement in worker performance (Locke, Feren, McCaleb, Shaw, & Denny, 1980). Based on the results of a number of highly successful lab and field studies, goal setting has been called "a simple, straightforward, and highly effective technique for motivating employee performance" (Latham & Locke, 1979, p. 80).

While it is clear that goal setting can be an effective motivational technique, we feel that comparatively little study has been directed toward careful investigation of the possible limitations of the approach. There are very likely no panaceas in any field of applied science (Locke, Sirota, & Wolfson, 1976). As such, it would seem that goal setting theory is subject to boundary limitations regarding to whom it applies and where it works best (Miner, 1980). The current study addressed this issue by examining the effectiveness of goal setting in an industrial organization that made extensive use of engineered performance standards.

The use of task standards is an outgrowth of the basic tenets of scientific management (see Taylor, 1967). These standards represent the time a trained employee working at a normal pace would be expected to complete a given task. They are usually based on time and motions studies or on historical performance trends. Within many industrial organizations work standards have been established for most production jobs. While these standards are used for advance cost estimates, manpower projections, and other planning requirements, they also serve another implicit function—they establish acceptable performance levels for workers (Maynard, 1971). In this sense, a standard is a goal for workers to try to achieve (Locke, 1978).

If achieving standards represents an acceptable performance level, then industrial organizations that make extensive use of task standards may encounter problems in implementing goal setting programs for workers. The basic proposition of goal setting theory states that there is a positive relationship between the difficulty of an accepted task goal and level of performance on the task (Locke, 1968). Considerable research has shown that hard, specific goals (if accepted) result in performance improvements (Locke, Shaw, Saari, & Latham, 1981). Performance standards certainly define specific goals; however, they may not always be difficult. While performing at standard level may be challenging for employees with low ability and work motivation, it wouldn't represent a challenging goal for a motivated and highly skilled employee.

Goals Versus Current Performance Levels

The objective of goal setting is to establish specific, challenging goals for all workers. Individuals are encouraged or required to have different goals dependent on their current performance level. The problem with goal setting in an organization using industrial standards is that the organization is sending mixed messages. The supervisor is trying to establish a challenging goal for the worker (often above standard performance level) while the organization has previously defined standard performance as acceptable.

One means of possibly reducing the above problem is for the supervisor to assign goals. The supervisor could then set goals based on current performance independent of existing standards. Research has shown that if goal difficulty is held constant, equal goal acceptance and performance improvements are obtained regardless of whether goals are assigned or set participatively (Dossett, Latham, & Mitchell, 1979; Latham & Saari, 1979; Latham, Steele, & Saari, 1981). However, there is some evidence to suggest that when both participative and assigned goals are set independently, participative goal setting may result in more difficult goals (Latham & Yukl, 1975a; Latham, Mitchell, & Dossett, 1978). However, given the current state of knowledge, it would be difficult to predict which method would be more effective in organizations with existing performance standards. Also, regardless of the method used, it is not certain whether or not workers would set or accept goals above standard.

While the best means of setting goals remains unclear, there is one sub-group of workers who might be expected to improve more as the result of a goal setting program in an industrial organization — low performers. Individuals who are currently performing below standard are not faced with conflicting messages when goals are established by or with the supervisor. In addition, it is possible that high performers are more likely to understand task requirements and have personal performance goals than are low performers. Thus, it seems reasonable to expect a goal setting intervention with production workers to have its greatest impact on low performers.

One recent study supports this contention for nonproduction workers. Pritchard, Bigby, Beiting, Coverdale, and Morgan (1981) found that for data transcribers, goal setting and feedback had a positive impact on poor performers but no impact on good performers. They argued that since the treatment was designed to increase motivation, and since the good performers were probably already motivated, the treatment had little impact on them.

The purpose of the current study was to examine the impact of goal setting and feedback on the performance of Navy industrial employees working within the context of existing performance standards. It was hypothesized that low performers (workers historically performing below standard) would have more difficult goals and would show greater performance improvements than high performers (workers historically performing at or above standard).

METHOD

The engine division of a Naval Air Rework Facility (NARF) served as the research setting for the study. Production workers in this division were involved in the overhaul of aircraft engines, components, and accessories. They were all civil service employees, predominately male, and most had a high school education.

Prior to the goal setting intervention, assigned tasks included a description of required work and the time allocated for the work (i.e., the performance standard). However, although workers knew how well they performed on individual tasks when they completed them, they were not provided with any summary feedback of their performance on all work during a given time period. Since feedback has been shown to be a necessary condition for goal setting to be effective (see Locke et al., 1981), it was first necessary to design an individual work measurement and feedback system.

A computerized system was developed to measure individual level performance using the existing NARF management information system. A weekly report was then generated for each employee providing performance feedback for the previous week. The performance measure was based on how well workers performed against standard and was calculated by taking the ratio of time expended on tasks in a given week to the total standards earned in that same week. This figure was then multiplied by 100. Thus, a rating of 100 meant that an individual completed all work within the standard time allocated. Ratings higher than 100 indicated performance better than standard and those lower than 100, performance below standard.

Twenty-two production shops in the engine division were included in the study. Each shop was supervised by its own foreman. Eleven shops were selected for the goal setting treatment and the remaining 11 were used as a comparison group. All shops were both spatially and structurally distinct subunits.

The 11 experimental shop foremen were trained in the use of the new worker feedback reports and in goal setting; six were trained to assign goals to subordinates while the remaining five were trained to set goals participatively with subordinates. The foremen were asked to arrive at different goals for different subordinates based on the worker's ability, motivational level and current performance. The foremen were initially resistant to the notion of setting challenging goals for workers who were already performing at or above standard. They felt that these employees were currently doing more than should be expected of them. However, the foremen agreed to proceed and give the program a fair chance.

. The 11 foremen in the experimental groups met individually with their subordinates to either assign a challenging performance goal or to arrive at such a goal participatively. In addition to receiving the weekly performance report, workers in the goal setting shops met with their foremen individually every 2 to 4 weeks to discuss progress towards their goals and possible work problems.

An 18-week period prior to the beginning of goal setting and feedback was used to establish a baseline level of performance for both the experimental and comparison workers. The 22-week period after program implementation was used to assess program effectiveness. Sixty-seven workers participated in setting their goals while 57 were assigned goals. The comparison shops were composed of 117 workers.

The weekly employee performance data were aggregated to form single pre- and post-treatment performance scores for each worker. Reliability coefficients, computed on the weekly performance measures for the baseline period, indicated that the data were sufficiently reliable for use as overall performance measures (coefficient Alpha = .75).

Research has shown that objective measures of goal difficulty are often better predictors of performance improvement than subjective measures (Yukl & Latham, 1978). For this reason, goal difficulty was operationalized as the difference between an individual's baseline performance score and his/her goal. This measure of goal difficulty allowed for the partial control of baseline individual differences in ability and motivation.

RESULTS

General Results

The initial analyses examined all workers independent of their baseline performance.

Manipulation check. In order to verify the treatment conditions, workers in both the assigned and participative groups were asked to respond on a 4-point Likert scale how much influence they had in setting their goals (1 = a lot of say; 4 = no say). Individuals in the participative condition ($\bar{X}$ = 1.3 reported significantly more influence ($p < .01$) than did workers in the assigned condition ($\bar{X}$ =, 3.0)

Performance Change. The mean performance levels for workers in the treatment and comparison groups are presented in Table 1. A test for homogeneity of regression coefficients yielded no significant differences across the groups. Therefore, an analysis of covariance was used to contrast the experimental and comparison groups with the baseline performance measure as a covariate. A significant main effect was found ($p < .01$) indicating differences in treatment performance levels across the groups. Follow-up tests indicated that both the assigned goal setting group (adjusted $\bar{X}$ = 108.3) and the participative group (adjusted $\bar{X}$ = 106.2) were significantly higher ($p < .05$) than the comparison group (adjusted $\bar{X}$ = 101.7). There were no significant differences between the two goal setting groups during the baseline or treatment periods.

Table 1

Mean and Adjusted Performance Efficiency Scores

	Mean Performance Efficiency				
Group	Baseline (B) Period	Test (T) Period	Performance Change (T - B)	Adjusted[a] Test Period	N
Comparison	99.5	102.4	+2.9	101.7 [b,c]	117
Experimental	97.5	106.4	+8.9	107.1 [b]	124
Assigned goals	99.3	108.9	+9.6	108.3 [c]	57
Participative goals	96.0	104.4	+8.4	106.2 [c]	67

[a] Adjusted to control for differences in baseline period performance.

[b] For the analysis contrasting the comparison and combined experimental groups, covariance F = 11.7, p<.001.

[c] For the analysis contrasting the comparison, assigned goals, and participative goals groups, covariance F = 6.2, p<.01.

Goal difficulty. A significant correlation was found between objective goal difficulty and degree of performance improvement ($r = .40$, $p < .001$). No difference was found between the average level of goal difficulty in the participative group ($\bar{X} = 8.1$) and the assigned group ($\bar{X} = 11.0$). This finding is consistent with the earlier finding indicating no difference in performance between the two groups in the treatment period. One interesting finding did emerge as to the actual goals set in the two groups. Seventy-three percent of the participative workers had a goal of 100 (or standard level of performance) whereas only 8% of the assigned workers had a goal of exactly 100. This distribution of goals at or different than 100 across the two groups was statistically significant (Chi Square = 30.2, $p < .001$). It thus appears that workers who had some influence in their choice of goals preferred a goal equal to existing organizational standards.

High Versus Low Performers

Workers in the experimental and comparison groups were divided into two categories based on their level of performance during the baseline period: (1) high performers were individuals whose average performance during the 18-week baseline period was at or above standard (i.e., 100), and (2) low performers were individuals whose average performance was below standard.

Performance change. The mean performance levels for high and low performers by different treatment groups are presented in Table 2. Two repeated measures analyses of variance were performed--one for experimental high performers and one for experimental low performers. No main or interaction effects were found for high performers indicating that there was no performance improvement in either the assigned or participative conditions. On the other hand, a main effect for time period was found for the low performers ($p < .01$), indicating that low performers in both the assigned and participative conditions significantly improved their performance as a result of goal setting.

Table 2

Mean Performance Efficiency Scores for
High and Low Performers

Group	Baseline Period	Test Period	Performance Change[a]	Adjusted Test Period	N
High Performers					
Experimental	114.0	118.3	+4.3	117.1 [b]	57
Comparison	111.2	111.1	- .1	112.5 [b]	52
Low Performers					
Experimental	83.5	96.3	+12.8	98.3 [c]	67
Comparison	90.1	95.4	+ 5.3	93.5 [c]	65

[a] The difference between the test period and baseline period performance scores.

[b] Covariance F = 2.8.

[c] Covariance F = 6.3, p< .05

Because regression towards the mean presented a potential confounding interpretation for the improvements with the low performers, analyses of covariance were also performed on these data. A test for homogeneity of regression coefficients revealed no significant difference among high and low performers across the three groups. Thus, an analysis of covariance was conducted separately for high and low performers using the baseline performance measure as a covariate. The results were identical to those reported earlier. High performers in the experimental and comparison groups did not differ while the low performers in the experimental groups (adjusted $\bar{X} = 98.3$) were significantly higher ($p < .05$) than the low performers in the comparison group (adjusted $\bar{X} = 93.5$). Overall, the results suggest than goal setting had a positive impact on the performance of low performers and no effect on high performers.

Goal difficulty. One factor that could explain the different effects of goal setting on low and high performers is goal difficulty. It was proposed that low performers would set (or be assigned) more difficult goals relative to their baseline performance level than would high performers. The results relevant to this hypothesis are given in Table 3. The mean goal difficulty level for all the low performers (15.8) was significantly greater ($p < .001$) than the mean level of goal difficulty for high performers (-.4). On the average, high performers had goals that were slightly lower than their baseline performance level, whereas low performers had average goals that were approximately 16 points above their baseline performance level.

Table 3

Mean Goal Difficulty, Goal Acceptance, and Performance Change for Experimental High and Low Performers

Experimental Group	Mean Goal Difficulty	Mean Performance Change	N
High Performers			
Assigned	5.0	+5.7	27
Participative	-5.3	+3.1	30
Total	-.4	+4.3	57
Low Performers			
Assigned	16.7	+13.0	30
Participative	15.0	+12.6	37
Total	15.8	+12.8	67

Analyses were also undertaken to compare goal difficulty for assigned and participative workers. Results indicated that mean goal difficulty was significantly higher ($t = 2.54$, $p < .05$) for high performers who were assigned goals (5.0) than for high performers who participatively set goals. Indeed, high performers who participatively set goals had an average goal that was more than five points below their baseline performance. No significant difference was found between the mean goal difficulty level of poor performers in the assigned (16.7) and participative (15.0) conditions.

Because the results indicated that a number of workers had negative goals (e.g., goals that were lower than their baseline performance), additional analyses were performed to assess the relation of positive and negative goals to performance change. Results indicated that 51% of the high performers ($\underline{N} = 28$) had goals that were lower than their baseline performance whereas only 5% of the low performers ($\underline{N} = 3$) had negative goals. Seventy-five percent of these high performers with negative goals ($\underline{N} = 21$) were in the participative condition. A significant positive correlation was found between goal difficulty and performance change for low performers with positive goals ($r = .41$, $p < .001$). However, no significant relationships were found for high performers with either negative or positive goals, although the correlation for the later group was marginally significant ($r = .27$, $p < .10$). Also, the performance change scores for high performers with positive goal ($\bar{X} = 8.1$) were higher than those for high performers with negative goals ($\bar{X} = .5$), although this difference was only marginally significant ($p < .10$). These findings suggest that goal setting was somewhat successful for high performers, but only if they had goals higher than their baseline performance.

DISCUSSION

These findings provide support for Locke's (1968) goal setting theory, although they also suggest that goal setting effectiveness may be contingent on contextual factors. First, there was a positive relation between goal difficulty and performance improvement; however, this relationship only held for workers whose goals were higher than their baseline performance. In addition, consistent with the hypothesis concerning low performers and an earlier study by Pritchard et al. (1981), goal setting was more effective with low performances than with high performers. This differential impact was reflected both in terms of greater goal difficulty and more performance improvement.

The goal setting process appeared to be affected by the NARF's use of engineered performance standards. This is supported by the lower mean goal difficulty level for high performers (relative to low performers) that occurred both when goals were assigned and participatively set. Assigned goal setting did result in more difficult goals for high performers than did participative goal setting; however, this difference was not reflected in significant differences in the degree of performance improvement for the two groups. The large proportion of goals that were participatively set at 100 (standard performance level) also suggests that organizational task standards can influence the goal setting process. Workers may have felt that 100% was the most reasonable goal for the organization to expect them to achieve--independent of their baseline performance. With the exception of these findings, participative and assigned goal settings yielded virtually identical results. This is consistent with a large number of lab and field studies (see Locke et al., 1981). The failure to find more difficult goals set in the participatively treatment groups may partially reflect the role of context. Where workers had some influence over their goals, they often opted for what they considered to be fair (i.e., standard) rather than what they felt would be challenging.

Some caveats seem in order. First, the sample size was not large, especially when it was broken down into subgroups. Second, the characteristics of the work force may have played an important role. Navy production workers have more job security than most private sector industrial employees. Thus, these workers may have felt more latitude in choosing negative goals. Finally, goal setting effectiveness was only assessed over 5-1/2 months. There is some evidence to suggest that goal setting effects are not sustained over longer time periods (see Ivancevich, 1976).

Overall, the results of this study support the general contention of this paper. Goal setting is an effective motivational technique for Navy production workers but is subject to contextual constraints. In this sense, it has both potential utility and potential problems. There are limitations as to conditions where goal setting works best and for whom it works best (Ivancevich, 1978). There is a need to follow Latham and Yukl's (1975b) suggestion that future research in goal setting begin developing more of a contingency framework. This study was one step in that direction.

REFERENCES

Bureau of Labor Statistics. United States Department of Labor, 1979.

Cummings, T. G., & Molloy, E. S. *Improving productivity and the quality of work life*. New York: Praeger, 1977.

Dossett, D. L., Latham, G. P., & Mitchell, T. R. The effects of assigned versus participatively set goals, KR, and individual differences when goal difficulty is held constant. *Journal of Applied Psychology*, 1979, *64*, 291-298.

Ivancevich, J. M. Effects of goal setting on performance and job satisfaction. *Journal of Applied Psychology*, 1976, *61*, 605-612.

Ivancevich, J. M. *Individual differences and organizational goal setting: Issues, research, and new directions*. Paper presented at The National Academy of Management, San Francisco, 1978.

Latham, G. P., & Locke, E. A. Goal setting: A motivational technique that works. *Organizational Dynamics*, 1979, *8*, (2), 68-80.

Latham, G. P., Mitchell, T. R., & Dossett, D. L. Importance of participative goal setting and anticipated rewards on goal difficulty and job performance. *Journal of Applied Psychology*, 1978, *63*, 163-171.

Latham, G. P., & Saari, L. M. The effects of holding goal difficulty constant on assigned and participatively set goals. *Academy of Management Journal*, 1979, *22*, 163-168.

Latham, G. P., Steele, T. P., & Saari, L. M. *The effects of participation and goal difficulty on performance*. TR-GS-6. College Park, Maryland: University of Maryland, 1981.

Latham, G. P., & Yukl, G. A. Assigned versus participative goal setting with educated and uneducated woods workers. *Journal of Applied Psychology*, 1975a, *60*, 299-302.

Latham, G. P., & Yukl, G. A. A review of research on the application of goal setting in organizations. *Academy of Management Journal*, 1975b, *18*, 824-845.

Locke, E. A. Toward a theory of task motivation and incentives. *Organizational Behavior and Human Performance*, 1968, *3*, 157-189.

Locke, E. A. The ubiquity of the technique of goal setting in theories and approaches to employee motivation. *Academy of Management Review*, 1978, *3*, 594-601.

Locke, E. A., Feren, D. B., McCaleb, V. M., Shaw, K. N., & Denny, A. T. The relative effectiveness of four methods of motivating employee performance. In K. Duncan, M. Gruneberg, & D. Wallis (Eds.), *Changes in working life*. New York: Wiley, 1980.

Locke, E. A., Sirota, D., & Wolfson, A. D. An experimental case study of the successes and failures of job enrichment in a government agency. *Journal of Applied Psychology*, 1976, *61*, (6), 701-711.

Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. Goal setting and task performance: 1969-1980. *Psychological Bulletin*, 1981, *90*, (1), 125-152.

Maynard, H. B. (Ed.). *Industrial engineering handbook*. New York: McGraw, 1971.

Miner, J. B. *Theories of organizational behavior*. Hinsdale, IL: Dryden Press, 1980.

Mitchell, T. R. Organizational behavior. *Annual Review of Psychology*, 1979, *30*, 243-281.

Newsweek. An economic dream in peril. Author. September 8, 1980, 50-69.

Patten, T. H. *Pay: Employee compensation and incentive plans*. New York: Free Press, 1977.

Pritchard, R. D., Bigby, D. G., Beiting, M., Coverdale, S., & Morgan, C. *Enhancing productivity through feedback and goal setting*. AFHRL-TR-81-7. Brooks AFB, TX: Air Force Human Resource Laboratory, May 1978.

Taylor, F. W. *The principles of scientific management*. New York: Norton, 1967. (Originally published, 1911).

Yukl, G. A., & Latham, G. P. Interrelationships among employee participation, individual differences, goal difficulty, goal acceptance, goal instrumentality and performance. *Personnel Psychology*, 1978, *31*, 315-323.

The views expressed in this paper are those of the author and are not necessarily those of the Department of the Navy.

THE DETERMINANTS OF GOAL CHOICE, WORK MOTIVATION AND TASK PERFORMANCE

James A. Riedel
Delbert M. Nebeker
Navy Personnel Research and Development Center
San Diego, California 92152-6800

Background

In many organizations goal setting has been found to be a powerful technique for influencing work motivation and performance (Locke, Feren, McCaleb, Shaw, and Denny, 1980). While the positive effects of goal setting on worker productivity have been demonstrated repeatedly, a major weakness of this approach is the failure to specify the process by which goals are set. The central argument of this paper is that the process of goal choice may be central to understanding the relationship among organizational context, goal setting, motivation and performance.

With few exceptions, there has been little research directed toward understanding the determinants of goal choice, acceptance, and commitment (Steers & Porter, 1979). This is unfortunate because goal setting does not take place independently of the work place; reward systems and other work setting characteristics come together to affect goal choice, acceptance, and commitment (Crawford, 1982). Some investigators, however, have attempted to use an expectancy theory model to explain goal choice and acceptance. For example, in both laboratory and field settings goal acceptance has been reliably predicted using expectancy and valence measures (Dachler & Mobley, 1973; Mento, Cartledge, & Locke 1980; Steers, 1975). In a related line of research, expectations of success and the value placed on the outcomes of goal attainment were found to be the principal determinants of "level of aspiration" (Frank, 1941; Hilgard, 1942/1958). These two factors are highly related to the core concepts in expectancy theory: expectancy and valance (Vroom, 1964).

In addition to a limited understanding of how people set goals, the relationship between goal setting and other motivational techniques is unclear. This, in part, may be due to the fact that there has been little integration of goal setting with motivational theories. The role of motivational techniques, such as monetary incentives and goal setting, in work motivation and performance is one of the most underresearched and poorly understood areas in organizational behavior (Opsahl & Dunnette, 1966; Lawler, 1981).

The need for a better understanding of the process of goal choice is evident. This process may provide insight into the relationship among goal setting, organizational context, motivation, and performance. Also, this process, if linked to motivation theory, should help to clarify the relationship among goal setting and other motivational techniques, such as monetary incentives. The purpose of this study is to offer a preliminary model of goal choice, work motivation and performance. This model is presented in Figure 1. As a preliminary test of its validity, selected elements in the model will be examined to determine its usefulness as an explanatory device.

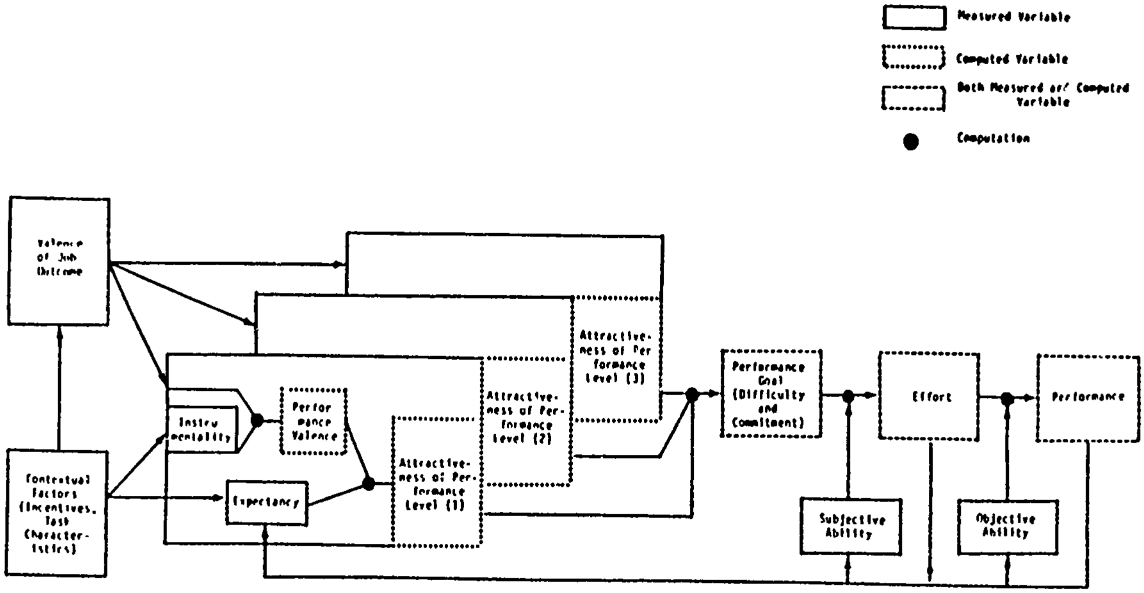

Measured Variable
Computed Variable
Both Measured and Computed Variable
Computation
Valence of Job Outcome
Contextual Factors (Incentives, Task Characteristics)
Instrumentality
Performance Valence
Expectancy
Attractiveness of Performance Level (1)
Attractiveness of Performance Level (2)
Attractiveness of Performance Level (3)
Performance Goal (Difficulty and Commitment)
Effort
Performance
Subjective Ability
Objective Ability

Figure 1. A Model of Goal Choice, Work Motivation, and Performance.

Work Motivation Model

Based on expectancy theory concepts and processes, the model shown in Figure 1 explains work motivation and performance as a cognitive process where an individual chooses, from alternative performance goal levels, the level perceived to be most attractive. This perception of attractiveness is based on various beliefs and feelings a person has regarding the likelihood that performing at certain levels will lead to particular job outcomes. Contextual factors, such as the opportunity to earn monetary incentives for good performance, will influence these beliefs and feelings. The hypothesized effect of the performance goal is to influence the amount of effort a person is wiling to expend in accomplishing the goal. Furthermore, an individual's self-assessment of ability as well as actual ability are presumed to moderate the relationships among performance goal, effort, and performance. Since in this model the goal concept is a major determinant of effort and performance, it is crucial to understand how people choose their performance goals.

The model suggests that contextual factors influence valence, an individual's anticipated satisfaction with particular levels of different job outcomes, and instrumentality, the expectancy that different performance levels are associated with different outcomes. Contextual factors also affect expectancy, a person's belief concerning the likelihood of achieving a particular level of performance if they tried their best. Valence and instrumentality combine multiplicatively to determine performance valence. Performance valence is a hypothetical construct that represents the anticipated satisfaction of performing at a given level of performance. The anticipated satisfaction for a given performance level is derived from its degree of association with particular job outcomes and the valence of those job outcomes to the individual. Performance valence combines multiplicatively with expectancy. This product becomes the perception of attractiveness for each performance level. In essence, each possible level of task performance acquires valence through its association with certain job outcomes and the anticipated satisfaction associated with these outcomes. This performance valence is then modified by a person's oelief concerning the likelihood of achieving that level of performance given his or her best effort. The result is a perception of attractiveness for each level of performance.

The model specifies that goal choice is based on a person's evaluation of the relative attractiveness of various performance levels. The model is flexible in that it accommodates alternative decision strategies (e.g., return on effort, maximization, value matching). In this study the return on effort approach, which assumes that people use an incremental decision rule in choosing a goal, was used in determining the goal choice prediction. With this approach the model would predict goal choice to be some measure reflecting the marginal gain in the attractiveness of performance for performing at a particular level. While this approach has been successfully employed to improve expectancy theory predictions of performance (Kopelman, 1977), empirical evidence is lacking concerning the relative accuracy of alternative decision strategies. Additional work is needed to determine whether the return on effort approach offers the best representation of the goal choice process.

The remainder of the model describes the process by which performance goals are translated into work motivation, and the work motivation into task performance. The hypothetical relationship of these concepts and the process by which goals are translated into performance is based on the broad theoretical position that Performance (P) equals the product of Ability (A) and Motivation (M); (P = A X M).

Though cognitions serve telic purposes, they are influenced by past behavior and experience. The model specifies feedback loops suggesting that a person's effort-performance expectancy is influenced by past expenditures of effort and performance. Also, past performance affects both a person's objective ability and their subjective estimate of their ability. These factors, in turn, affect future effort and performance. The model is dynamic in that the source of purposive action is cognitive activity, though not necessarily conscious, that is influenced by past action and its consequences.

Method

Subjects

One-hundred and thirty experimental subjects participated in this study. Their average age was 21 years. Seventy-one of the subjects were female and 59 were male. Approximately 40% were high school students and the remaining 60% were undergraduate college students. Some data for six subjects were missing and therefore were unavailable for some of the analyses.

Procedure

The present study was part of a larger work simulation study designed to investigate the effects of alternative incentive magnitudes on performance (Riedel, Nebeker, & Cooper, 1985).

Subjects were recruited for part-time employment to perform a clerical transfer task. The 130 subjects who qualified for the job were assigned randomly to 1 of 7 experimental conditions differing in terms of the magnitude of incentive offered for various levels of performance. They worked 5 days, 4 hours a day, for a total of 20 hours at a rate of $4.40 per hour.

Research questionnaires were administered three times: after assignment to an experimental condition, at the start of the third day, and at the start of the fifth day. These questionnaires contained the expectancy and goal items needed for evaluating the model predictions. The quality and quantity of performance was recorded daily. A detailed description of the experimental procedure, treatment conditions, constructs and measures, and method of wage and incentive payment can be obtained from the author.

Results

Manipulation Check

Incentives and performance. It was expected that subjects in the incentive conditions would perform better than subjects in the nonincentive groups. To test this hypothesis, an analysis of variance was performed with treatment condition as the dependent variable.

The results of this analysis suggest a significant treatment effect, $F(6,120) = 3.27$, $p < .005$. A planned comparison of the performance means for the incentive and nonincentive groups revealed a significant difference, $t(120) = 3.87$, $p < .001$.

Incentives and instrumentality. Subjects were asked the amount of pay they expected to receive if they were to perform at alternative levels of performance. Judging from the responses, the performance-pay relationship was accurately perceived by most subjects across the treatment groups. For all conditions the reported pay instrumentalities approximate the actual relationships between pay and performance.

Model Predictions

The central research question pertained to the capacity of the model to account for the process of goal choice and task performance. To evaluate the model, goal choice (level) was predicted by the model, using a return on effort decision algorithm. This prediction was compared with self-reported goal choice. Also, the ability of the model to predict performance was evaluated by correlating the predicted performance with actual task performance. The results of these analyses are summarized below, first for goal choice and then task performance.

Goal choice. Prior to evaluating the goal choice prediction, responses to the self-report goal choice question were examined. Twenty-six subjects selected a single quantitative goal (e.g., 5 units per hour), 39 subjects selected a quantitative goal with a range (e.g., 5-7 units per hour), and 60 subjects set non-quantitative goals (e.g., to do my best). To increase the sample size for the single quantitative goal category, single goals for subjects with quantitative range goals were computed by averaging the upper and lower anchors of their range goal response. This resulted in 65 cases with a single quantitative production goal.

The model prediction of goal choice correlated significantly with self-reported goal choice ($r = .51$) and task performance ($r = .50$), both significant ($p < .001$). It was of interest to determine if the relationship between this goal measure and performance maintained for the entire sample, including people without quantitative goals. For the entire sample the correlation between the predicted goal choice and performance is ($r = .54$) and for people without quantitative goals ($r = .57$), both significant ($p < .001$). Also, the results of a t test indicate no significant mean difference in performance between those people who set quantitative goals and those people who did not set a quantitative goal. It appears that the goal choice prediction from the research model, based on the return on effort algorithm, relates highly to self-reported goal choice. The model also provides a significant prediction of performance, regardless of whether a subject reported setting a quantitative goal.

Performance. The capacity of the model to predict performance was evaluated by correlating the predicted performance with actual task performance. The model prediction of performance was significantly correlated with actual performance on the task ($r = .46$, $p < .001$). While this prediction was slightly better for subjects who set quantitative goals ($r = .55$, $p < .001$) than subjects who did not ($r = .39$, $p < .001$), the difference between these correlations was not significant. These findings provide preliminary support for the validity of the research model in predicting performance.

Discussion

The central purpose of this study was to improve our understanding of the process of goal choice. Overall the goal choice process specified in the research model was supported by the findings. First, monetary incentives were found to influence pay instrumentality. Second the cognitive components of goal choice which were specified in the research model predicted self-reported goals and performance, suggesting that the process of goal choice may be linked to expectancy theory concepts and processes.

The effect of incentives on pay instrumentality indicates that the treatment affected individual perceptions about the amount of pay associated with alternative levels of performance. Results show that the pay instrumentalities approximate the actual relationships between pay and performance, indicating that the pay contingencies were perceived quite accurately. In terms of the model, the effect of the treatment was to increase instrumentality and thereby increase performance valence, the anticipated satisfaction of performing at a given level of performance. The performance valence for a given performance level is derived from its degree of association with particular job outcomes and the valence of those job outcomes to the individual. It can be concluded that the experimental treatment was very effective in strengthening this association.

The goal choice process specified in the research model was supported by the findings, suggesting that expectancy theory concepts may be useful in understanding the cognitive components of goal choice. The combination of the expectancy constructs produced a reasonably accurate prediction of goal choice. The predicted goal choice was significantly correlated with the actual self reported goal. This finding suggests the interpretation that goal choice is a cognitive process where an individual chooses, from alternative performance goal levels, the level perceived to be most attractive. This perception of attractiveness is based on various beliefs and feelings a person has regarding the likelihood that performing at certain levels will lead to particular job outcomes. The results indicate that contextual factors, in this case the opportunity to earn monetary incentives for good performance, influence these beliefs and feelings.

This study has contributed to a better understanding of the relationship between organizational context and goal setting as they relate to work motivation and performance. The findings suggest that the process of goal choice is central to understanding how contextual variables influence goals, motivation, and performance. Moreover, the research model provides a useful starting point for investigating the relationships between organizational context and employee cognitions and perhaps for integrating goal setting with expectancy theory.

References

Crawford, K. S. (1982). *Goal setting with industrial workers: The impact of contexteal factors.* Unpublished doctoral dissertation, University of California, Irvine, California.

Dachler, H. P., & Mobley, W. H. (1973). Construct validation of an instrumentality-expectancy-task-goal model of work motivation: Some theoretical boundary conditions. *Journal of Applied Psychology*, *58*, 397-418.

Frank, J. D. (1941). Recent studies of the level of aspiration. *Psychological Bulletin*, *38*, 218-226.

Hilgard, E. R. (1958). Success in relation to level of aspiration. In C. L. Stacy and M. F. DeMartino (Eds.), Understanding human motivation (pp. 235-241). Cleveland: Howard Allen. (Original work published 1942).

Kopelman, R. E. (1977). Concepts, theories and techniques across-individual, within-individual and return on effort versions of expectancy theory. *Decision Sciences*, *8*, 651-662.

Lawler, E. E., III. (1981). Pay and organization development. Reading, MA: Addison-Wesley.

Locke, E. A., Feren, D. B., McCaleb, V. M., Shaw, K. N., & Denny, A. T. (1980). The relative effectiveness of four methods of motivating employee performance. In K. D. Duncan, M. M. Gruneberg, & D. Wallis (Eds.), *Changes in working life: Proceedings of the NATO International Conference* (pp. 1121-1157). London: Wiley.

Mento, A. J., Cartledge, N. D., & Locke, E. A. (1980). Maryland vs. Michigan vs. Minnesota: Another look at the relationship between expectancy and goal difficulty to task performance. *Organizational Behavior and Human Performance*, *25*, 419-440.

Opsahl, R. L., & Dunnette, M. D. (1966). The role of financial compensation in industrial motivation. *Psychological Bulletin*, *66*, 94-118.

Riedel, J. A., Nebeker, D. M., & Cooper, B. L. (October 1985). *The influence of monetary incentives on goal choice, goal commitment and task performance* (NPRDC Tech. Rep. In review). San Diego: Navy Personnel Research and Development Center.

Steers, R. M. (1975). Task-goal attributes, *n* achievement, and supervisory performance. *Organizational Behavior and Human Performance*, *13*, 392-403.

Steers, R. M., & Porter, L. W. (1979). *Motivation and work behavior*. New York: McGraw-Hill.

Vroom, V. H. (1964). *Work and motivation*. New York: Wiley.

The views expressed in this paper are those of the authors and does not necessarily reflect the views of the Department of the Navy.

THE EFFECTS OF REWARD MAGNITUDE AND DIFFICULTY OF PERFORMANCE STANDARDS UPON INDIVIDUAL PRODUCTIVITY

Delbert M. Nebeker[1]

Navy Personnel Research and Development Center
San Diego, CA 92152-6800

Management systems that control productivity and performance are critical to the success of organizations. One form of control system known as a Performance Contingent Reward System (PCRS), is receiving increasing attention in the U.S., particularly when it involves financial incentives. This is partly a function of the critical productivity problem we face in this country, and partly because recent evidence shows financial incentives to have a strong positive impact on performance (e.g. Nebeker, Neuberger, 1985; Locke, Feren, McCaleb, Shaw & Denny, 1980). In spite of this evidence the use of financial incentives as a means to increase productivity remains a controversial issue (Belcher, 1974; Lawler, 1983). If financial incentives are to be used effectively as a means to improve worker efficiency, we must have a better understanding of how they operate. We need to know how reward systems should be designed to maximize their value.

The design of reward systems can vary along a number of different dimensions. Theses include the following:

1. Objectivity of performance measure. The degree to which the performance measure is measured objectively as opposed to subjectively.

2. Performance aggregation level. The number of people include in the performance measure who share a reward.

3. Performance standard. The difficulty of the performance level required to earn a reward.

4. Sharing rate. The percent of earnings "saved" by performing above standard that is given as a reward.

5. Performance period. The length of time that performance data is accumulated before a reward determination is made.

6. Feedback type. The method of providing performance feedback.

7. Feedback period. The length of time between performance feedback.

[1] The views expressed in this paper are those of the author and are not official and are not necessarily those of the Department of the Navy

8. Performance-reward function. The shape of the function relating reward to performance.

9. Incentive period. The length of time following the performance period before payment is made.

Little is known about the optimal values for these parameters in various work situations. Virtually no empirical research directly addresses these issues in work environments. Theory is only slightly more helpful. If we were going to get answers about the optimal values for these parameters we were going to have conduct some parametric studies. Our field experience had suggested that two of the more important parameters in the design of reward systems were standard difficulty and reward magnitude.

In reviewing the literature we found the following: There is disagreement over whether performance standards ought to be made difficult to reach (Locke, et al., 1981, Barnes, 1980) or attainable for most workers and therefore easy for many (Peters & Waterman, 1982, Motowidlo, et al., 1978). Evidence for both points of view can be cited. One possible reason for the apparent contradiction is that the effects of the reward magnitude have not been adequately considered in research on standards or goal difficulty. It is quite likely then, that the affects of easy or difficult performance standards are moderated by the magnitude and/or the attractiveness of the rewards available for reaching and exceeding these standards (Matsui, Okada & Mizuguchi, 1981).

The instrumental learning and conditioning literature (c.f. Logan, 1970,p.90-91) posits that increasing magnitudes of reward have "diminishing returns" on performance (at least for rats). This suggests that very large rewards are likely to have a lower marginal utility than moderate rewards. Some of our own preliminary research supports this contention with people at work. Practical applications of reward systems in real organizations, however, show wide variability in the amounts of reward offered for performance above standard. Examples in business and industry can be found with sharing rates ranging from 10% to over 100%. It is reasonable to assume then that systems designed to pay very large rewards are likely to produce marginally less improvement and be less cost effective than systems that pay more moderate amounts.

The present research was designed to help us understand the interactive effects of these two variables by exploring the joint effects on productivity of varying degrees of reward magnitude, as defined by sharing rate, and standard difficulty. Furthermore, it is expected that worker ability will affect the relationship between standard difficulty and reward magnitude.

Method

Research setting

This research was conducted in the Organizational Systems Simulation Lab (OSSLAB) at the Navy Personnel Research and Development Center. The OSSLAB is designed to create a high fidelity simulation of real computerized work environments. Microcomputers are used as workstations so that individuals can be hired to do real work tasks under experimentally controlled conditions.

Subjects and design

Twenty-four employees (8 males,16 females) were recruited and hired (at $4.89 per hour) to provide technical support to the Navy Personnel Research and Development Center. Their job was to enter and maintain references in a data base for searching and retrieving the scientific literature. The Ss were required to be keyboard proficient before being hired. They worked two 4-hour shifts a week for eight weeks. A Work sample test given to measure ability revealed no significance differences between the two shifts.

The research design called for the Ss to perform their work under three reward conditions: (1) Baseline or control; (2) small incentives, wherein 15% of the wages saved by performing above standard were paid to the employee as a bonus and; (3) large incentives, wherein 50% of the wages saved by performing above standard were paid to the employee as a bonus. Furthermore, the shifts were designated as either the easy standard group or the difficult standard group. The easy standard group had a performance goal or standard set at the 20th percentile of the group's baseline performance level. This meant that 80 percent of the group was already exceeding the standard when the incentives were introduced. The difficult standard group had their performance goal, or standard, set at the 90th percentile. This meant that only 10 percent of the group were exceeding the standard when the first incentive was introduced. These values were chosen to match two interesting findings. Barnes (1980) demonstrates that standards set by usual industrial engineering methods typically produce standards that only 10% of the people exceed without performance standards and feedback or incentives. The difficult standard shift was designed to match this condition. The easy standard was chosen to match the finding that 80% of workers believe they are performing above average. Thus the standard would be consistent with there own self-concept.

The final design was a 2 X 3 factorial design where standard difficulty level was a between-subject factor and incentive level was a within-subject factor.

Procedure

On the first day of work the Ss reported to the job site and were welcomed by their first and second level supervisors. They were given an overview of the work they would be doing with an emphasis being placed upon the value of their job and how important accuracy and quality were to the information system a long with speed.

They were then trained on the use of their work stations, (IBM PC-XT microcomputers). This included the use of both hardware and software. Upon completion of this training their shift was completed for the day and they were excused to leave. On the second day they were given additional training on the task, allowed several practice items and then given the work sample test. Following the work sample they began their assigned work of entering and maintaining the references in the data base. The remainder of their employment consisted of their performing the task each workday. The only variations in this schedule were: (1) the introduction of the small incentive during the third week and the introduction of the large incentive during the sixth week; and (2) the administration of work perception questionnaires on four occasions spread throughout their eight weeks of employment; (3) the readministration of the work sample test at the end of the second week.

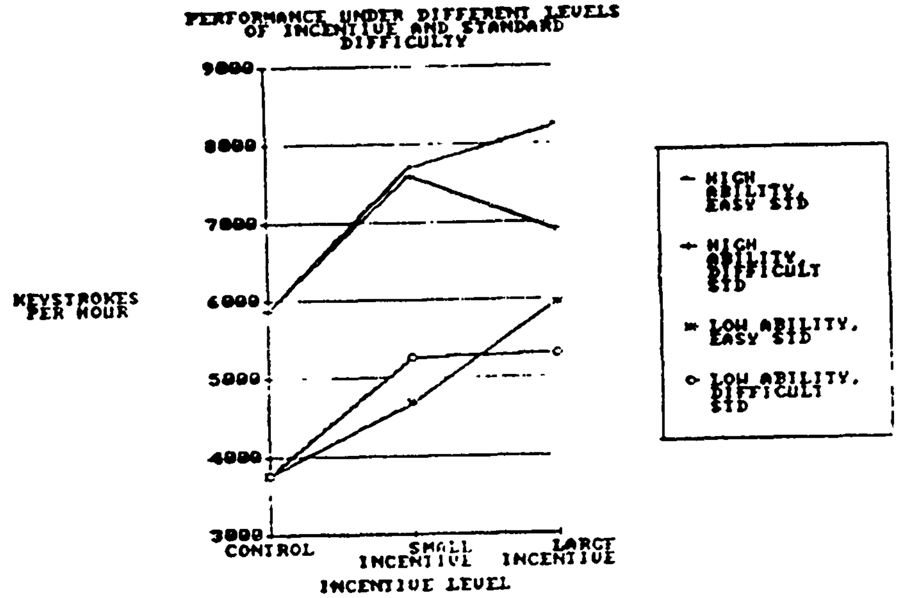
Performance was measured by keystroke rate, the number of keystrokes per hour. At any time during the experiment the workers could chose to view, on their screens, one of several reports of their current and "to-date" performance. During the baseline/control condition, prior to the introduction of the incentives, these reports included only raw performance information such as keystroke rate, hours on the tasks and regular pay. Following the introduction of the incentives, however, the reports added: (1) A listing of standards and the workers current and to-date performance efficiency against these standards (e.g. keystrokes per hour/standard keystrokes per hour); (2) the current and to-date bonus earned for exceeding the performance standards; (3) current and to-date total earnings

Results and Discussion

Daily performance means were used as the chief dependent variable in a series of moderated multiple-regression equations. In these equations the ability score was entered first, as a subject-covariate factor; then the the dummy coded treatment main effects; followed by the two-way interactions and; finally the three-way interactions.

The results of these analyses reveal a highly significant multiple R at each step and an overall R=.89 (R^2=.79; p<.001). In this regression three significant predictors of keystroke rate were found. Ability, as measured by the work sample test, accounted for a large portion of the performance variance (Beta=.583;*F*=42.42;*df*=11/202; *p*<.001). Also contributing significant portions of the variance to the prediction of performance were the interaction of standard difficulty with the large incentive manipulation, (Beta=1.213;*F*=8.21;*df*=11/202;

p<.005) and a three-way interaction between ability, standard difficulty, and large incentives (Beta=-1.274;F=9.928;df=11/202; p<.005). The relationships produced from these effects are shown in the figure below.



As can be seen the independent variables and their interactions produce some interesting effects on keystroke rate. First, it is obvious that there are large ability differences in performance and interaction effects for incentive level, standard and ability. High ability workers perform significantly better in all conditions than low ability workers. Second, when performance standards are relatively easy, increases in incentives result in performance increases for all ability levels at each level of incentive. When the standards are difficult, however,_ performance for_the_high ability workers, first increases for the small incentives and then <u>decreases</u>, when the large incentives are introduced. Finally, with small incentives and the difficult standards, the low ability workers perform better than their counterparts with easy standards. With the introduction of the large incentives, however, their performance, does not improve, while the easy standard group's does, allowing the easy standard group to out- perform the difficult standard group.

From these results the following points appear clear: (1) Performance improves substantially with the introduction of both standards and incentives regardless of whether or not the standards are difficult or easy and whether or not the incentives are large or small. (2) It appears that the amount of improvement with easy or difficult standards depends upon whether or not small or large incentives are being offered for exceeding them. The best performance, contrary to Locke et al.'s position (1981), occurs when the standards are at the 20th percentile and the sharing rate is 50%. (3) High ability worker's behavior under conditions of both high standards and large incentives seem to be either discouraged by the opportunity to earn incentives, or discount the value of increasing performance and actually reduce their performance. Either of these explanations requires a view of small incentives as providing only an achievement or competence motive for reaching the goal. When large incentives are introduced, however, financial motives for reaching the goal may become dominant. Under these conditions the question of discouragement and/or relative value may be significant factors affecting performance. Which of these explanations (or some other possibility) is superior remains for future research to determine.

References

Barnes, R. M. (1980). Motion and time study: Design and measurement of work. New York: Wiley.

Belcher, D. W. (1974). Compensation administration. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Lawler, E. E., III. (1983). What ever happened to incentive pay? (Tech. Rep. G83-6(37)). Los Angeles: University of Southern California, School of Business Administration.

Locke, E. A., Feren, D. B., McCaleb, V. M., Shaw, K. N., & Denny, A. T. (1980). The relative effectiveness of four methods of motivating employee performance. In K. D. Duncan, M. M. Gruneberg, & D. Wallis (Eds.), Changes in working life: Proceedings of the NATO International Conference (pp. 1121-1157). London: Wiley.

Logan, F. A. (1970). Fundamentals of learning and motivation. Dubuque, IA: Wm. C. Brown.

Matsui, T., Okada, A., & Mizuguchi, R. (1981). Expectancy theory prediction of the goal theory postulate, "the harder the goals, the higher the performance." Journal of Applied Psychology, 66, 54-58.

Motowidlo, S. J., Loehr, V., & Dunnette, M. D. (1978). A labor..tory study of the effects of goal specificity on the relationship between probability of success and performance. Journal of Applied Psychology, 63, 172-179.

Nebeker, D. M., & Neuberger, B. M. (1985). Productivity improvement in a purchasing division: The impact of a performance contingent reward system. Evaluation and Program Planning, 8,121-134.

Peters, T. J., & Waterman, R. H., Jr. (1982). In search of excellence: Lessons from america's best-run companies. New York: Harper & Rowe.

Results of an External Evaluation System Development in the Canadian Forces

Commander R.H. Kerr CF
and
Mr. Duane Tyerman

Background

The External Evaluation Process as part of the ISD (Instructional Systems Development) model is under active study within several commands in the CF (Canadian Forces). One such study, which is still ongoing, was described in MTA Proceedings (Kerr et al, 1984); this short addendum reports results and analysis of the findings in the studies involving 79 Dental Technicians (graduates and supervisors) and 467 Administrative Clerks (graduates and supervisors). Action proposed and underway is discussed in the last portion of this report.

System Description

The system employed a mail-out questionnaire approach (individuals also responded by mail). Question formats required the graduates and supervisors to assess task completion in accordance with a question grid providing for eight possible response patterns per task. Both trades were assessed on 40-50 tasks. Response sheets were manually inout into computer files and subsequently edited and analysed using special programmes designed specifically to be 'user digestible'. Details of the formats and outputs are described in the previous paper.

Results - Input Analysis

Ninety-eight percent of the questionnaires were returned and 42% of the response sheets contained recording errors. Further analysis resulted in only 18% of task responses being spoiled out of a total of 43,680. Although an 18% spoilage rate is considered acceptable by the authors, revision to the input formats have been made and considerable improvement is expected in reducing recording errors. A contract has been awarded to a research firm to utilize optical mark reader equipment incorporating the editing function. Because of the relative uniqueness of the input format, and the multi-track aspects of response patterns possible in such a paper-based system, no other option appears available in a 'mail-out' setting. The majority of personnel responding found the format easy or very easy to complete and spent a mean time of 32 minutes completing the questionnaire.

As indicated in the previous paper, test-retest reliability was attempted with a sample of 60 of the original population, by a readministration of the same instrument one month later. After editing, 74% of responses were identical to the original administration of those sampled. Because of the multi-track aspects of response possibilities and because some response change is expected (e.g. personnel may now have performed a task not performed at the first administration) this response is taken to be reasonably reliable, and with form revision, line correlations should improve.

Processing and Outputs

Processing was described in the previous paper and outputs were generated and provided to Training Managers. The External Evaluation business could provide some earth-shaking insights into training continua but not in these instances! Cases involving performance deficiencies beyond the limited level were virtually non-existent and those tasks reported as being not required were easily explained away as being infrequently performed. A good example was the completion of a casualty form by a junior administrative clerk. Infrequently completed tasks such as these are being examined for the degree of emphasis that these tasks have on course lengths. Further action may be taken to alter job specifications which may eventually effect the level at which the training is conducted. The authors are convinced that their next populations should be chosen to expose and quantify areas where some more serious problems are known to exist, which would enable more components in training continua to be examined.

Training Managers on the whole were pleased with and understood the outputs. The emphasis on job-orientation utilized in this system took much explaining, mainly because the audience for outputs were training managers, not performance technologists as discussed in the previous paper.

Further Analysis and Direction

The question of external evaluation in the CF has received renewed interest in 1985 at National Defence Headquarters and other Commands have instituted their own programmes and studies. The authors believe that External Evaluation as applied to Pilot Training may demand different emphasis than that required for Naval Technician training in that output analyses have different descriptors in order for alleviable action to take place in correcting performance deficiencies. The major demand appears to be in developing a system which will monitor and correct 'overtraining' in order to optimize training system efficiency. Apart from additional questions being asked in the field using an existing External Evaluation system to 'flag' a possible overtraining area, the authors at this stage feel that the internal evaluation and professional design processes would provide more impact in guaging and correcting overtraining. Experimental designs could then be utilized to verify, with the assistance of an External Evaluation System whether differing training methods or standards result in acceptable field performance. The authors have been tasked to examine this area in 1986.

Conclusion

This brief resumē of proceedings is intended to provide interim results on the development of an External Evaluation system within a CF context. Future developments using improved technological advances such as optical scanning techniques and use of video display terminals as inputs to external evaluation systems harbinge a heightened future for this often neglected process.

References:

Kerr, R.H. et al (1984). External Evaluation Revisited - an Experimental Update. Proceedings of the 26th Military Testing Association, 1984, 2, 763-768.

Occupational Learning Difficulty:
A Construct Validation Against Training Criteria

Joseph L. Weeks
Air Force Human Resources Laboratory

Michael D. Mumford
Georgia Institute of Technology

Francis D. Harding
Advanced Research Resources Organization

Occupational learning difficulty is defined as the time required to learn to satisfactorily perform occupational tasks. It is expressed in terms of a quantitative index which is produced on the basis of information obtained from structured job analysis questionnaires developed and administered by the United States Air Force Occupational Measurement Center. For any given job specialty, a learning difficulty (LD) index is derived by combining relative ratings of task learning difficulty obtained from senior-level technicians and benchmark ratings of task learning difficulty obtained from external occupational experts. The adjusted task ratings resulting from this combination are then weighted and aggregated to produce an occupational-level index of learning difficulty which can be meaningfully compared across job specialties. A more detailed description of the derivation procedure has been provided elsewhere (Weeks, 1984).

Once an LD index was available for most Air Force enlisted job specialties and its reliability and validity had been demonstrated, it served as a job-centered, frame-of-reference for various management decisions. In this context, there are both primary and special applications of learning difficulty information. Primary applications involve both personnel and training management.

For example, to the extent practical, the order of job aptitude requirement minimums are established so as to correspond to the order of job specialties in terms of learning difficulty. This application contributes to the optimal allocation of talent by ensuring that job specialties which are highest in learning difficulty are manned by enlistees having the highest aptitudes.

Learning difficulty indexes are also applied during the initial job-offer process. Air Force enlistees are assigned to job specialties at military entry processing stations on the basis of a computer-based, person-job match algorithm known as PROMIS (Hendrix, Ward, Pina, & Haney, 1979). One of the policies implemented by PROMIS is to offer the most difficult job specialties to the most talented enlistees. This process obviously requires information concerning enlistee aptitudes and information concerning job difficulty. Within the PROMIS system, job difficulty is defined in terms of the LD index.

Another primary application involves decisions concerning mode of training. Enlisted job specialties are designated as either category A, B, or C skills. For category A skills, the mode of training is formal, resident or school-house training. For category C skills, the mode of training is on-the-job training (OJT). For category B skills, the mode of training can be either formal resident training or OJT. The LD index is one of several inputs to the decision process associated with determining mode of training.

In addition to these primary applications, there have been noteworthy special applications. The LD index was used as an empirical basis for justifying Air Force personnel quality standards in response to inquiries by the House and Senate Armed Services Committees during the development of the 1985 Defense Appropriations Bill. Furthermore, it has been advanced as an empirical basis for Air Force job and training requirements during investigations by oversight committees such as the Government Accounting Office and the Air Force Audit Agency.

Problem

Because of the importance of such applications, the validity of the LD index is an issue of considerable interest. Burtch, Wissman, and Lipscomb (1980) were the first to seriously address this problem. Their approach consisted of correlating relative ratings of task learning difficulty by senior-level technicians with benchmark ratings by occupational experts. Observed correlations ranged from .54 to .96 for 100 different job specialties. These results provided evidence of the convergent validity of the benchmark ratings. Although this study was comprehensive in that separate analyses were conducted for several different job specialties, it cannot be considered sufficient by itself. Validation efforts must take into account the functional role of learning difficulty information in management applications. For all the applications previously described, the occupational-level index rather than task-level ratings of learning difficulty served as the referent for management decisions. Consequently, there appears to be a need to evaluate the validity of the occupational-level index of learning difficulty.

Method

Because the LD index is not applied to predict some criterion, construct validation is considered to be more appropriate than predictive validation. With construct validation, the goal is to evaluate the intrinsic meaning of some measure of interest. This is accomplished by evaluating both convergent and discriminant validity (Campbell & Fiske, 1959). Because the LD index is a measure of learning time, strong relationships with the training time for an initial-skills course associated with a specialty would be expected. Also, because the LD index is a measure of a job property, strong relationships with measures of personnel attributes would not be expected.

In an independent research project devoted to the development of a covariance-structure model of Air Force technical training, it was necessary to collect information concerning occupational learning difficulty as well as measures of student input, course content, and training outcome variables for several initial-skills courses. Therefore, it was decided to extend that effort to include an examination of the construct validity of the LD index. Only analyses relevant to the construct validation of the LD index will be discussed here. Details concerning the development of the covariance-structure model of initial-skills training are provided by Mumford, Weeks, Harding, and Fleishman (1985).

Samples of subjects and courses used for the present analyses were identical to those employed by Mumford et al. (1985). Courses were selected from among approximately 200 initial-skills courses administered by various technical training centers under Air Training Command. Courses were selected to provide a representative sample with respect to numerous criteria including course content, student flow, training costs, and aptitude area. Subjects were sampled so as to provide a minimum of 50 students for each course. These procedures provided a total of 5,970 students and 48 initial-skills courses.

The independent variable examined in the present study consisted of the LD index obtained for the job specialty associated with each of the 48 initial-skills courses. For each specialty, the index consisted of an aggregate value obtained by deriving the average learning difficulty of first-term positions in the specialty. This procedure was considered appropriate because initial-skills courses are designed to provide training for tasks likely to be encountered during the first term of service.

A wide variety of dependent variables falling into three broad categories were examined. The dependent variables included those which were expected to be highly related to the LD index as well as those which were not. For dependent variables in each category, detailed descriptions of the source of data and measurement process are provided by Mumford et al. (1985).

The first category of dependent variables consisted of measures of student inputs. These measures included (1) students' average scores on the composite of the Armed Services Vocational Aptitude Battery which serves as the basis of the aptitude requirement for entry into the course, (2) average reading grade level as measured by the Air Force Reading Abilities Test, (3) the average academic motivation of students in the course as indexed by the average number of difficult high school courses completed, (4) average educational level of students in the course, (5) educational preparation as indexed by the average number of recommended high school course prerequisites completed by students in the course, and (6) the average age of students in the course.

The second category of dependent variables consisted of measures which represent outcomes of training. These measures included (1) average final course grades for students in the course as indexed by the average score on end-of-block written tests, (2) average number of hours of special individualized assistance (SIA) provided students by course instructors, (3) the number of academic counseling sessions provided students, (4) the number of nonacademic counseling sessions, (5) washback time as indexed by the average number of retraining hours provided students, and (6, 7) the academic and nonacademic student attrition rates for the course.

The third category of dependent variables consisted of measures of properties of the initial-skills course and are described as course content variables. This category included (1) course length (in hours), (2) course diversity as reflected by the number of different units of instruction, (3) expert's average rating of the abstract knowledge requirement of the course,

(4) the expected student attrition rate for the course, (5) the number of students per instructor, (6) the average number of months of instructional experience of the instructors assigned to the course, (7) average ratings of the quality of instruction provided by course instructors, (8) manning requirements as reflected by the availability of a selective reenlistment bonus for the specialty associated with the course, (9) length of academic day (in hours), (10) the number of instructional aids employed per instructional hour, (11) the percentage of training hours devoted to hands-on instruction, (12) the frequency of formal feedback per hour of instruction, (13) practice or the number of hours devoted to a unified body of material, (14) student flow or total number of students passing through the course per year, and (15) the reading difficulty of course materials.

Analyses undertaken to examine the construct validity of the LD index were straightforward. Scores on the various student input, training outcome, and course content variables were correlated with the LD index. With regard to analyses involving student input variables, it is important to note that the LD index for the job specialty associated with a course was assumed to be applicable to all students in the course. Once the correlational analyses were completed, the pattern of relationships indicated by the correlational data were examined to evaluate the discriminant and convergent validity of the LD index.

Results and Discussion

Table 1 presents the mean and standard deviation of each dependent variable and its correlation with the LD index. Examination of the student input variables indicates that all observed correlations were statistically insignificant. Because the LD index is a measure of a job property and the student input variables are measures of personnel attributes, this outcome was expected. The fact that this expectation was confirmed by the results of the present analyses lends support to the discriminant validity of the LD index.

All observed correlations for the training outcome variables were statistically insignificant. This outcome was also expected. As a result of efforts to develop the covariance-structure model of training, we have gained some insight into the complex interrelationships that combine to influence training outcomes. Largely because of the instructional systems design process which uses information concerning job tasks as one basis of course design, the LD index is conceived of as having a primal influence on course content. However, numerous student input and course content variables combine to influence training outcomes. Consequently, the relationships between the LD index and training outcomes are indirect being moderated by several other variables.

The strongest relationships produced by the LD index were with various course content variables. This was expected because as previously indicated, the design of initial-skills training is guided by job content. For instance, because the LD index represents learning time, it was expected that a positive relationship would exist between the LD index and course length. The fact that a moderate positive correlation was observed for these two variables tends to argue for the convergent validity of the index. Moreover, it would be expected that the LD index would be positively related to indices of course subject matter difficulty to the extent that it is an intrinsically meaningful index. The moderate positive relationships observed between the LD index and course diversity, abstract knowledge requirement and expected attrition rate support this expectation.

In addition to these relatively direct relationships, it was expected that the LD index would yield a number of more diffuse relationships. For instance, it might be expected that fewer students per instructor and more experienced instructors would be a means of compensating for task learning difficulty. The observed relationships between the LD index and these two variables lends support to this expectation. Further, it might be expected that the overall quality of instruction would be low for tasks of high learning difficulty. Again, the observed relationship between these two variables supports this expectation. The positive relationship between the LD index and manning requirements may be attributed to the fact that specialties higher in learning difficulty generally require training which is highly valued in the civilian work place. The loss of military-trained technicians to the civilian labor market would, in turn, lead to a greater demand for personnel.

Table 1. Means and Standard Deviations of Dependent Variables and Bivariate Correlations with the Learning Difficulty Index

VARIABLES	MEAN	STANDARD DEVIATION	r
Student inputs (N = 5,970 students)			
Aptitude Composite Score	68.800	16.30	.046 NS
Reading Grade Level	11.400	1.00	.064 NS
Academic Motivation	39.000	13.40	.069 NS
Education Level	2.180	.45	-.012 NS
Educational Preparation	1.670	.94	-.115 NS
Age	20.100	2.20	.035 NS
Training Outcomes (K = 48 courses)			
Final Course Grade	85.200	7.61	.076 NS
SIA Time	6.560	15.40	.048 NS
Academic Counseling	1.450	3.47	.043 NS
Nonacademic Counseling	.170	1.56	-.094 NS
Washback Time	11.100	51.50	.066 NS
Academic Attrition	.028	.16	.014 NS
Nonacademic Attrition	.004	.06	-.002 NS
Course Content (K = 48 courses)			
Course Length	413.900	309.30	.501 **
Course Diversity	54.800	43.30	.536 **
Abstract Knowledge Requirement	2.420	.98	.467 **
Expected Attrition Rate	.096	.03	.450 **
Student-Instructor Ratio	9.100	4.80	-.556 **
Instructor Experience	32.500	14.70	.591 **
Instructor Quality	2.500	.15	-.314 *
Manning Requirements	.440	.49	.423 **
Day Length	.480	.44	-.033 NS
Instructional Aids	.270	.10	.161 NS
Hands-on Practice	.410	.13	-.186 NS
Frequency of Feedback	.340	.12	-.191 NS
Practice	8.520	3.17	-.074 NS
Yearly Student Flow	1943.600	2662.50	-.200 NS
Reading Difficulty	10.980	.63	.031 NS

** - Probability ≤ .01 that observed r is a random deviation from a population r of zero.

* - Probability ≤ .05 that observed r is a random deviation from a population r of zero.

NS - Probability ≥ .06 that observed r is a random deviation from a population r of zero.

Conclusion

In reviewing the results of these analyses, it seems reasonable to conclude that the occupational learning difficulty index displays construct validity. The index was found to be moderately related to both course length and a number of indices of course subject matter difficulty lending support to the convergent validity of the index. Moreover, the finding that strong relationships did not exist between the LD index and a number of student input and training outcome variables tends to support the discriminant validity of the index. Overall, the evidence produced by these analyses appears to indicate satisfactory construct validity for the occupational-level index of learning difficulty.

References

Burtch, L. D., Lipscomb, M. S., & Wissman, D. J. (1982, January). Aptitude Requirements Based on Task Difficulty: Methodology for Evaluation (AFHRL-TR-81-34). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. Psychological Bulletin, Vol. 56, pp. 81-105.

Hendrix, W. H., Ward, J. H. Jr., Pina, M. Jr., & Haney, D. L. (1979 September). Pre-enlistment Person-Job Match System (AFHRL-TR-79-29). Brooks AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory. (AD-A078 427).

Mumford, M. D., Weeks, J. L., Harding, F. D., & Fleishman, E. A. (1985). An Empirical System for Assessing the Impact of Aptitude Requirement Adjustments on Air Force Initial-Skills Training (AFHRL-TR-). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. (In press)

Weeks, J. L. (1984, November). Occupational Learning Difficulty: A standard for determining the order of aptitude requirement minimums. (AFHRL-SR-84-26). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. (A417 410).

News about CAT in the German Federal Armed Forces (GFAF)

Wolfgang Wildgrube
Psychological Service, Bonn, West Germany

Introduction

At the last annual MTA-conference 1984 in Munich, a workshop about "CAT in Germany" had presented the first activities and experiences used computerized testing in the GFAF (named: CAT I). Since 1983 the GFAF evaluated two experimental installations for computerized testing at the recruiting centers in Munich and Hannover choosing the practical approach by conventional testing (Wildgrube, 1985 b; see also the contributions of Angermüller and Kulling).

Meanwhile CAT made an important step forward in the GFAF. First a lot of paperwork to prepare computerized testing is accomplished so that in 1987 CAT systems can be operational at the four recruiting centers for volunteers and the recruiting center for officer candidates compiling the conventional procedure to computer application. Second a major change is planned at the recruiting centers for draftees. The final goal will be to accomplish medical examination and psychological testing - supplemented by psychological counseling - at the same day so that each draftee knows, after one day at the recruiting center, the date, location, and unit for the time in the service respectively for his beginning basic training. Therefore individualised testing by computer is necessary for this one day examination for draftees. The corresponding pilot project started in January 1985 in Hannover using one of the CAT installations.

Aptitude Classification Battery

In the GFAF the Aptitude Classification Battery (EVT - German abbreviation) is in use as the standard entrance examination for draftees and volunteers (similar the ASVAB; officer candidates have a special test battery). Parallel to the CAT developments goes a major revision of the EVT battery, so that the GFAF starts in January 1986 with the following revised test battery:

Figure Matrices Test (FMT)	20 Items/8 Alternatives	18 Min.
Word Relation Test (WBT)	20 Items/5 Alternatives	4 1/2 Min.
Arithmetic Reasoning Test (RT)	20 Items/Input of the Results (paper/pencil for notices)	14 Min.
Spelling (Orthographical Test; RST)	60 Items/4 Alternatives	12 1/2 Min.
Mechanical Ability Test (MT)	20 Items/5 Alternatives	13 Min.
Electrotechnical Comprehension Test (EKT)	20 Items/5 Alternatives (Pretest after RT: 8 Items	20 Min. 6 Min.)
Reaction Test (RP)	64 Items/6 Alternatives Input of the Results	3,41 Min.
Radio Test (FT)	150 Items/3 Alternatives Input of the Results	3,30 Min.
Signal Test (SigT)	18 Items/4 Alternatives Input of the Results	3 Min.
Doppler Test (DopT)	20 Items/3 Alternatives Input of the Results	4 Min.

After the six conventional subtests, up to now carried out by paper pencil, follow four special tests presented by maschines ("apparative" tests). This fixed sequence of subtests will be used during the routine application of the EVT by computer. Changes are possible at any time by interrupts of the proctor, for example omitting of the signal test or stopping after the electrotechnical comprehension test.

New Hard- and Software

Concerning the rapid developments and changes in the area of Personal Computer the definition of new hardware and furthermore, software for CAT was necessary (named: CAT II). Besides the change from PC with 8 bit processors to PC with 16 bit processors the new equipment will contain as an expansion of CAT I further the four "apparative" tests, namely reaction test, signal test, doppler test, and radio test, so that the whole Aptitude Classification Battery can be presented by computer.

There was conformity about that standard hardware and software were not sufficient for CAT at fixed locations/recruiting centers in the GFAF. At the end of the last year a booklet with the detailed requirements was worked out in cooperation between the GFAF and the German firm ZAK. The new hard- and software will be delivered in December 1985. One installation of the new equipment is assigned for the recruiting center of

draftees in Munich, the other one is a twin-configuration for the recruiting center in Hildesheim which should be delivered in Spring 1986.
In addition to the actual developments by the firm ZAK various firms present at the moment different concepts and corresponding financial estimations to realise the requirements of the GFAF concerning CAT for example in fifty recruiting centers for draftees.

The new CAT equipment has the following characteristics:

- A local area network (LAN) is in use with 15 work stations for examenees applicable different test batteries at each station.
- The testing session will be monitored at a central place by an IBM AT with harddisk. A second IBM AT - also in the LAN linked - is located in a seperate room and will be used for input of personnel/biographical data and for output of results and furthermore the second central place is prepared as back up for the master.
- Back up-arrangements are prepared at different levels, e.g. the central places for monitoring (2 IBM AT), the work stations (15 stations available), power interrupt (additional equipment and software tools), test results (internal twin storage at disk). There is no break allowed longer than 30 minutes as well as the loss or the repeat of a whole subtest.
- A work station can be described by these characteristics:
  + a white screen with black items, 768 x 1024 pixels;
  + a headset for voice output;
  + a special keyboard with a part for the compiled paper pencil tests (ten digits and the green and red function keys) and with four special parts for the "apparative" tests (e.g. reaction test).
- At the central place of the CAT station monitors the proctor the test session, using different menues. A maximum of four tasks are active so that an overview about the state of testing at any time is possible.
- The second proctor sits in the seperate room, types in the personnel data (at the moment the system provides 18 menues), or starts the output by a matrix printer (the different output is tailored for the requirements in the recruiting center, at subtest-, person- or item-level). The reference criterion for all data is the personnel-id-number (similar the social security number) so that all data (personnel data, biographical data, test data/results) are stored in a data base at the end of the whole psychological examination. The possibilities for transfer to a mainframe computer is considered, at subtest-level online and at item-level via tapedrive for follow up studies.
- The testing session starts for each examenee with an introduction or learning phase. There is the chance to get practice in the use of the keyboard and to learn the kind of testing. If an example solved or typed in incorrectly the program presents the example item once more. This phase and the example items before each subtest are supported by voice

output via headset. The whole text (without the items within the subtests) is prepared by deltamodulation, stored at the harddisk at each work station, and then monitored (screen and voice) by program during the test session.

CAT Center

At the recruiting centers the testing station is only installed for the application of the tests. After given in the password the proctor starts the session using the different menues, while a special password is necessary for data handling or for use of utilities. All aspects of change of the items or the itempool or modification of the software or the testing procedures will be carried out in the CAT center which is located at the Federal Armed Forces Office in Bonn.
Until to the end of this year the following hardware will be delivered: 1 central place (IBM AT), 1 work station, printer, plotter (for graphical items), tape drive (for data transfer to mainframe at item level). Furthermore the essential software will be prepared for Bonn:

- Source codes of the whole application software.
- Compiler for Basic, FORTRAN, Pascal, C (the greatest part of the software is written in C).
- Tools used by the firm ZAK for the software developments.
- Item editor for the input of graphical/nongraphical items, instructions, voice modules, and for storage in an itempool respectively data base.
- Item editor for assembling different subtests, modifying testing procedures, inserting instructions, and creating new test batteries.
- Utilities for data handling and for the management of data in a base.
- Statistic software for simple analyses at the personal computer.

All changes and modifications are prepared in the CAT center in Bonn and the floppies containing the newest version are distributed to the recruiting centers for the daily routine testing. In addition the CAT center is the link between GFAF and firm so that all hardware troubles first are reported to the CAT center. So the CAT center will be the central place for all aspects of CAT in the GFAF concerning hardware, software, itempools, data of the examenees, as well as for the scientific evaluation.

Experiences and Results

The two pilot installations in Munich and Hannover (CAT I) are in use daily and the CAT results are applied to selection and classification decisions. Now the data transfer is available from CAT to mainframe computer via tape drive for different statistical analyses.
Besides the computerized testing information has been collected by an ad-hoc-questionnaire. Here are some results from selected questions collected in Hannover (also classified by the four educational levels) and Munich.

	Hannover (N=223) level of education 1	2	3	4	T	Munich (N=175) T
1. When I entered the testing room and saw the testing equipment, I was curious as to what expected me						
1. I was very curious	2	53	78	32	165	136
2. I was not so curious	-	22	18	9	49	32
3. I was not curious at all	-	2	5	2	9	7
14. Comparing the computerized test version to the paper and pencil test, I think						
1. that I like the computerized form better	1	61	80	32	174	134
2. that I like the paper and pencil form better	1	11	15	8	35	33
15. I often play video games (e.g. "Star War")						
1. yes	1	14	17	8	40	39
2. no	1	59	80	31	171	129
16. I have experience with home computers						
1. yes	-	6	22	8	36	21
2. no	2	69	75	31	177	149
17. I participated in this test with pleasure						
1. if yes, why	2	70	92	34	198	145
2. if no, why not	-	5	3	4	12	18

The results indicate a high acceptance of computerized testing respectively this kind of non-group testing by computer. The examenees prefer the CAT application even if they have no experiences with home computers or video games. So will be soon computerized testing without any problems similar the conventional paper pencil testing.

CAT offers the chance for recording more data as well as paper pencil testing, so that more detailed analyses can be made. One aspect is the item solution time which is recorded for each item. But at time new models in testing theory and basic research are necessary to interpretate this time-based data in addition to the ability parameter.
An other important point concerns the difference in the test scores during a day period shown in the following table.

Test (time)	Scores			Time used per subtest (seconds)		
	morning	noon	P	morning	noon	P
WAT (11.00)	13.86	13.01	.002	234	246	.000
FDT (11.00)	14.46	13.76	.031	505	510	.504
RT (11.00)	9.99	9.13	.008	826	823	.690
MT (11.30)	11.59	10.70	.001	712	711	.890
RST (12.00)	28.10	26.00	.004	172	175	.016
EKT (12.00)	6.63	5.55	.000	552	561	.202

There are significant and relevant differences between the two groups (sample sizes approximately equal) in the scores of all six subtests. Remarkable are the significant values for the two subtests "word analogy" and "spelling" concerning the time used per subtest. The table above shows for the other subtests very similar time for solving the items in a subtest, while the scores are different. Further research is needed for this aspect so that different norms will be used for the morning and for the noon session if necessary.

Wildgrube, W. (1985, a): Computerized Testing in the German Federal Armed Forces (FAF): Empirical Approaches; in: Weiss, D.J. (Ed.) Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference, Computerized Adaptive Testing Laboratory, Department of Psychology, University of Minnesota, April 1985, p. 353 - 359.

Wildgrube, W. (1985, b): Computerized Adaptive Testing (CAT) in Germany - General Topics -; in: BMVg - P II 4 (Eds.), Proceedings - 26th Annual Conference of the Military Testing Association, München, Fed. Republic of Germany, 5 - 9 Nov. 1984, coordinated by the Psychological Service of the Federal Armed Forces - BMVg, P II 4, Bonn, 1985, p. 151 - 156.

EFFORTS TOWARDS THE IMPROVEMENT OF THE COMPUTERIZED ADAPTIVL SCREENING TEST (CAST)

Deirdre J. Knapp, Rebecca M. Pliske, and Richard M. Johnson
U.S. Army Research Institute for the
Behavioral and Social Sciences[1]

The Computerized Adaptive Screening Test (CAST) was designed by the Navy Personnel Research and Development Center (NPRDC) under the sponsorship of the Army Research Institute (ARI) to provide a prediction of prospects' Armed Forces Qualification Test (AFQT) scores at recruiting stations. This paper briefly discusses the development of CAST and summarizes current efforts to enhance its utility.

Background

Armed Forces applicants fall into certain mental categories on the basis of their AFQT scores. AFQT, which is intended to be a measure of trainability, is derived from a linear combination of subtest scores (i.e., WK, AR, PC, and one-half of NO) on the Armed Forces Vocational Aptitude Battery (ASVAB). In the Army, individuals who score at or above the 50th percentile (mental categories 1, 2, and 3A) are eligible for special options and benefits such as the 2-year Enlistment Option and the Army College Fund. Applicants who score between the 31st and 49th percentiles on AFQT (mental category 3B) qualify for enlistment but are not eligible for special options. Lastly, those individuals who score below the 31st percentile (mental categories 4A, 4B, 4C, and 5) are regarded as being low priority candidates for enlistment.

It is vital that recruiters have access to information which predicts prospects' AFQT performance for several reasons. For example, recruiter's missions specify not only the number of recruits to be enlisted, but also the quality of those recruits as determined by mental category classification. Further, if a prospect appears to have virtually no chance of producing an acceptable AFQT score, the recruiter may choose to discourage him or her from further interest in the Army. The recruiter can then spend more time selling the Army to more promising prospects. Finally, if a prospect appears to be of average quality, the recruiter may not want to spend much time describing special options and benefits to the individual. On the other hand, if the recruiter does not sell the options and benefits to those individuals who are likely to be eligible for them, he or she is failing to use a powerful sales tool. Clearly, recruiters can enhance their performance if they effectively use a valid predictor of prospects' AFQT performance.

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the view of the US Army Research Institute or the Department of the Army.

Description

As its name indicates, CAST is a computerized adaptive test. Adaptive tests are constructed so that they are tailored to fit each examinee. This is done by administering items which have optimal discriminability given a particular examinee's ability. This can be compared to traditional testing where all examinees respond to the same items, regardless of differences in examinee ability. Hence, adaptive tests are more efficient to use than traditional tests because a comparable amount of information can be gained from fewer test items.

The item pool for CAST was developed by researchers at the University of Minnesota (cf. Moreno, Wetzel, McBride, & Weiss, 1983) for use in the development of a computerized adaptive version of ASVAB. Researchers at NPRDC developed the software capable of administering items predictive of AFQT on the Army's microprocessor system known as JOIN.

The current operational version of CAST consists of 78 word knowledge (WK) items and 225 arithmetic reasoning (AR) items. All items are multiple choice with a maximum of five response alternatives. The items were developed using the three-parameter logistic ogive item response model (Birmbaum, 1968); thus each item has three parameters (discrimination, difficulty, and guessing) associated with it. Test items for CAST were chosen so that the discrimination parameter values would be greater than or equal to .78; the difficulty parameter values would range between +2 and -2; and the guessing parameter values would be less than or equal to .26. The ability estimates yielded by CAST are based on the Bayesian sequential scoring procedure discussed by Jensema (1977). The stopping rule is ten WK items and five AR items.

Validation Information

There are three validation efforts associated with CAST. The initial validation study was conducted at the Los Angeles Military Entrance Processing Station (MEPS) with a sample of 312 U.S. Army applicants (Sands & Gade, 1983). The correlation between CAST scores and AFQT scores was .85. The second data collection effort took place in Army recruiting stations in the midwestern region of the U.S. during the first two months of 1984 (Pliske, Gade, & Johnson, 1984). CAST scores were linked to subsequent AFQT performance via the social security numbers (SSN's) of the prospects. More specifically, recruiters recorded prospect SSN's thus allowing the researchers to locate the appropriate MEPS records. Matching records for 1,962 prospects were located and the resulting validity estimate was .80.

The most recent estimate of CAST's validity is based on data which is being collected from a sample of 60 Army recruiting stations during January through December of 1985. This sample was selected to be representative of the population of approximately 2,000 Army recruiting stations in terms of geographic location, population density, and ethnic composition. The correlation between CAST and subsequent AFQT performance, based on preliminary analyses of the first six month's of data, is comparable to those obtained in the earlier studies (r=.82; n=2,240).

Clearly, CAST is a valid predictor of AFQT performance. Of course, one problem with computerized adaptive testing is the need to have a computer standing by to administer it. Since recruiters do not always have this luxury, there remains the need to use the Enlistment Screening Test (ESI), a paper-and-pencil predictor of AFQT. The initial validation information regarding EST was provided by Mathews and Ree (1982). Although their data yielded a healthy validity estimate of .83, a cross-validation of the test has not been reported. Consequently, the recruiting stations which have been providing CAST validation data over the past year have also been asked to record the SSN's and EST scores of all prospects who ta^ke EST rather than CAST. Based on six month's of data, the validity estimate is .79 (n=685). As expected, the validity of ESI has been reaffirmed.

Proposed Improvements

Currently, efforts are underway to improve three specific aspects of CAST. The first aspect concerns the kind of information that the test provides to the recruiter. The second ana third aspects involve the test's item pools and stopping rule. At the present time, CAST provides bar charts that give information about performance on the WK and AR subtests and the examinee's predicted AFQT percentile score. There are two fundamental problems with providing only point predictions of AFQT scores to recruiters. The first problem is a function of the statistical naivete' of most recruiters. Because the great majority of recruiters do not understand the concept of correlation, they do not adequately understand the nature of the point prediction that they are given. Hence, recruiters complain that predicted and actual AFQT scores often fail to be exactly the same. A second problem with the use of point predictions concerns the way in which recruiters utilize information from CAST. As indicated earlier, recruiters are primarily concerned with the prospects' subsequent classification into one of three groups of mental aptitude categories. Given these considerations, it seems that recruiters would be better served if CAST provided output that reported the odds associated with a given prospect falling into each of the three critical mental categories.

Two approaches to category prediction are being studied using data from the on-going validation effort described above. The first approach models the strategy that recruiters probably follow. That is, one uses the point prediction, which is based on a regression model, to determine the mental category to which an individual will likely belong. The second approach is based on classification analysis. In contrast to regression analysis, classitication analysis provides subtest weights that optimize category, rather than point, predictions. Table 1 shows the percentage of cases which were classified into each of three categories on the basis of these two approaches. A comparison of the two approaches reveals that they differ with respect to where their prediction errors occur. Both approaches are good at identifying individuals who fall into categories 1-3A (Approximately 75% of the people who are predicted to be in 1-3A actually are in 1-3A). Classification analysis, however, is much better than regression analysis at identifying individuals who are in categories 4A and below (80% versus 55% accurate prediction). This advantage is at the expense of a somewhat poorer ability to identify prospects who are in category 3B.

TABLE 1

Percentage of Cases Classified
Into Critical ASVAB Categories

		PREDICTED AFQT CATEGORY*		
		1-3A	3B	4A AND BELOW
ASVAB	1-3A	76/77	23/18	1/5
AFQT	3B	25/27	60/38	15/35
CATEGORY	4A and below	4/4	41/16	55/80

* Regression analysis to left of diagonal; classification analysis to right of diagonal.

Regardless of which method of category prediction is used, the information provided can be presented in a way that would be logical to recruiters. For example, CAST software could be changed to report the probabilities associated with an examinee falling into each of the three critical categories. The recruiter could then compare the odds, and make an informed judgment concerning his or her subsequent course of action. The important point here is that presenting CAST results in such a fashion would make it clear to recruiters that, although those results can be very useful, they are not infallible.

Turning to the subject of CAST's stopping rule, the questions to be asked are twofold. First, assuming that the stopping rule is to be based on number of items, what is the optimal number of subtest items to administer? Second, would it be better to base the stopping rule on the precision of the ability estimate rather than on the number of items being administered? Data from the latest validation study have been used to examine the first question. The version of CAST which is being used in the 60 experimental recruiting stations administers five more items per subtest than the operational version of the test. It also records response time so that the time it takes to administer various subtest length combinations can be compared.

Validity estimates and average completion times were computed for six subtest length combinations (5, 10, and 15 WK items; 5 and 10 AR items). Validity coefficients ranged from .79 to .85. Completion times ranged from just over 10 minutes to just over 18 minutes. Given that the validity estimate associated with the current stopping rule is .82 and the average completion time is a little over 12 minutes, it appears that an increase in subtest length would not be justified. A substantial increase in completion time is required for even a small increase in validity.

With an adaptive test, an ability estimate and the variance associated with that estimate is computed each time an examinee answers a new test item. Rather than stop the subtests after a given number of items are administered, the test software can be altered to end the subtests once the variance estimate has dropped to a given value. The advantages and disadvantages of altering CAST to rely on the latter type of rule need to be evaluated carefully. Also, the determination of an optimal variance criterion would require further study.

The two item pools currently contained in CAST will be expanded within the next 2-3 years. Given the extensive use of the test, it is important that the item pools are large enough to prevent the frequent recurrance of particular test items. The possibility of developing items which will provide optimal discrimination at the critical AFQT cutpoints is also being considered.

Closing Remarks

CAST is a very good test which we are seeking to make even better. At the present time, our efforts are primarily aimed at changing the software to yield information that will be of the greatest use to recruiters. Our plans over the next couple of years are aimed at insuring the continued integrity of the test itself.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick (Eds) Statistical Theories of Mental Test Scores. Reading, Mass: Addison-Wesley.

Jensema, C.G. (1977). Bayesian tailored testing and the influence of item bank characteristics. Applied Psychological Measurement, 1, 111-120.

Mathews, J.J. & Ree, M.J. (1982). Enlistment Screening Test Forms 81a and 81b: Development and Calibration (AFHRL Report No. 81-54). Brooks Air Force Base, Texas: Air Force Human Resources Laboratory.

Moreno, K.E., Wetzel, C.D., McBride, J.R., & Weiss, D.J. (1983). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests (NPRDC) Report No. 83-27). San Diego, CA: Navy Personnel Research and Development Center. (NTIS No. ADA 131683).

Pliske, R.M., Gade, P.A., & Johnson, R.M. (1984). Cross-Validation of the Computerized Adaptive Screening Test (CAST). Alexandria, VA: U.S. Army Research Institute.

Sands, W.A., & Gade, P.A. (1983). An application of computerized adaptive testing in U.S. Army Recruiting. Journal of Computer-Based Instruction, 10, 87-89.

DEVELOPMENT OF A PORTABLE COMPUTERIZED PERFORMANCE TEST SYSTEM

R. S. Kennedy, W. P. Dunlap,
R. L. Wilkes, & N. E. Lane
ESSEX Corporation
Orlando Florida

Abstract

The ethics and pragmatics associated with developing an automated performance test system to study the effects of various treatments make repeated measures in small groups of subjects the customary research paradigm. In such cases, test stability, reliability, and factor structure take on extreme significance. In a Navy program 80 percent of 150 tests failed to meet minimally acceptable psychometric requirements. Recent findings with our battery show: acceptable psychometric properties in terms of both differential stability and reliability for both the long and the short battery; two factors available for the 7.5-minute test battery; four for the 15-minute battery and correlation with the WAIS. The factorial richness of the battery is adequate and goes beyond the factors that can be conveniently measured by more traditional paper-and-pencil tests into motor speed dimensions that may have important practical implications for assessment of concurrent functional capacity. Both factorial richness and correlation with the more global cognitive capacity construct, IQ, might be improved by the inclusion of subscales indexing verbal and arithmetic abilities, however, adding factors is not without penalty. The trade-offs of these issues (testing time, factor structure, stability, stabilization time, and reliability) are discussed. About a dozen validation studies are presently ongoing. What remains is to demonstrate functional validity in the detection of human functional capacity deficits in a real-world setting. This should be our primary mission in subsequent work.

INTRODUCTION

Exotic work environments often include factors (i.e., weightlessness, motion, fatigue, etc.) that disrupt performance. Furthermore, these settings are typically populated by limited numbers of highly critical workers. Kennedy and Bittner (1977) have observed that two connected concerns associated with the measurement of performance under such conditions are the lack of sensitive tests and a general unwillingness to expend the time and effort necessary to standardize such tests. A program designed to evaluate performance measures (PETER) was undertaken by the Naval Biodynamics Laboratory, New Orleans, LA (Kennedy & Bittner, 1977; Bittner & Carter, 1981; Kennedy, Bittner, Harbeson, & Jones, 1982), and more than 150 performance tasks have been examined for suitability in repeated measures research. Detailed descriptions of the evaluation process and task metric selection criteria may be found in Bittner, Carter,

Sponsored by NASA No. 9 16982 and NSF No. 00559

Kennedy, Harbeson,, and Krause (1984). A listing of 30 tests which survived this test and evaluation process appears in Bittner et al. (1984), but for the most part the tests that were studied were paper-and-pencil tests. The easy availability and economy of portable high speed computers suggests that innovative methods for automated data collection and analysis must be explored. Features that recommend microbased testing systems include capabilities for fully automated test battery administration and data storage, as well as portability and reduced size and weight. Automated and portable microprocessors capable of administering and storing performance measures and responses provide the obvious vehicle. The purpose of this report is to provide a descriptive overview and a brief prospectus of the Automated Performance Test System (APTS), and report our recent progress in the engineering analysis of the battery in terms of stability, reliability, factor structure, feasibility, and predictive validity. Ongoing validation studies which examine sensitivity to treatments are being recounted elsewhere at this meeting (Johnson, Kennedy, Merkle, Smith, & Bittner, 1985).

RECENT TEST DEVELOPMENTS

Preliminary Study

Method and Analysis. The "best" six tests from the PETER program (Bittner et al., 1984) were programmed on a portable microprocessor and administered along with tests in their original paper-and-pencil formats. Twenty three Casper College male and female students were tested over four replications on a 6.0 minute computerized battery. The group means, standard deviations, and 4X4 intersession correlation matrices were calculated for each task in each testing mode. Task group means and standard deviations were examined across sessions for evidence of task stabilization. Intersession correlations were assessed for evidence of task differential stability. Rapid stabilization was expected since at least theoretically comparable practice was received within both modes of testing.

Results and Discussion. The data showed that all tasks in both modes give good evidence of stability by the fourth session, with high reliability efficiencies (r >.85) for 3 min. of testing. Improvement, averaged across all tasks from sessions 1 to 4, was approximately 20%. The amount of improvement for paper-and-pencil testing (22.4%) was, in general, comparable to the amount of improvement demonstrated in the microbased testing mode (19.3%). Typically, paper-and-pencil testing produced higher scores across test sessions relative to microbased testing; however, from the data the acquisition curves for both modes are strikingly similar. Furthermore, the task standard deviations provide good evidence that none of the tasks in either mode has reached a ceiling. Clearly, all indicators point to good and comparable metric characteristics for the paper-and-pencil and microbased versions of each task. The factor structure obtained from the analyses of computerized test versions in each of the four sessions indicates the presence of two well-identified factors in the computer battery. Factor 1 is clearly a "motor" factor, probably related to response speed; as such, it affects performance on Pattern Comparison and Grammatical Reasoning. Factor 2 is just as definitively a "cognitive" factor with its importance for various tests changing with practice. The clarity of analyses under this constraint is encouraging. With respect to the paper and-pencil tests, there is reason to believe that these are essentially the same factors as for computerized versions, but the computerized versions appear to stabilize earlier and to be more clearly defined. It should be noted that in both the microbased and paper and-pencil analysis a possible third factor gave indications of emerging. The nature of the factor is unknown; however, the "automaticity" of responses characteristic of well-practiced skills (Ackerman & Schneider, 1984) is a likely potential explanation. Also, a significant general factor "g" may become evident within both modes of presentation. These findings are described in detail elsewhere (Kennedy, Wilkes, Lane & Homick, 1985).

Stabilization Study

Method and Procedure. Thirty-one male and female college students were recruited for participation. Prior to testing, subjects received a brief introduction to the purpose of the study and were advised regarding the general procedures associated with data collection. Ten (10) paper-and-pencil batteries and ten (10) microbased batteries were administered per subject, and the subjects were also tested individually with the Wechsler Adult & Intelligence Scale (WAIS). Six of the tests previously recommended above as a "mini-battery" for environmental research were included for examination. Four additional tests were also studied. The group means, standard deviations, and intersession correlation matrices were calculated for each individual paper-and-pencil and microbased test. Group means and standard deviations were examined for evidence of test stabilization, and intersession correlations were assessed for evidence of differential stability. Task definition (magnitude of r after stabilization) was determined directly and then adjusted according to the Spearman equation for test length and called average stabilized "reliability-efficiency" (to a 3-minute base). Predictive validity was assessed by comparison with an individually administered test of intelligence. Factor analyses used the principal factors method with squared multiple correlations as community estimates, followed by normalized varimax rotation.

Results and Discussion.

o Stability of Means - The overall impression of the 10 tests was that continued improvement occured over all tests but is slowed down in those tests considered stabilized. The Sternberg test appears to stabilize by Trial 4. The Preferred Hand Tapping and Non-preferred Hand Tapping tasks stabilize rather late in practice, by about Trials 7 or 8, whereas, the Two Hand Tapping task appears stable by Trial 4. The Pattern Comparison test stabilizes very rapidly, by at most Trial 3; whereas the Manikin test stabilizes later at Trials 6 to 7, as does Code Substitution. Gramatical Reasoning stabilizes rapidly, by Trial 3, and Reaction Time by Trial 5. The Landolt C test of dynamic visual acuity did not appear to stabilize over the 10 test days. There appears to be a strong learning component in this test as it is presently structured on the computer; thus it's stability is insufficient to be retained in the test battery in its current form. Clearly, the acuity test is a candidate for improvement in future versions of the performance battery.

o Standard Deviations - The standard deviations are constant and give no evidence of ceiling effects.

o Differential Stability - The Sternberg, the three Tapping tasks, Pattern Recognition, and Manikin reach apparent differential stability by Trial 3. Code Substitution, Gramatical Reasoning, and Reaction Time reach differential stability somewhat later, but apparently by Trial 6. The average reliabilities of these tests are all quite high. The intercorrelation matrix for the Moving Landolt C, on the other hand, does not give any indication of reaching differential stability.

o Factor Analysis - We recognize the limitations of performing a factor analysis with such a small sample, but are somewhat encouraged by the good stability and high reliability of the tests and plan for these results to be advisory. Factor I loads on Pattern Comparison and Code Substitution of both computer based and paper-and-pencil versions, and also loads on the Manikin and Fitts Histogram test which did not have dual versions. Factor II appears to be a Motor Speed factor because it loads on Reaction Time and the three Tapping tasks as well as the Sternberg, but loads on none of the paper and-pencil tests; therefore, Factor II may represent a construct that can be measured via computerized testing but not by standard paper-and-pencil tests. Factor III is probably best thought of as a Motor Control factor because it loads on Tapping tasks and Spoke and Aiming. Factor IV is a pure Gramatical Reasoning factor, loading on both forms of this test.

o Correlations with WAIS - The microcomputer based tests clearly correlate most strongly with Performance IQ and less strongly with Verbal IQ. The strongest simple correlation between the computerized tests and Full Scale IQ was for Gramatical Reasoning, although Non-preferred Hand Tapping was fairly high. The R squared values indicate that a substantial proportion of Performance IQ variance can be predicted from the computerized battery, but may also suggest that a more verbal subtest would improve the relation to Full Scale IQ.

o Paper and Pencil Tests - These tests essentially replicated the microprocessor based tests and a more complete description of these findings appears in Kennedy, Dunlap, Jones, and Wilkes (1985).

o General - All but one of the microcomputer tests could be recommended for a mid-ranged (<10 min.) battery. Based on the factor analysis, we would suggest Pattern Comparison, Sternberg, Tapping, and Grammatical Reasoning. We would also propose that each test be administered twice as long in order to improve the reliability and thus afford an opportunity for improved sensitivity.

Conclusions

The philosophy of our approach to performance test development involves three different phases. The first is to deal with only tests or tasks that can be shown to be psychometrically sound. This requires that we demonstrate stability of means and standard deviation within few administrations, and most important, that differential stability, the symmetry or constancy of trial-to-trial intercorrelations, be shown to occur quickly and at high values. The second phase is to show that the battery has factorial multidimensionality and that the subscales cross-correlate with earlier performance tests and other recognized instruments of ability. Finally, it is necessary to demonstrate and document sensitivity to factors known to compromise performance potential in laboratory and ultimately real world situations.

The cross correlation between the Portable Human Assessment Battery and the WAIS IQ measures was particularly interesting in that a substantial portion of the performance subscale variance could be accounted tor by this self-contained self-administered short battery of computerized tests. If arithmetic and verbal subtests were implemented we would be able to substantially improve the correlation with full-scale IQ. Of very great interest were the strong relations shown between Nonpreferred Hand Tapping and the verbal WAIS subscale. This was the second highest intercorrelation with the full-scale IQ. Jensen and Munro (1979) hypothesized that complex reaction times should be better predictors of general intelligence (g) than simple reaction times; and certainly this procedure is borne out by the present data, in that the Sternberg, a complex motor response task, has a higher correlation that simple reaction time with the WAIS IQ. Furthermore, Jenson and Munro (1979) found that motor speed showed as strong a relation as complex reaction time to g, an unexpected finding, but one that might relate to the strong relation between tapping and IQ in the present results. A comparison with other mental tests is in order (e.g., Armed Services Vocational Aptitude Battery (ASVAB). The extreme length of the ASVAB may be thought to improve the stability of its factors, but actually the second session of the ASVAB may be no more stable than Session 2 of the computer tests we reported, the latter taking 15 min (cf., McCormick, Dunlap, Kennedy & Jones, 1985). The Wechsler Adult Intelligence Scale (WAIS) takes 1-2 hrs. to administer and purports to sample two factors; these may or may not overlap with the four of the computer battery we report. In future studies in this program we will attempt to anchor our tests against these other better-known tests. For example, if only the computer tests are considered it may be possible to sample two motor and three cognitive factors, each within 6 min. and with reliabilities greater than r = .70% for a 3-min. base.

Based on the data reported above, we believe that the following four points are arguable and we would like to offer the following hypotheticial situations for speculation. (1) A Job Sample representing real-world work is likely to take > 100 hours to reach stability, and if a single (composite) score (e.g., correct detection) based on 60 min. of testing would be used to characterize performance, it is unlikely that such a score would have retest reliability greater than r = .60, and it might be lower. If many scores are broken out (e.g., hits, RMS error, miss distance), the individual reliabilities are likely to be lower than r = .30. Alternatively, the microprocessor that I have used would probably take < 1 hour (probably 30 min.) to reach stability on the test(s), and this total score is likely to have retest reliability (r > .90) for 12 min. of testing, as are each of the subtest scores. (2) High reliability does not assure sensitivity, but lack of reliability assures insensitivity. Most tests advocated for use in unusual environments or with toxic substances have neither been checked for reliability nor stability. (3) Our microprocessor based battery total score correlates well with global measures of intelligence. The single best predictor of job performance in all military jobs is a global measure of intelligence. Our battery probably shares considerable variance with military jobs and job performance. (4) After stability of performance on a Job Sample and stability of our battery's performance, if both are corrected for attenuation, a large proportion of the one hour test (perhaps 80%) would be shared with the 12 min. test, although the former may take 250 times as long to stabilize.

REFERENCES

Ackerman, P. L., & Schneider, W. *Individual differences in automatic and controlled information processing*. Urbana, IL: University of Illinois, Department of Psychology, 1984. (Report No. HARL-ONR-8401)

Bittner, A. C., Jr., & Carter, R. C. (1982). *Repeated measures of human performance: A bag of research tools*. (Research Report No. NBDL-81R011). New Orleans, LA: Naval Biodynamics Laboratory (NTIS No. AD A113954)

Bittner, A. C., Jr., Carter, R. C., Kennedy, R. S., Harbeson, M. M., & Krause, M. (1984, in press). Performance Evaluation Tests for Environmental Research (PETER): Evaluation of 112 measures. *Perceptual and Motor Skills*. (NTIS No. AD A152317)

Bittner, A. C., Jr., Smith, M. G., Kennedy, R. S., Staley, C. F., & Harbeson, M. M. (1984). Automated Portable Test (APT) System: Overview and prospects. *Proceedings of "On the Use of Microprocessors in Psychology" held in conjunction with Psychonomic Society*, San Antonio, TX.

Jensen, A. R., & Munro, E. (1979). Reaction time, movement time, and intelligence. *Intelligence*, *3*, 121-126.

Johnson, J. H., Kennedy, R. S., Merkle, P. J., Smith, M. G., & Bittner, A. C., Jr. (1985). *Microcomputer-based field testing for human performance assessment*. Paper presented at the 27th Annual Conference of the Military Testing Association, San Diego, CA.

Kennedy, R. S., & Bittner, A. C., Jr. (1977). The development of a Navy Performance Evaluation Test for Environmental Research (PETER). In L. T. Pope & D. Meister (Eds.), *Productivity Enhancement: Personnel Performance Assessment in Navy Systems*. San Diego, CA: Navy Personnel Research & Development Center. (NTIS No. AD A056047)

Kennedy, R. S., Bittner, A. C., Jr., Harbeson, M. M., & Jones, M. B. (1982). Television-computer games: A "new look" in performance testing. *Aviation, Space, and Environmental Medicine*, *53*, 49-53.

Kennedy, R. S., Dunlap, W. P., Jones, M. B., & Wilkes, R. L. (1985). *Portable human assessment battery: stability reliability, factor structure, and correlation with intelligence tasks*. Orlando, FL: Essex Corporation. National Science Foundation, Final Technical Rep;ort 85-3.

Kennedy, R. S., Wilkes, R. L., Lane, N. E., & Homick, J. L. (1985). *Preliminary evaluation of a microbased repeated measures testing system*. Paper presented at the 56th Annual Meeting of the Aerospace Medical Association, San Antonio, TX.

McCormick, B. K., Dunlap, W. P., Kennedy, R. S., & Jones, M. B. (1982). *The effects of practice on the Armed Services Vocational Aptitude Battery*. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. Behavioral Report No. ARI-1. (Final Report on Contract No. MDA903 82 M3943).

LEADERSHIP PREPAREDNESS IN NEWLY COMMISSIONED NAVAL OFFICERS

Leanne Atwater
Alice Crawford

Navy Personnel Research and Development Center
San Diego, California 92152-6800

Background

The growing concern for our national security and the concomitant expansion in the size of the Navy point to an increasing need for effective military leadership as expressed by the Secretary of the Navy. While leadership in general is one of the most thoroughly analyzed concepts to be found in the research literature, military leadership has had considerably less systematic attention. Military leadership, specifically leadership training for newly commissioned Navy officers, is the focus of this paper.

Navy personnel have generally viewed themselves first and foremost as leaders. They see the need for strong leaders--those who have the initiative, courage, and knowledge to think and act in situations where military objectives may not be easily recognized--as more critical than ever before.

There is concern, however, that many of the events during the past several decades have eroded traditional leadership values. The result has been an overemphasis on management-based, theoretical frameworks designed by social scientists who do not really appreciate leadership in a military context. Articles on this topic, written by the uniformed military are common. Sarkesian (1985), for example, traces the evolution of the corporate management model back to the McNamara years, which focused on cost-effectiveness and econometrics and shifted the focus of leadership in battle to the pursuit of management goals. He feels that the negative impact of this shift was felt in Vietnam. Byron (1985), believes that in peacetime the leadership demanded in combat situations is forgotten and instead, good managers are rewarded. To summarize, many feel that the Navy is losing its military leadership capability, which will in turn affect battle readiness.

An obvious paradox exists in the Navy community in light of all the concerns voiced when it comes to what should be done to develop these ideal leaders. Prior to the inception of this research effort, the authors conducted numerous interviews with senior Navy officers and a common theme emerged. Leadership training is conducted when everything else is finished--it is the last priority--and many, given their way, would throw it out altogether.

In part, this attitude toward leadership training stems from priorities. To teach leadership, something else must go. But this attitude seems to stem also from a lack of agreement as to what military leadership really is, and how (or even if), it should be trained. This definitional problem can be remedied if leadership is viewed as consisting of three dimensions--management, interpersonal skills, and warriorism. While warriorism may or may not be trainable, the majority opinion of leaders represented in a recently published book about military leadership (Taylor & Rosenbach, 1984) is that leadership skills (i.e., managerial and interpersonal skills) can be learned. This study addresses issues relevant to "teaching" these leadership skills to newly commissioned naval officers.

Current Leadership Training in the Navy

The Navy currently provides some training for new officers to help them assume their leadership role. Each of the commissioning sources has a leadership curriculum for naval officers as part of their education and preparation. The specialty schools (e.g., Surface Warfare Officers School (SWOS)) continue the leadership training process.

Until recently, the leadership training program at SWOS was a 2-week course designed as part of the Leadership Management and Education Training (LMET) program implemented Navy-wide in 1978 in an attempt to standardize the Navy's leadership training. Since 1978 many of the LMET courses have been dropped or shortened. The reasons for this are varied, but reflect to some extent the attitude mentioned earlier that leadership training is low priority training. It also reflects a desire to shorten the overall training pipeline and to get officers into the fleet sooner.

Purpose

If officer training is to be accomplished in the most efficient manner, it is imperative that the training provided be relevant, effective and not redundant. This study set out to address two primary questions. Does training make a difference? That is, are students learning anything they feel will be relevant to their leadership positions? If so, what are they learning and where, and can leadership training be provided more efficiently?

METHOD

Sample

The sample for this study consisted of 581 newly commissioned junior officers at SWOS. Four hundred and eighty-three students were questioned on their first day of SWOS about leadership training and preparation at their commissioning source. Ninety-eight additional students were questioned before and after participating in LMET. Table 1 presents a description of the sample.

Table 1

Description of SWOS Sample

	First Day of SWOS (N = 483)	Pre/Post LMET Sample (N = 98)
Commissioning Source		
USNA	196	3
OCS	37	68
NROTC	250	21
Academic Major		
Science/Engineering	300	41
Humanities/Other	183	59
Average Age	23	24
% choosing <u>surface</u> as 1st choice community	64	55
% intending to make Navy a career		
Yes	25	18
No 19	20	
Unsure	65	62

Questionnaires

Four versions of a questionnaire were designed to assess leadership training, leadership preparedness, and self-rated leadership abilities, and to collect a number of biographic and demographic characteristics of the students (e.g., academic major, age, sex, commissioning source). Among the four versions the following topics were also measured: managerial style, achieving style as measured by the Manifest Need Questionnaire, and action vs. state orientation. These topics will not be addressed in this paper.

Questionnaires were administered in classroom settings by SWOS instructors. Administrators were given a set of instructions to read to the students before filling out questionnaires. Students were assured that all information was confidential.

RESULTS

Does Leadership Training Really Make a Difference?

An important question in this study had to do with the merit of leadership training--does it really make a difference? Results indicated that leadership training is important to preparedness. A number of different analyses formed the basis of this conclusion. First, the relationship between new students' reports of their training and preparedness were correlated. As expected, the correlations between these two indicators were quite high (.5 to .7). At least in the students' minds they see their training as related to their levels of preparedness.

Second, analyses indicated that students' perceptions of their leadership training and preparation were more related to their feelings of preparation to go to war if necessary than were their self-rated leadership abilities. Preparation to go to war is not only determined by ability; training has an impact.

Third, a number of open-ended comments provided by new students indicated that they felt they needed more leadership training, especially in the interpersonal aspects, in order to best assume their role as division officer.

Fourth, 98 students were questioned before and after LMET about their leadership training and preparedness. On the basis of paired t-tests, 14 of the 21 items measuring different aspects of leadership preparedness showed significant improvements as a function of LMET. Many items had differences greater than a standard deviation. Specifics of these differences will be discussed in a later section.

What Leadership Training is Provided?

Training at the Commissioning Sources

The level of leadership training new officers received at the three major commissioning sources was of particular concern in this study. Table 2 presents a list of the leadership training and preparation topics assessed and indicates the areas where officers felt <u>most</u> and <u>least</u> trained and prepared. In general, officers entering SWOS felt they were trained "to some extent" (the midpoint on the scale which ranged from 1, "not at all," to 5, "to a very large extent") in most aspects of leadership. The overall levels of perceived preparation were somewhat higher ($\bar{x}$ = 3.6).

Table 2

Leadership Training and Preparation Topics Measured and Extreme Averages By Commissioning Source

	New SWOS Students[a]						Greatest Improvement after LMET
	USNA		OCS		NROTC		
	Trng	Prep	Trng	Prep	Trng	Prep	
	(N = 196)		(N = 37)		(N = 250)		(N = 98)
a. Making the transition from the Naval Academy, OCS or NROTC to the operational Navy							
b. Taking responsibility for a division of enlisted personnel							
c. Understanding Navy procedures and protocol			3.5				
d. Relieving the division officer in your first division officer assignment	2.8	2.9			2.7	3.0	X
e. Knowing how to motivate enlisted personnel							X
f. Performing the paperwork requirements as a division officer (PMS, PQS, etc.)	2.5	2.7	2.5	3.0	2.6	2.8	
g. Managing your time and setting priorities when you have a heavy workload	4.4	4.3	3.6	3.9	3.6	3.9	
h. Talking to a large group of people who work for you							
i. Briefing your superior, or the CO about an issue in your division							X
j. Counseling subordinates about personal matters			2.4				
k. Counseling poor performers			2.4	3.3			X
l. Handling alcohol and drug abuse problems among your subordinates						3.3	
m. Resolving conflicts among your crew members					2.7		
n. Listening effectively				3.9		3.9	
o. Managing stress (i.e., lack of sleep, disappointing your boss, overwork, conflicts)	4.1	4.2	3.4				
p. Communicating with people effectively				3.9		3.9	
q. Demonstrating concern for your subordinates							
r. Setting goals	4.1	4.2				3.7	
s. Planning work	4.1	4.1				3.5	
t. Interacting with Chiefs in your division	2.7	3.1					X
u. Rewarding and disciplining your subordinates							X

[a]Only areas with highest and lowest average levels of training or preparation are presented in this table.

The levels of training differed across commissioning sources in 18 of 21 areas. For the most part, Naval Academy (USNA) graduates felt better trained than those from Naval Reserve Officer Training Corps (NROTC) or Officer Candidate School (OCS), and OCS graduates generally felt least trained. A noteworthy exception to this was training in how to interact with chiefs in the division. In this area, NROTC graduates felt best trained and Naval Academy graduates felt they had received little training.

The aspects of leadership in which new SWOS students, across commissioning sources, felt they had received the most training and were most prepared were "managing time" and "setting priorities with a heavy workload." Naval Academy graduates, in addition, felt well trained and prepared in terms of managing stress, setting goals and planning work. Graduates of OCS and NROTC, however, felt more prepared in areas of listening effectively and communicating effectively with people than they did in setting goals and planning work.

Areas where new SWOS students felt least prepared were "relieving the division officer" and "performing paperwork requirements." Those from OCS also reported little training in terms of "counseling subordinates," although they felt prepared to some extent to do this.

Although not reflected in Table 2, when examining the levels of training and preparation on the more interpersonal aspects of leadership (e.g., knowing how to motivate subordinates, briefing superiors, counseling subordinates), the levels of training were rather low, and the levels of preparation were moderate. This was especially true for graduates of NROTC and OCS.

It appears that in general, students leaving the commissioning sources feel better prepared in terms of managerial type skills (goal-setting and planning) than in the interpersonal aspects of leadership (counseling and disciplining).

LMET Training at SWOS

The results of LMET training improve this state of affairs to some extent (see Table 2). Leadership areas which showed the greatest improvements as a result of LMET were "motivating enlisted personnel," "talking comfortably before a large group," "counseling poor performers," "interacting with chiefs in your division," "rewarding and disciplining subordinates," and "relieving the division officer."

Also of interest were two items in which perceptions of the levels of training decreased from time 1 to time 2. These areas were "understanding Navy procedures and protocol" and "performing the paperwork requirements as a division officer." It seems likely that LMET served as a realistic preview and made officers aware of the large amount they didn't know in these areas.

Self-Ratings of Leadership Ability

While the majority of students' perceptions of their leadership training differed significantly across commissioning sources, self-ratings of leadership ability in eleven areas did not differ.

The area in which all students felt most capable was in "doing whatever it takes to get the job done." Aspects of leadership in which students felt least capable were "speaking comfortably in front of a group," and "motivating subordinates to do jobs they don't want to do." In general, officers feel more able to perform managerial duties than to handle the more interpersonal aspects of their leadership role.

The leadership ability ratings of Academy graduates were somewhat surprising. While Academy graduates felt better trained and better prepared than graduates of the other two commissioning sources, they did not rate their leadership abilities any higher. This seems to dispel, to some extent, the stereotype that Academy graduates have trouble adjusting in their first tour as division officers due to an over-confidence. It also suggests that abilities are not the only factor relevant to how well prepared officers feel they are. Training does play a role.

Additional Findings of Interest

Who or What Influenced Leadership Skills?

New SWOS students were asked the extent to which classroom instruction, examples of military leaders, experience leading others, as well as things they learned before they entered USNA, OCS or NROTC influenced the leadership skills they had acquired. Academy and NROTC graduates felt examples of military leaders and experience leading people were most influential. OCS graduates felt they had been most influenced by things they learned before attending OCS. OCS graduates also tended to feel their levels of training had been low in comparison to their levels of preparation. This supports the conclusion that OCS graduates feel their leadership skills are less a function of their officer training than do those who go through NROTC or the Naval Academy. (This is not surprising when the length of these officer training programs is considered (i.e., four months for OCS as opposed to four years for USNA and NROTC).) All three groups felt classroom instruction was least influential, though they still reported it influenced them "to some extent."

Assuming Different Leadership Roles

Functioning effectively as a naval officer encompasses a number of different leadership roles. The officer must function as a leader of others, a technician, a professional, and, at times, a warrior. SWOS students were asked how well prepared they felt to assume each of these roles (see Table 3). Students from all commissioning sources, at SWOS for the first day, felt most prepared to assume the role of professional, and least well prepared to assume the role of technician. One might expect that 14 weeks of intense SWOS training would improve officers' confidence in their preparation to assume the role of technician, but the sample of students questioned on their last day of SWOS also felt least prepared to assume the role of technician when compared to the other three roles. This may be because officers are trained to manage technicians rather than to develop technical expertise. Alternatively, their training may serve to make them aware of how much there is to know.

The overall levels of preparedness to go to war if necessary were fairly high ($\bar{x} = 3.6$). All 21 aspects of leadership training were significantly related to preparation to go to war. The training variables most highly correlated with preparation to go to war were in areas of motivating enlisted personnel, handling alcohol and drug abuse problems, resolving conflicts and planning work. These correlations were all approximately .26.

Table 3

Average Levels of Preparation to Assume Diverse Leadership Roles
Before and After SWO Training by Commissioning Source

	New SWOS Students				SWOS Graduates			
Leadership Role	USNA (N = 196)	OCS (N = 37)	NROTC (N = 249)	Overall (N = 482)	USNA (N = 5)	OCS (N = 86)	NROTC (N = 34)	Overall (N = 125)
Technician	3.4	3.2	3.1	3.2	4.4	2.9	3.0	3.0
Professional	4.3	3.9	3.9	4.0	4.2	3.6	3.9	3.7
Leader	3.9	3.4	3.7	3.8	3.8	3.5	3.7	3.5
Prepared to go to war	3.8	3.8	3.5	3.6	4.0	3.6	3.6	3.6

Academic Major and Leadership Preparedness

The Secretary of the Navy, John Lehman, and others have suggested that a strict science and engineering curriculum may be too academically narrow to provide new officers with the well-rounded education they need to be good leaders. To address this question, correlations were computed between academic major (science and engineering vs. other) and students' perceptions of their leadership training, preparation and ability. (These correlations were partial correlations controlling for commissioning source.) No significant relationships were found between academic major and training, preparation or ability. Since the measures of leadership are all self-report, they must be treated with some caution, but nonetheless, no relationships emerged.

DISCUSSION

It was encouraging to discover that students felt their training influenced their levels of preparation to become leaders, and that they believed classroom instruction was useful. It was also interesting that academic major had no relationship to students' perceptions of their leadership preparation or leadership abilities. The stereotype that science majors have narrow academic experiences and, therefore, tend to be less sensitive to interpersonal concerns was not supported.

Of particular interest were the training issues addressed. It appears that each of the commissioning sources are preparing new officers to some extent to assume their leadership responsibilities. The bulk of this training seems to impart managerial skills rather than the more interpersonal leadership skills. LMET at SWOS furthers this leadership preparation with a positive impact on the interpersonal dimensions. If the Navy's goal is to provide the most efficient training pipeline, it appears that the commissioning sources would do well to concentrate formal leadership training in the managerial skills, leaving the interpersonal skills to LMET. This would be an improvement over providing minimal training in all areas at both schools. It also seems that an interpersonal perspective in managerial training would be worthwhile (e.g., planning work and setting goals for subordinates). Further, while the length and method of the training experiences differ at the various commissioning sources, they would do well to agree on a standard set of leadership issues to be addressed in leadership training and do them as well as possible in the time frame provided.

While the findings from this study are based on self-report they are suggestive of issues worthy of future pursuit. This work will be followed up with input from SWOS instructors as to students' leadership abilities, an evaluation of the shortened LMET curriculum and optimally, a one-year follow up of individual performance in the fleet.

REFERENCES

Butler, M. C., Bruni, J. R., Hartman, F. A., & Hilton, T. F. (1984). Manifest needs among health care professionals: Dimensionality and reliability. Paper presented at the Annual Meeting of the APA, Toronto.

Byron, J. L. (1985). W[illegible]riors. Proceedings, June 1985, pp. 64-68.

Sarkesian, S. C. (1985). Leadership and management revisited. The Bureaucrat. Spring, 1985, pp. 20-24.

Steers, R. M., & Braunstein, D. N. (1976). A behaviorally-based measure of manifest needs in work settings. Journal of Vocational Behavior, 9, 251-266.

Taylor, R. L., & Rosenbach, W. E. (1984). Military leadership: In pursuit of excellence. Boulder, Colorado: Westview Press.

Leadership Development of USAF Aircraft Maintenance Officers

Captain Michael A. Morabito
437th Organizational Maintenance Squadron, Charleston AFB, SC

Captain Benjamin L. Dilla
Air Force Institute of Technology, Wright-Patterson AFB, OH

A sample of 320 Air Force aircraft maintenance officers (AMOs) were surveyed using the updated version of Yukl's Managerial Behavior Survey (MBS), to measure leader behavior of the AMO's superior officer, and other scales focusing on the AMO's perception of his/her own leadership development. Specific development methods used by AMOs and the perceived importance of each were explored. Furthermore, suggestions were collected on ways to improve development methods available to them in the Air Force. The leadership development activities were correlated with the superior's leader behavior and with demographic and organizational variables. The personal factors of age and rank were found to be associated with leadership development. Participation in 8 of 19 leadership activities correlated significantly with the degree of importance placed on the activities. Analysis of the MBS results indicated certain categories of the superior's leader behavior were significantly associated with the perceived leadership development of the AMO.

Introduction

Leadership is a constant concern in the military environment, yet very little is known about how leaders are developed. The overall objective of this research was to identify methods of leadership development used by Air Force officers, specifically junior aircraft maintenance officers (AMOs), with special interest in the impact of the immediate superior's leader behavior on the junior officer's development.

For the purposes of this research, leadership is defined as a dynamic, goal-directed process of influence between leader and follower, including the interaction of each with the situation (Yukl, 1981). Leadership development is broadly defined as any method or activity used by individuals to enhance their personal ability to influence subordinates toward goal accomplishment.

Leadership Theory and Research

The subject of leadership has been addressed by many scholars from ancient times to the present. One of the major approaches which has been key in understanding the concept is the leader behavior approach. Research at Ohio State University resulted in identification of the classic dimensions of consideration and initiation of structure (Halpin and Winer, 1957), while efforts by the University of Michigan Survey Research Center produced the distinction of job-centered and employec-centered leadership (Likert, 1961) and, later, the four categories of support, interaction facilitation, goal emphasis, and work facilitation (Bowers & Seashore, 1966). These categorizations, although important in developing an understanding of leadership and still used today in the context of some of the situational theories, are not without their problems.

One of the key issues has been the recognition that ". . . these broadly defined categories provide too general and simplistic a picture of leadership. They fail to capture the great diversity of behavior required by most kinds of managers and administrators" (Yukl, 1981, p. 120). This realization led Yukl and his colleagues to develop a more comprehensive categorization, which currently includes 13 dimensions. According to Yukl, "The advantage of the new taxonomy is that it has a larger number of more specific behavior categories than earlier ones, and it includes most behaviors found to be important in leadership research" (Yukl, 1981, p. 128). Because of these advantages, Yukl's Managerial Behavior Survey (MBS) was selected for the description of superior officers' leader behavior in this research.

Military Leadership Development

Studies of the process of leadership development in the military services has focused primarily on professional military education (PME) or other formal programs. The Air Force has essentially come full circle in their philosophy of educating leaders in the past 12-15 years. Reports in the 1970s expressed concern by senior officers on the need for development of technical and management skills (Dobias, 1974; Robinson, 1974) and suggested better pre-commissioning training and even on-the-job training. Then came concern in the late 1970s, still present today, that the Air Force had gone too far in the technical training of young officers, making them more occupational than truly professional (Gosnell, 1980). Another recent Air Force study of particular interest emphasized the important role played by commanders in serving as leadership models for their subordinates (Benton, 1981). This important concept of mentoring has been much neglected and even resisted by the Air Force with its emphasis on formal programs of education and training.

Research within the U.S. Army reinforces this emphasis on the supervisor-subordinate relationship. In analyzing problems in junior officer development, Wellins and others (1980) noted that a key issue was that "the senior officer may not take time to supervise, guide, and correct the performance of the new lieutenant" (p. 5). Another important study conducted within the Army concluded that the most successful leader training results from experiential processes ("learning by doing") rather than analytical or procedural processes (Shriver et al., 1980).

Perhaps the most revealing observation from a review of the literature on leadership theory and leadership development is that there has been little if any research to identify methods which present or prospective leaders actually use for leadership development. To identify these methods and their relative usefulness for a specific group of leaders was the focus of this research.

Methodology

Sample

The original population of interest was all U.S. Air Force aircraft maintenance officers serving in the grade of lieutenant or captain. For practical considerations, this was narrowed to those serving in the continental United States (CONUS) in the three largest major commands--Military Airlift Command (MAC), Strategic Air Command (SAC), and Tactical Air Command (TAC). Generalization of the findings of this research should be limited to these specific population parameters. According to official sources, the population size was 730 officers. Air Force officials authorized surveys for a random sample of 320 individuals to allow for a 95 percent confidence level.

Procedure

Surveys were mailed to officers at their duty addresses, accompanied by a cover letter signed by the Dean of the School of Systems and Logistics, Air Force Institute of Technology (AFIT/LS). Participation was voluntary, and respondents were assured of anonymity. Respondents were asked to complete and return an optical scanning sheet for the standard survey items. Space for comments and suggestions was provided at several points in the questionnaire booklet. Respondents were instructed to return all materials in a postage-paid return envelope.

Measures

The MBS, provided by Dr. Gary A. Yukl of the Business School, State University of New York at Albany, was used to assess subordinate perceptions of their superior officer's leader behavior. This instrument measures the frequency of 130 leader behaviors (ten items in each of thirteen categories).

The rest of the survey was composed of items written especially for this research effort. It included: (a) standard demographic information, including sex, age, source of commission, rank, major command, prior service, and organization/level (7 items); (b) perceived extent of leadership development (4 items); (c) immediate superior's leadership effectiveness (1 item); (d) perceived importance of leadership development activities (18 items); and (e) extent of involvement in leadership development activities (15 items). Items on PME completed by different methods were grouped together in section e, producing a smaller number of items than in section d.

Results

Respondent Profile

From the random sample of 320 officers, 185 usable surveys were returned, for a return rate of 57.8%. This was considered reasonably good since the survey package contained 27 pages and 190 items. The sample was predominantly male (84%) with the median age category being 30-34 years. Most received their commission from Officer Training School (58%) vice ROTC (37%) or USAFA (4%). Consistent with this pattern, the majority had prior enlisted service (56%). Almost half (48%) were captains, with about an equal split among the rest between second lieutenant (26%) and first lieutenant (25%). Half were assigned to TAC (51%) with the rest divided almost equally between SAC (25%) and MAC (24%). The largest number were assigned to Organizational Maintenance or Aircraft Generation Squadrons (37%); almost two thirds (65%) were assigned to the unit level versus the DCM staff or a higher headquarters.

Leadership Development Activities

Respondents were asked to rate the perceived importance of leadership development activities on a five-point scale ranging from "not important" (1) to "extremely important" (5). They were later asked to indicate their completion of certain activities (e.g., Squadron Officer School) or extent of involvement in other ongoing endeavors (e.g., number of hours per week spent in personal study of leadership). Results are summarized in Table 1.

Table 1
Leadership Development Activities

ACTIVITY	MEAN IMPORTANCE RATING	INVOLVEMENT (Average or Percent)
Leadership of NCOs	4.21	6.60 Times/Week
Observation of Superiors	4.15	4.19 Times/Week
TDY Deployments	4.15	3.83 Weeks/Year
Leadership of Airmen	4.08	6.65 Times/Week
Leadership of Peers	3.91	4.07 Times/Week
ACSC in Residence	3.86	0.0% Have Attended
Other Leadership Activities	3.39	2.23 Hours/Week
ACSC by Seminar	3.12	8.6% Have Completed
Graduate Degree	3.07	29.7% Have Completed
Personal Leadership Study	2.90	1.73 Hours/Week
ACSC by Correspondence	2.78	5.4% Have Completed
Sports Leadership	2.73	1.27 Hours/Week
Other AF-Related Activities	2.73	0.67 Hours/Week
Community Leadership	2.71	1.02 Hours/Week
Professional Org. Leadership	2.68	1.02 Hours/Week
Church Leadership	2.66	0.69 Hours/Week
SOS in Residence	2.61	34.6% Have Attended
LPDP	2.53	15.1% Have Completed
SOS by Correspondence	2.29	34.0% Have Completed

The most important activities to the AMOs were their working experiences with NCOs, airmen, superior officers and peers, and on TDY deployments. The least important in their eyes were the formal programs which are typically emphasized in official circles. These included the Lieutenants Professional Development Program (LPDP) and PME, especially by correspondence.

Chi-square tests were employed to determine if any significant relationships existed between the rated importance of leadership development activities and the extent of involvement in those activities. Eight of the nineteen leadership development activities were found to have a statistically dependent relationship at the .05 level of significance. They were: leadership of NCOs, TDY deployments, other leadership activities, graduate degree, personal leadership study, other Air Force related activities, professional organization leadership, and church leadership. These activities spanned the spectrum of involvement ratings and level of involvement, so no pattern was obvious. The findings could be interpreted to say that the officers perform a given activity because they perceive its importance in developing their leadership ability or that the activity is rated as being important simply because the officers are involved in it.

Managerial Behavior Survey Results

Before using the MBS results, internal consistency reliabilities were computed for each of the thirteen scales; all were .86 or greater. Also, means and standard deviations of the scale variables (formed by summing the ten item scores for each scale) were computed to provide a picture of the average leader behavior of the superior officers. Results are summarized in Table 2. Of interest in the present context was the fact that "developing" behavior was one of the lowest rated categories.

Table 2
Managerial Behavior Survey

Leader Behavior Scale (10 items each)	Cronbach's Alpha	Mean	Standard Deviation
Supporting	.93	2.80	.726
Problem Solving & Crisis Mgt.	.90	2.78	.632
Interfacing	.92	2.68	.669
Representing	.92	2.67	.700
Informing	.89	2.64	.684
Consulting & Delegating	.91	2.58	.706
Monitoring Operations	.86	2.51	.656
Harmonizing & Team-Building	.93	2.49	.747
Recognizing & Rewarding	.92	2.48	.770
Planning & Organizing	.91	2.40	.755
Motivating Task Commitment	.89	2.34	.731
Developing	.92	2.27	.797
Clarifying Roles & Objectives	.90	2.21	.765

Next, a set of correlations were computed between the AMO's perceived leadership development and the reported leader behavior of the superior officer. The hypothesis was that some types of leader behavior (e.g., developing, consulting and delegating, recognizing and rewarding) would be positively related to the subordinate's leadership development. In fact, all of the correlations between the MBS scales and self-rated leadership development were negative. Nine of the thirteen were statistically significant at the .05 level; however, the largest negative correlation was only -.19 (for "supporting" behavior).

Personal Determinants of Leaoership Development

The final major analysis was accomplished to determine if there were any relationships between the demographic items and leadership development by the AMO. A one-way ANOVA was used for each of the categorical variables with perceived leadership development as the dependent variable. The ANOVA for rank resulted in significant differences between group means (p<.02), while there was a marginally significant difference for prior enlisted service (p<.09). In both cases, greater experience (as indicated by higher rank and prior service) was associated with greater leadership development.

Discussion

Our major conclusion is that experience is the best form of leadership development. The AMOs in this study placed the highest importance on their working relationships with airmen, NCOs, and peers, their observation of superior officers, and their experiences on TDY deployments. Furthermore, these were the activities in which they reported the greatest involvement. These results were also supported by demographic data which indicated greater perceived leadership development for those having higher rank and those having prior enlisted service. It would appear that the old adage, "there's no substitute for experience", is indeed true when it comes to leadership.

A striking result on the opposite end of the spectrum was the comparatively low ratings given to formal development programs such as LPDP and PME, especially for the correspondence versions. These trends were supported by many of the written comments which emphasized the need for experiential training.

Even other varieties of leadership experience, e.g. in the community, churches and professional organizations, received relatively low ratings versus those categories which were specific to the job. These trends would seem to support the now well-accepted situational aspect of leadership. Experience, to be most effective, should be specific to the job environment.

A surprising outcome was the negative correlations observed between the superior's leadership behavior and leadership development of the subordinate. It was expected that at least some forms of leader behavior would enhance the AMO's leadership development. The negative relationship observed could be explained in at least two ways. They could indicate that if a junior officer has developed well as a leader, then the superior does less leading and allows the junior officer to lead. Another possible explanation is that the more the AMOs feel they have developed as leader, the more critical they are of the superior's leader behavior.

This research focused on rather global issues of leadership development, specifically in the aircraft maintenance career field. Much has been learned; however, many questions remain. Future research should further examine the relative worth of different methods of leadership development; such trends should be examined in various settings with different groups of leaders. A criterion measure more objective than self-rated leadership development would be particularly useful in follow-on work.

References

Benton, J.D. (1981). Promoting leadership in the Air Force's management environment. Unpublished research report, No. 0230-81, Air Command and Staff College, Maxwell AFB, AL.

Bowers, D.G., & Seashore, S.E. (1966). Predicting organizational effectiveness with a four-factor theory of leadership. Administrative Science Quarterly, 11, 238-263.

Dobias, L.J. (1974). An analysis of management development in the Air Force. Unpublished research report, No. 0785-74, Air Command and Staff College, Maxwell AFB, AL.

Gosnell, W.L. (1980). The Air Force is making occupationalists of its junior officers. Unpublished research report, No. MS071-80, Air War College, Maxwell AFB, AL.

Halpin, A.W., & Winer, B.J. (1957). A factorial study of the leader behavior descriptions. In Stogdill, R.M., & Coons, A.E. (Eds.), Leader behavior: Its description and measurement (pp. 39-51). Columbus: Bureau of Business Research, Ohio State University.

Likert, R. (1961). New patterns of management. New York: McGraw-Hill.

Robinson, G.D. (1974). Schooling the middle manager. Unpublished research report, No. 5406, Air War College, Maxwell AFB, AL.

Shriver, E.L., and others. (1980). Development of a leader training model and system. Unpublished research report, AD-A082 730, Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

Wellins, R.S., and others. (1980). Analysis of junior officer training needs. Unpublished research report, AD-A096 034, Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

Yukl, G.A. (1981). Leadership in organizations. Englewood Cliffs, NJ: Prentice-Hall.

IMPLEMENTING TECHNOLOGY IN A MILITARY ENVIRONMENT

William L. Maloy and Nancy N. Perry

Chief of Naval Education and Training
Pensacola, Florida

INTRODUCTION

The Chief of Naval Education and Training established the Training Technology Implementation (TTI) Office on his staff to transition proven technologies into Navy training settings systematically. The charter of this organization involves placing specific technologies and techniques in the training environment, and, more importantly, developing successful strategies and overall architecture for implementation to guide future efforts.

This paper briefly describes three ongoing projects in order to set the context for examining several lessons learned by the TTI Office. Out of these lessons surface broad policy issues which, if dealt with, promise more orderly implementation of training technologies and techniques in the future.

IMPLEMENTATION PROJECTS

Automated Maneuvering Training Board System

The first TTI initiative grew out of the Defense Science Board's Summer Study of 1982 pronouncement that not enough technology was finding its way into the military training environment. The Chief of Naval Research agreed to provide investment capital for this endeavor. The Chief of Naval Education and Training was assigned the task of designing the implementation architecture and directing the implementation process. Representatives from the Chief of Naval Operations, the Chief of Naval Education and Training, the Office of Naval Research, and the Navy laboratories identified several of the most promising Research and Development (R&D) products as candidates for this implementation project.

Meanwhile, researchers at the Navy Personnel Research and Development Center were developing a new method for guiding the individualization of ship maneuvering interactions; the Automated Maneuvering Board Training System (AMBTS). Because of the inherent difficulty involved in mastering maneuvering board skills, the implementation of AMBTS at the Operations Specialist (OS) "A" School in Dam Neck, Virginia was selected for the initial demonstration project. Implementation architecture included training effectiveness and cost-effectiveness analyses, life cycle costing, and acceptability.

Performance of students using AMBTS is being compared to that of students receiving the traditional instruction. Additionally, a sample of students will be followed into the fleet to determine if performance differences persist in the job environment.

AMBTS in Pierside and Reserve Training

The second initiative grew out of the technology of the first, but focuses upon the pierside and reserve community components of the training continuum. The issues here are not the effectiveness and efficiency of initial skill training but rather maintenance of skills, recertification of individuals returning to sea after protracted shore assignments, and qualification of teams whose players must learn to coordinate their individual skills to maximize team efforts. The fact that the Reserves must accomplish this training at remote sites exacerbates these problems for them.

AMBTS is the medium for accomplishing the training; however, a method for assessing individual performance and prescribing follow-on or refresher training must be added to the capability of the basic program. Through this application of the technology, we hope to learn how to apply what is known about the retention of learning. This will heip us design acceptability strategies which can be applied in the fleet and the reserve community where there is a good deal of skepticism about technology's value as an instructional tool.

Training Continuum Strategies

The third initiative is in an entirely different training arena, but builds upon the training concepts of the second initiative. In this case, the implementation model focuses not upon the learning and maintaining of technical skills but on the educational requirements of the officer accession continuum. We have selected an instructional continuum that begins with a Navy school preparing potential NROTC students for the college environment, to the NROTC itself, to the follow-on warfare schools, and, ultimately, to the fleet.

The process for making this continuum operate more effectively includes: (1) assessing individual capabilities at each school level against required developmental and professional skills; (2) intervening at each point along the way to close deficiency gaps through the application of computer-based modules and other good support practices (including peer tutoring); and, (3) providing follow-on assessment and deficiency prescriptions to the next schoolhouse which can then continue the remediation/development process.

Through this interaction among those responsible for training along the continuum, we seek to add a "feedforward" as well as a feedback process that establishes an accountability for value added at each follow-on school. If we are successful at packaging these techniques, we will have in place the basic ingredients of a true training continuum which should become the basis for all Navy training in the future. Perhaps we can then, at last, reduce the impact that manpower and personnel policies have upon training performance and concentrate upon helping people learn what they need to know -- without wasting time on blame.

TTI has other initiatives underway. These include the identification of applications for smoke generators in Reserve training, the employment of artificially generated signals in gram analysis training in Navy settings, and the development and implementation of a Navy-related functional skills math and reading curriculum.

The purpose of all these early initiatives is to package existing technologies and instructional materials into meaningful instructional delivery systems. These systems are tailored to the unique instructional environments along the continuum and enjoy the credibility and acceptability of schoolhouse instructors and fleet supervisors. As we continue to improve this packaging process and couple it with such delivery concepts as requalification, we come closer to exploiting our ability to intervene in the capability of people to perform. We can then allocate our resources on continually assessing, building upon, and sustaining skill development as the cornerstone for fleet readiness.

IMPLEMENTATION LESSONS

During planning, initiating, and executing several implementation projects, TTI encountered recurring issues and problems. By examining and trying to solve them, we uncovered lessons that should be heeded during systematic implementation. These lessons are consistently found and persistent enough that even the obvious ones should be discussed and all should be incorporated into overall implementation models. Attending to them leads to nothing more than good implementation practices.

1. The difficulty of implementing technology is continually underestimated.

This is a self-evident truth, but one that is generally ignored. Researchers and operators in both the military and civilian communities continually underestimate the difficulty of implementing technology. Researchers and operators have different goals and interests; therefore, just recognizing that the process is not likely to solve the problem.

Perhaps this is why we are seeing a new breed of professional; and the lesson is we must encourage the development of these people -- people who can stand, however shakily, with one foot in the the research laboratory and the other in the operational environment. These specialists, employing these emerging implementation skills, can bring the transitioning process to a level of sophistication which will build a climate of operator acceptance and ownership. Maintaining this climate throughout dissemination, will best contribute to the sustainability of new training technology and techniques in the classroom.

2. The capability of technology to improve training should not be exaggerated.

This second lesson highlights a principle which has long been discussed with regard to innovation but is often overlooked in practice. The training equipment that can be uncovered in military salvage depots is testimony to our willingness to embrace new technologies as panaceas and our subsequent disillusionment with them.

Similarly, in our zeal to adhere to the letter of our instructional systems design law we have too often confused familiarity with mastery. We have also overlooked the impact of forgetting on initial performance in an operational setting. It is now apparent that exaggerated claims of improved performance, often accompanied by poorly utilized computer technology, helped to undermine the credibility of those endeavors with operators who knew better. This frequently resulted in animosity toward the ISD process and the techniques and technologies it promoted.

At the beginning of every implementation endeavor we are well advised to realistically formulate promises and use technology in supporting roles and as tools where appropriate. Our initiatives at TTI incorporate this advice.

3. Goals determining hardware and software uses must be carefully defined.

It is important to have goals that determine hardware and software uses. What we are learning, however, is that we must use great care not to structure our goals too narrowly around near term cost and training benefits as we perceive them which may, in turn, restrict our our ultimate flexibility.

As we get better at recognizing technology linkages, there will be numerous opportunities to build one system upon another. In this way we can expand the applications for individual team and crew training uses as well as leapfrog those applications from the schoolhouse to shipboard environments. However, it will be difficult to fully exploit these possibilities if we have been overly restrictive in our early-on goal setting and front-end analysis processes. We must improve our abilities to perform trade-off analysis and take risks.

4. Technical skills must be acquired in order to implement technologies successfully.

Trainers and implementers may have an edge over others when it comes to analyzing the capabilities of technologies to improve training, but specific hardware decisions are often out of our realm of expertise. We need people who know about hardware configurations -- people who can provide guidance in ordering equipment and helping us with hook-up, operating, and maintenance problems.

The role of these people includes the development of quality controls in the acquisition, use, and maintenance of hardware and materials, and monitoring vendor performance. There is no reason to tolerate the shoddy performance that many of us have experienced in training systems technology. We should use these people to develop stringent, enforceable quality controls across the board.

POLICY ISSUES

Our early attempts to move technology from research to implementation brought two policy issues into bold relief. If our initiatives are to be optimally successful we must deal with both the issue of a funding continuum to accompany the R&D continuum and with procurement issues.

Funding Continuum Issue

As an R&D product moves from exploratory research to engineering development, to prototype demonstration, to pilot testing, and, finally, to full-scale implementation, a systematic funding continuum to assure this orderly progress is essential but currently lacking. To establish this funding continuum we must begin by identifying clearly stated, broadly based umbrella-type research categories under which specific projects can be grouped. These categories should be derived primarily from senior officers responsible for training who should review the categories periodically in order to renew their commitment or change them as a result of changing military priorities.

Specific R&D projects will fall under these umbrella categories where they will be prioritized and appear as a totally funded package through the 6.4 funding level. This portion of the continuum must then be coupled with follow-on operating dollars, initially in the form of wedges. As the budget year comes closer, these wedges will give way to the more specific dissemination and life cycle requirements emerging from the 6.4 accomplishments.

Without this kind of funding continuity we will never fully realize the potential that research-based instructional techniques and technologies hold for improved performance.

Procurement Issues

Software and hardware procurement and maintenance issues can no longer tolerate policy neutrality. In the early stages of transitioning R&D from the laboratory to the classroom, trainers must be formally involved in the procurement process. There are too many training issues involved in purchasing decisions (and many quality control initiatives) to leave technology skilled trainers out of the process.

Further, we must weigh the advantages and disadvantages of the centralized/decentralized issue in upgrading and maintaining software and courseware. To date, each implementation project has found its own, not always satisfactory, solution to this problem. We must consolidate the knowledge gained by these projects and prepare ` standardized, supportable policy on this issue.

Finally, policy matters associated with life cycle costing greatly concerns those responsible for training in the classroom and at pierside. How these policies evolve is closely tied to the credibility and acceptability technology applications will enjoy among our military colleagues.

CONCLUSION

We in Navy training have embarked upon implementation in a formal way. We believe this to be good, sensible, and even wise. However, we suspect that our initiatives, our experience, and our policy concerns differ little from others'.

The task before us, then, is to learn how to learn from each other; to share what we know; and to build on each other's experiences. How well we do these things will determine how well we fulfill our goals in the struggle to be more effective in a world of scarce resources where so much depends on human performance.

Research Needs Assessment and Technology Transfer in USAREUR

Karol Girdler and LTC Ford McLain
U.S. Army Research Institute

In October 1984, the Army Research Institute (ARI) Field Unit in the U.S. Army Europe (USAREUR) ended its programmatic research activities, and it was redesignated as a Scientific Coordination Office (SCO). The primary mission of the organization became research needs assessment and technology transfer. Technology means scientific knowledge in the form of hardware, processes, methods or ideas. Here, technology transfer refers to the application of technology produced by ARI researchers for a new user and in some cases, putting old results to new use. Since one-third of the U.S. Army is stationed in Europe and numerous ARI technologies are developed for use by the active Army in the field, it is essential that products be transferred and needs be considered within this setting. While identifying research needs and promoting technology transfer, we have identified gaps in what should be a comprehensive research and development (R&D) management process. The purpose of this paper is to clarity the need for an R&D management approach which is accountable for research including needs assessment and product utilization with feedback from the user.

Highlighted here are our conclusions developed as we searched for a model against which we could test our experiences in needs assessment and with which to guide and judge our effectiveness in disseminating knowledge and facilitating implementation of research products in an operational setting. In searching for a model, we were, in part, seeking to define the role of and the support system needed for those who carry out these functions in the military setting.

Dissemination of research results, utilization and the relationship to needs assessment are not new ideas. A great deal has been written on ways to disseminate knowledge and get it used, and about the critical variables, barriers and gateways to be encountered. Glaser (1983) compiled an extensive review of research on the topics and reviewed several models. The bibliography alone is 162 pages. The demand for better needs assessment analysis and product utilization in the military, in particular, have been addressed by military psychologists such as Freda (1980) and Shields (1977). For example, Shields found that some technology is under-utilized or totally rejected by the intended users because the need for the product did not exist, the product was flawed from a user standpoint, or the researcher or developer did not attend to the complete process of R&D management.

What is new here are our discoveries of what is possible now, and what further support is required as we perform these functions in an operational environment. Also unique is that we may be the only unit in the R&D community with these functions as the main mission of the organization.

This paper is an unofficial document intended for limited distribution to obtain comments. The views, opinions, and/or findings contained in this document are those of the authors and should not be construed as the official position of ARI or as an official Department of the Army position, policy, or decision.

We have speculated and forayed into our environment with a number of simultaneous actions in accomplishing our mission. We have had successes, and at all times we have come back to the central question of how our activities do fit or should fit into a system of R&D management and a model of knowledge utilization. Our aim is to do more than "muddle through" with occasional immediate successes; it is to influence the further institutionalization of our activities and provide wider support for such action throughout the military and the R&D community. Within the military, centralized management of R&D may be too big to handle; but, a change in the sequencing of and habitual relationships within the R&D community and between R&D people and users still needs to be considered.

One way to provide institutionalization of R&D management for this purpose is on a case-by-case basis, that is, a case management system for research products. If such longitudinal management existed in a clearer, more concise form, to include greater involvement with the users, we would be in a better position than we are now. However, our experience suggests to us that product case management may not be the most effective procedure. Such a potentially regulation-based system could solidify horizontal communication in organizations or result in paperwork drills. In addition, the concentration on a product-based system could limit the perception and creativity of those responsible. The longitudinal mindset of such a system could lead managers to see R&D management as a process where each product has a definite start and end unconnected to other products and issues. Rather than a product case management process, the ultimate support system for scientists engaged in these activities would recognize research needs assessment and technology transfer as two sides of one coin, the activities occurring in a cycle and occurring simultaneously for many products. The common element to be managed in this setting of multiple activities could be not products, but concepts in relationship to central issues.

We are interested in institutional support for the scientist operating in the role we are defining - one who deals simultaneously with a number of user needs and research products and seeks to bring products to the intended user, as well as others who could benefit from the knowledge. At the same time, information is transmitted about new users and new needs to those scientists producing technology. These needs might be answered with new research or existing products which can be implemented. This in turn may suggest new needs relevant to the product implemented, as well as to other research, so that the full cycle has occurred. The cycle is not performed for one product and one need. The scientist may have several needs or sets of needs in mind for various organizations, or several possible products and be at various stages of the cycle in different settings. R&D management processes could create a support base for performing these activities. The research process which receives the most support and management now is the normal scientific method or process of problem review, hypotheses generation, etc. The role being described here of making broader linkages and connections takes us to areas which are previous to problem review and which follow research project completion, and even follow product utilization.

The continuing questions are: How do we organize support for the scientist in this role? Who in effect is a research manager? How do we organize the work itself to keep it flexible and avoid bureaucracy which could stifle communication needed to identify new uses for technology? The role is that of a gatekeeper of established channels of information exchange, and of a gateway creator who must ascertain focal points where "gates" can be created between the huge systems of the military and the R&D community. Where are the interfaces between the two systems most permeable and facilitation most likely to allow information to be passed and used? In this age of information, it might require only a few people to manage these aspects of the R&D process given access to the right tools and the proper vantage point or perspective. The individual in this role needs access to the entire R&D community to assist in managing research in this sense. The database or information network to support such work needs to include not only technical information about current products and their development, but information about issues and trends of primary importance to the military and information about new concepts of technology.

A few examples may help show the reader where we see opportunities to organize successful exchanges of information, to facilitate the process, and suggest management needed for a support base. Keep in mind that exchange means information passed both ways. In effect, the scientist involved in this aspect of research has two major groups of "clients:" The R&D community and the military user.

There are a number of ways new research needs are identified. As noted above, need identification is often performed simultaneously with implementation or information dissemination. At our office in particular, one form this multiple operation takes is what we call Technical Advisory Service (TAS). In this function, military users make requests for our assistance based on their perceived operational needs, and their knowledge of our expertise which we have made known in the community.

Our first example concerns TAS being carried out at the Warrior Preparation Center (WPC). The WPC is a joint operation by USAREUR and the United States Air Force in Europe (USAFE) to support command and battle staff training. The center provides training for the operational level of warfare for corps through Army Groups in joint exercises with the Allied Tactical Air Force (ATAF). Our organization had originally approached WPC to possibly assess needs at this new center which might be addressed by ARI research. After the WPC staff members became familiar with the SCO mission, functions, and areas of expertise, they began to identify operational and developmental requirements which could be addressed by the ARI SCO.

In this case, establishing a relationship and facilitating discussions with WPC resulted in benefits to both organizations. We have been able to provide short term TAS yielding ideas and answers for WPC operational development, specifically in terms of data collection and data analysis. The work was done by one SO psychologist and by one scientist "loaned" from ARI HQ for an intensive six week consultation to WPC. By applying our consultation skills at WPC and other organizations, we build credibility for ARI in USAREUR and lay the foundation for a positive reception of future introduction of ARI technology, and we create support for the process of our work. The WPC organization helps us do our job of identifying research needs as they continue a dialogue with our office, in which they describe exercises, issues of measurement, and performance at corps and echelon above corps level. In the future, we believe we might be able to introduce to WPC results from ongoing ARI research on corps performance and use feedback from WPC as guidance for further research and development. The lesson learned is that consultation skills and establishment of networks or relationships provide the basis for our role.

Our second example concerns training technology being developed by ARI for the Army Training Battle Simulation System (ARTBASS), which is to be implemented as the Army's primary system for training battalion command groups in the command and control of combat operations. ARTBASS uses a computer to simulate unit actions and weapons effects, and to calculate the changing status of personnel, equipment, ammunition and fuel in simulated combat. The system is portable, and the Department of the Army plans to field enough systems to train every maneuver battalion in the active Army and every Reserve component.

In our larger efforts to clarify issues and needs in command and control below corps in USAREUR, we attempted to ascertain who the users for ARTBASS training technology are in USAREUR and how the training technology being developed by ARI might be fielded along with the new system it supports. Previous discussions with military users in the field had led us to identify an important training issue in USAREUR: Supportive training products frequently arrive much later, if ever, than new training systems or new equipment. The need to make connections was clear. However, our experiences in the transfer of training technology for the ARTBASS raised more issues than answers about the R&D process vis-â-vis our role.

Our attempts to make linkages in USAREUR and to have prototype ARI training products for the ARTBASS examined were well-received by the 7th Army Training Command (7ATC) in USAREUR, the primarily responsible organization for implementation of ARTBASS. Likewise, the Army Materiel Command Europe (AMC) encouraged our coordination with them in their role in fielding the system. Our informal information exchange and coordination appeared to be holding up well within USAREUR; however, a subsequent meeting with a representative from the U.S. Program Manager for Training Devices (PM TRADE) office left us with questions about our role. The representative from the agency primarily responsible for overseeing development and initial implementation of ARTBASS was unaware of who ARI was and why coordination in the USAREUR environment was desirable. Coordination for fielding of training technology developed by ARI was not on the list of "things to do" when fielding new devices. Formally, at least to this representative, ARI was not considered as a contributor to the overall project. When we were questioned about how centralized coordination of training technology and new devices was done in the States before fielding to USAREUR, we had to respond that we did not know. The representative's proposed efforts to rectify the lack of coordination were in terms of a centralized, one product or one case fix, rather than a systems approach. Our coordination with the U.S. representative virtually came to a halt.

At this point we are left wondering. We have been told numerous times that new devices and equipment often arrive in USAREUR without programs of instruction or other types of supportive training materials. At the same time, we were often able to point to finished ARI products which could assist in these situations. It fell to the SCO in Europe to pursue the connections. We have not yet answered the questions. Should the coordination role be assumed by a central R&D management level? Are many decentralized efforts necessary? When and with whom? We continue to build our linkages in Europe and pursue the expansion of linkages in the entire R&D community. It is possible, however, that we will find ourselves outside the system as we create our informal network.

Our third example concerns a product for which our lateral communications have identified potential new users, and for which we are assisting in a test and evaluation process to boost user involvement in the product's development. The success we have in making connections is based largely on the fact that a new type of program has recently been initiated and based at HQ ARI to provide research product management and which provides some formal mechanisms to support a role such as we are developing.

The Joint Service R&D Program sponsored by DOD was initiated to offer end users the capability to participate in the development of prototype technologies. One product from this program under development jointly by the Navy and Army is the Personal Electronic Aid for Maintenance (PEAM). This highly portable, interactive device would replace maintenance manuals and provide an aid to performance of Organizational Maintenance. The initial application is for the Turret Mechanic of the M1 Abrams Tank.

Our involvement with the Joint Service program combined with our awareness of needs and issues in Europe has permitted the SCO to make connections which may expand the use of PEAM to address a related but different need beyond its planned application. We are currently involved in evaluating the feasibility of expanding the portable, step by step electronic guide into versions for use by German maintenance mechanics who do not use the English language manuals well. In addition, other products based on the concept of easy access to procedural knowledge might become involved in addressing the maintenance issue.

We see the PEAM example as an important aspect of the total R&D process. Within the greater R&D community when these types of connections are made, they tend to be as a result of personal efforts and interests. In the case of PEAM, the Joint Service R&D Program sets the stage for a broader view of research and development. When this type of R&D management is combined with the flexibility to share information and communicate about an issue laterally through military operational settings and R&D organizations, good ideas can possibly surface more easily and become reality.

Our final example concerns our activities at 7ATC where informal linkages built through our use of consultant skills has created gateways through which finished ARI products have been delivered to users.

7ATC is responsible for planning, developing, managing, and coordinating training requirements and programs for USAREUR. An initial visit with 7ATC staff members showed that they were generally unaware of ARI, and yet very interested in the type information ARI was capable of providing. After our learning about 7ATC's mandate and operations from staff members and written materials, we briefed down through the 7ATC chain of command, addressing what we identified as their staff concerns, to establish a formally recognized gateway to this organization.

We began our work with one directorate at a time, first establishing needs and delivering information, to establish a "track record" within the organization of delivering what we promised, before going to the next area. We currently play a consulting role, visiting 7ATC on a routine basis, and being called in to address specific problems. This allows for follow-up of previous work, and identification of new needs. It must be recognized that there is only limited organizational memory within 7ATC directorates. When a staff member leaves, our work essentially leaves with him. Likewise, when a new staff member comes on board we must brief him on our mission and provide a package of materials that identifies what has, and can be done to support staff members' job requirements.

There is still no routine delivery of ARI products at 7ATC for several reasons. The 7ATC staff carries a very heavy workload, and does not have the luxury of reading ARI reports, even when these are readily available. We have learned through experience that staff information needs must be clearly defined, and applicable ARI material discussed one-on-one if the information is to be recognized and applied to the 7ATC mission, and eventually in USAREUR. Using this approach, we have been able to identify and address a number of information needs using past research products and current ARI draft research materials. We have also been able to identify future products, both short-term, and long-term, which address 7ATC needs, and have provided these as they were released. Based on research product feedback and new operational needs at 7ATC, the SCO currently identifies needs for incorporation into ongoing ARI research, thus continuing the needs assessment - technology transfer cycle.

In conclusion, we are looking at research needs assessment, information dissemination, and utilization, which comprise a scientific field of continuing development - one in which a number of models have been suggested but none are definitive. We are attempting to create not only gateways to exchange information, but relationships which can be accessed through the gateways and through which the user can become a fuller part of the R&D cycle to influence products both before and after their creation. As we perform the activities, we are constantly examining our actions and assessing how they fit or do not fit smoothly into overall R&D management. We are searching for system solutions to facilitate our roles. If we do not, we will always be treating the military needs on a band-aid basis, and when we, the gatekeepers leave, the gates may close. By defining our role and identifying the types of support and credibility needed institutionally for the role, we hope to get consistently more use from each research dollar on a long-term basis. The payoff is in effective, user-oriented research management.

References

Freda, J. S. (1980). Army Training Technology Transfer: A Systems Model. (Research Report 1241). U.S. Army Research Institute.

Glaser, E., Abelson, H., & Garrison, K. (1983). Putting Knowledge to Use: Facilitating the Diffusion of Knowledge and the Implementation of Planned Change. San Francisco: Jossy-Bass Publishers.

Shields, J. L. (1976, October). "Training Technology Transfer." Proceedings of the 15th Annual Operations Research Symposium. Ft Lee, VA.

The Reduction of Standard Errors of Equipercentile Test Equating through Negative Hypergeometric Presmoothing*

Benjamin A. Fairbank, Jr. Ph.D.
Performance Metrics, Inc.
5825 Callaghan Rd., Suite 225
San Antonio, Texas 78228

I. INTRODUCTION

Test equating is the process of finding which scores on two or more similar tests correspond to the same level of ability in the population of potential examinees. The need for test equating arises as a result of many considerations. It is often valuable to have more than one version or form of a test. When more than one version or form of a test is available, the particular form taken by an examinee should not affect the examinee's expected score.

The replacement of operational tests requires equating when the scores on the new tests are to be used in the same predictive or evaluative equations or in the same manner as were the old scores. Test equating may be carried out in any of a large number of different ways, some of which are of recent origin and are technically sophisticated, and some of which have been in use for several decades (see Holland & Rubin, 1982). This report addresses only equipercentile test equating as applied to two equivalent groups (Angoff, 1971). The terms "reference test" and "experimental test" are used to indicate, respectively, the test whose score metric is to be used for the results of both tests, and the test whose score is to be converted to the units of the other test. For example, if an existing test known as Form K is to be replaced by a similar test known as Form M, Form K would be the reference test and Form M would be the experimental test.

As with any procedures having the goal of estimating population characteristics based on data obtained from a sample, there are always sample-dependent errors present in test equating. If an equipercentile equating were to be done twice with similar samples, the results would differ. The extent of such differences has been estimated by Lord (1982) and their magnitudes appear as the standard errors of equipercentile equating. As with all standard statistical procedures, the size of the expected errors decreases linearly with the square root of the sample size. It is thus operationally impractical to reduce errors beyond a certain amount by increasing sample sizes. For example, decreasing the error to one-fourth the size of the error associated with a given sample size would require using a sample 16 times the size of the original sample. As a consequence, practitioners of equipercentile test equating have looked for other ways to reduce equating errors. They have most frequently used the methods of smoothing.

Smoothing

Two general classes of smoothing methods were used. A third class is made up by combining a smoothing method from the first class with one from the second class. First, presmoothing is defined as the process of smoothing the observed score frequency distributions prior to the equating. Second, postsmoothing is defined as the process of

*This report is excerpted from a technical report by B. Fairbank, Jr. titled "Equipercentile Test Equating: The Effects of Presmoothing and Postsmoothing on the Magnitude of Sample-Dependent Errors," AFHRL-TR-84 64, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas, 78235. The research here reported was supported by a contract between the Air Force Human Resources Laboratory and Performance Metrics, Inc. The opinions are those of the author. The full technical report is available from the author.

smoothing the equipercentile points after equating. Third, combined smoothings involve presmoothing and postsmoothing applied consecutively. The common intent of all three smoothing methods is to remove small sample-dependent fluctuations from the nonsmoothed equatings so that the small sample equatings will more nearly approximate the asymptotic equatings, or those which would result from the use of samples so large that the sample-dependent errors approach zero. The extent to which the various methods achieve this common intent is investigated by this research. Seven presmoothing methods, seven postsmoothing methods, and five combined smoothing methods were used as follows:

A. Presmoothing Methods
  1. 3-point moving medians
  2. 5-point moving medians
  3. 3-point moving weighted averages
  4. 5-point moving weighted averages
  5. 5-point moving weighted averages with root transformation
  6. 4253H Twice
  7. negative hypergeometric

B. Postsmoothing Methods
  1. linear regression
  2. quadratic regression
  3. cubic regression
  4. orthogonal regression
  5. logistic ogive
  6. cubic splines
  7. 5-point moving weighted averages

C. Combined Smoothers
  1. negative hypergeometric + orthogonal regression
  2. negative hypergeometric + quadratic regression
  3. negative hypergeometric + 5-point moving weighted averages
  4. 3-point moving weighted averages + 5-point moving weighted averages
  5. negative hypergeometric + cubic splines

The final presmoothing method (see Keats & Lord, 1962; also Lord & Novick, 1968, pp. 515-520) is one devised explicitly for smoothing or fitting frequency distributions of test scores. The distribution is the negative hypergeometric, whose appropriateness is derived from a binomial error model of test scores. The model assumes several technical conditions, one of which i equivalent to the assumption that all of the items on the test whose score distribution is being fit are equally difficult. That condition is known to be false in the case of most tests, but the fit of the negative hypergeometric is still good enough to make it promising for further study (Keats & Lord, 1962).

Objectives

The aim of the present effort was to evaluate the effects of various different methods of presmoothing, postsmoothing, and combined smoothings on the accuracy of test equating. The study was exploratory in nature, designed to determine which method hold the most promise for operational use.

II. METHODS

General Plan

The plan underlying this investigation was to use three different approaches to determine the effectiveness of each of 14 unitary smoothing methods and five combined smoothing methods. The first approach used simulated tests and examinees; the second and third used data from tests administered to examinees under operational conditio The advantage of simulated tests and examinees is that all quantitative aspects of t tests and examinees are completely specified, and it is possible to know in advance t results of theoretically errorless equatings or those equatings which are unaffected sample dependent errors. Operational data, of course, have the advantage that they a [illegible] conditions typical of the ones under which smoothing methods would [illegible] data are not based on an ideal model, as are the data from simulations; rathe they contain all of the departures from theory that may be found in operational tes settings.

The first of the three methods of evaluation involved comparing each of t smoothed equatings with a known errorless equating. The known errorless equating wa based on a method that yielded results typical of an equating using an infinitely larg sample. The method requires deriving a distribution of expected score frequencies, th [illegible] being that which would result from administering the test to a sample so larg that the observed proportions at each score were observed essentially without error. T [illegible] the simulated test administrations were used for that method. The secon [illegible] was a similar comparison of sample and criterion equatings, but in place of dat [illegible] on simulations and on an errorless equating, the comparison used operational obtained data and an equating based on an unusually large sample size. The third meth was to use the statistical jackknife (Mosteller & Tukey, 1977) to estimate the size o standard errors of smoothed and unsmoothed equatings using operationally obtained dat and simulated data. Those errors were also compared to standard errors computed means of the formula given by Lord (1982).

Each of the six simulated examinations was "taken" by 100 groups of 2,000 simulees [illegible] of two methods was used to administer a test in simulation. The first method [illegible] to that given by Ree, (1980). The second method involved taking a sample of 2,00 observations at random from the Expected Observed Score Distribution (EOSD, see belo [illegible] test score distributions were tabulated for each simulated administration. [illegible] each test length, 100 equipercentile equatings and smoothings were then performed usi [illegible] methods described below.

The method used to establish the criterion equatings for the simulations used in t [illegible] is based on the EOSD for each test. An algorithm developed by Lord a Wingersky (1983) was used to prepare distributions of expected observed scores for ea of the six simulated tests. In an EOSD, each score has associated with it a proportion [illegible] not a frequency. The distributions model the result of administering the te [illegible] large number of examinees and observing the relative frequency of ea [illegible] comparing the small sample equatings (N=2,000) with those that result from [illegible] (i.e., those based on the EOSD), the extent of improvement result [illegible] is directly observable. The criterion equatings, then, are the unsmoothe equipercentile equatings which result from using the EOSDs in the unsmoothe equipercentile method.

[illegible]

The operational data that were used were taken from a set of test scores for [illegible] large sample sizes (approximately 100,000 examinees) for three roughly parallel form [illegible] subtests. Among those subtests were two forms of one test of lengt [illegible] and two forms of another test of length of 20 items. In addition to t frequency distributions of test scores for all examinees, there were available 100 samp of 2,000 scores for each of the four subtests. The samples were drawn at random witho replacement from the larger samples of 100,000 examinees.

For the operational data, criterion equatings were established by using the f sample of 100,000 examinees. As with the simulated data, the criterion equatings we unsmoothed equipercentile equatings. As with the simulated data, 100 reduced samp equatings were made for each of the test pairs both without smoothing and with each the [illegible] methods.

Equatings

All test equatings were performed using the equipercentile method described by Lindsay and Prichard (1974). For the unsmoothed equatings and the equatings to which only postsmoothing was to be applied, the raw frequency files were equated. When the equatings involved presmoothing, the smoothed frequency estimates were equated. Following the equatings and smoothings, each test or simulated test had associated with it a criterion equating, an unsmoothed equating, and 19 smoothed equatings, one for each of the smoothing methods used.

Analysis of Equating Results

Each of the five tests, three simulated and two operational, had associated with it one criterion equating, 100 unsmoothed equatings based on sample sizes of 2,000 (called the "small sample"), and 100 sets of 19 smoothed equatings based on the same samples. The question of interest is the effect of the smoothings on the accuracy of the equatings. A deviation is a difference between an equated score obtained with a small sample and an equated score based on a criterion equating. At each observed (i.e., integer) score on the experimental test, the corresponding score on the reference test was found using the criterion equating. The equated scores were found as decimal fractions not rounded to the nearest integer. The score corresponding to the same experimental test score was then found for the unsmoothed small sample equating and for each of the 19 smoothed equatings. The differences between the equated score based on the criterion equating and the equated score based on the small sample equatings were found for each possible score on the experimental test, for the unsmoothed and for the smoothed equatings, for all 100 replications. These differences, or deviations, were the raw data used for evaluating the smoothings. For each of the 100 small sample equatings, the deviations at each score were combined across equatings to give a general measure of deviation at each score. Three such deviation measures were computed.

The first measure is the Root Mean Square Deviation (RMSD), found by taking the square root of the sum of the squares of the deviations across all 100 samples. The second measure is the Average Absolute Deviation (AAD), or simply the mean of the absolute value of the deviations computed across all samples. The third measure is the average of the signed values of the deviations (ASD), found by taking the mean of the deviations across all 100 replications. ASD differs from AAD in that the absolute values are not found before the mean is computed. ASD is sometimes called "bias," or "statistical bias," but in the context of testing the term "bias" denotes other phenomena and so is less appropriate than "ASD."

III. RESULTS AND DISCUSSION

It was found that with some smoothing methods, especially the presmoothing methods, smoothing resulted in large increases in the deviation measures for very low test scores. In some cases the increases were so great that graphing them required such a large rescaling of the figures that the more important deviations in the middle ranges of the test could not be represented. These large induced deviations are seen as being of little interest because they occurred at score values which were lower than the guessing level on a test, and so were not associated with meaningful measures of ability. In order to show the more relevant deviations effectively, the figures do not present information on the levels of RMSD, AAD, or ASD at test scores below the guessing level for each test.

Of the 14 smoothing methods, negative hypergeometric smoothing was uniformly most effective in reducing root mean square error. The results of that method are shown in Table 1.

Table 1. Summary of the Averaged Effects of Presmoothing by the Method of Negative Hypergeometric

Test Length	Proportion of Mean Deviations RMSD	 AAD	 ASD
		Simulated Tests	
15	.891	.903	2.919
30	.865	.867	3.596
50	.852	.861	3.453
		Operational Tests	
20	.905	.908	2.008
25	.966	.989	7.479
Mean	.896	.906	3.891

Note. Tabled values represent RMSD, AAD, and ASD as proportions of the values found without smoothing. Averages taken over all samples at all scores above chance level.

To evaluate the effects of smoothing, particularly its effects on deviations, it is helpful to consider such deviations within the context of the accuracy of ability or achievement tests more generally. The standard errors of equating discussed in this report are not the only measurement errors which arise in the testing process. There are also standard errors of measurement that are intrinsic to any test which is not perfectly reliable. The following formula (Allen & Yen, 1979) relates reliability (R), standard error of measurement (SE), and test score standard deviation (SD).

$$SE = SD * \sqrt{1 - R} \ .$$

Thus, the standard error of measurement for the experimental test of length 15, based on a reliability (KR-20) estimate of .80 and a standard deviation of 3.28, is 1.47. Similarly, the standard error of measurement for the experimental test of length 30 is 2.20, and that for the experimental test of length 50 is 2.74. The corresponding average standard errors of equating, as estimated by Lord's formula, are .15, .30, and 51. Thus the standard error of equating ranges from approximately only 10 to 20 percent of the standard error of measurement.

The results of smoothing by the negative hypergeometric are the only ones which show consistent improvement in RMSD and AAD as a consequence of smoothing. The effects are particularly impressive with the simulated tests, presumably in part because the criterion equatings for those tests are nearly perfect, not estimated from very large samples. The gains are not uniform across the tests. On the shorter tests at lower scores, the measures of RMSD and AAD actually increase as a consequence of using the negative hypergeometric. The beneficial effects of the negative hypergeomtric do not extend to the measures of ASD. The ASD increases both globally and locally, sometimes quite dramatically. These increases were expected at the lower end of the test, where guessing is a factor, but increases at the upper end were not expected. It must be noted, however, the ASD figures were low initially, so that a tripling of ASD may still denote an acceptably low level.

Why is it that the negative hypergeometric smoothing method outperforms the other presmoothers? It is likely that it is in part because that smoother takes into account all of the information in a distribution's mean and standard deviation in arriving at the smoothed frequency for each point. Although the negative hypergeometric does require the assumption that all items are equally difficult, an assumption usually contradicted in practice, its success as a presmoother indicates that its use is robust against violation of this assumption. Furthermore, among the seven presmoothers investigated, only the negative hypergeometric is based on a mathematical model of testing.

The present study is limited in several respects, all of which may tend to reduce its generalizability to other applications.

First, only five tests were used: two operational and three simulated. Generalizations to other tests may be inadvisable if the tests do not statistically resemble those used for this study.

Second, the tests used, especially the simulated tests, may be more similar to each other than are most operationally equated tests. Generalization to less similar tests is of questionable appropriateness.

Among the presmoothing methods, the negative hypergeometric and, by extension, other smoothers of the same beta binomial family, deserve consideration for operational use. If any of the presmoothers studied here is to be adopted, then the negative hypergeometric would be the most appropriate. It has the effect of reducing RMSD by about ten percent, a benefit which could also be achieved by increasing sample size by about 20 percent.

References

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Belmont, CA: Wadsworth.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education, 508-600.

Holland, P. W., & Rubin, D. B. (1982). Test equating. New York: Academic Press.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. Psychometrika, 27, 59-72.

Lindsay, C. A., & Prichard, M. A. (1974). An analytical procedure for the equipercentile method of equating tests. Journal of Educational Measurement, 8, 203-207.

Lord, F. M. (1982). The standard error of equipercentile test equating. Journal of Educational Statistics, 7, 165-174.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1983). Comparison of IRT observed score and true score equatings. (RR-83-26-ONR), Princeton, NJ: Educational Testing Service.

Mosteller F., & Tukey, J. W. (1977). Data analysis and regression. Reading, MA: Addison-Wesley.

Ree, M. J. (1980). AVRAM: Adaptive vector and response automation method. Applied Psychological Measurement, 4, 277-278.

WEIGHTED LIKELIHOOD ESTIMATION OF ABILITY IN ITEM RESPONSE THEORY WITH TESTS OF FINITE LENGTH.

Thomas A. Warm, Ph.D.
U.S. Coast Guard Institute

A new method of statistical estimation, Weighted Likelihood Estimation (WLE), was discovered in Warm (1985a), and a new theorem of mathematical statistics was proved in Warm (1985b). The theorem states that WLE has zero first order bias, in contrast to Maximum Likelihood Estimation (MLE) and Bayesian Modal Estimation (BME) which are both biased. WLE was applied to ability (θ) estimation in Item Response Theory (IRT).

METHOD

Using Monte Carlo methods, WLE(θ) was compared to MLE(θ) and BME(θ) on 12 conventional tests with 10 to 60 items, a-parameters of 1 or 2, and normally distributed b-parameters. The three estimators were also compared on two tailored tests with optimal b-parameters. One tailored test had an infinite item bank and all a = 2.00 . The other tailored test simulated a finite item bank with declining a-parameters. For all tests all c = 0.20 .

RESULTS

Partial results are presented in the figures below. For complete results, see Warm (1985b).

In both conventional and tailored tests WLE(θ) was less biased than both MLE(θ) and BME(θ). In addition WLE(θ) had small variance over the entire range of the θ-scale, as well as small mean squared error even at non central θ.

DISCUSSION

The relative unbiasedness of WLE(θ) makes this estimator particularly appropriate in applications of IRT for which the parameter invariance property is important.

Two new insights for MLE(θ) were discovered: 1) natural, rational bounds, and 2) a conditional analogy to the attenuation paradox in tailored tests with high a-parameters.

The heart of WLE(θ) is a weighting function, w(θ), which is multiplied times the likelihood function, and the product maximized. This weighting function, which removes the bias and uncontrolled variance of MLE(θ), is a function of θ and the item parameters, and is specific to each test. It was shown to be equal to the square root of test information for the one- and two-parameter models of IRT, and equal to a closely related function for the three-parameter model.

REFERENCES

Warm, T.A.(1985a) Weighted Likelihood Estimation of Ability in Item Response Theory. Technical Report CGI-85-01, U.S. Coast Guard Institute, P.O. Substation 18, Oklahoma City, OK 73169.

Warm, T.A.(1985b) Weighted Likelihood Estimation of Ability in Item Response Theory With Tests of Finite Length. Ph.D. Dissertation, University of Oklahoma, Norman, OK. (Also available as Technical Report CGI-85-08, U.S. Coast Guard Institute, P.O. Substation 18, Oklahoma City, OK 73169.)

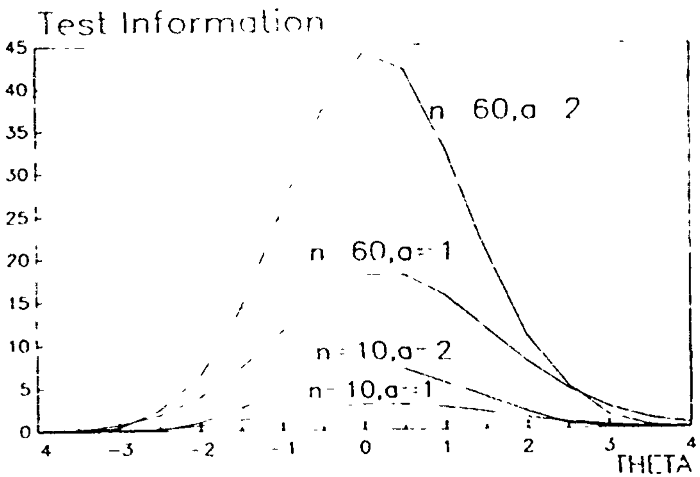


Figure 1. Test Information curves of the shortest (n=10) and longest (n=60) with all a=1 and all a=2.

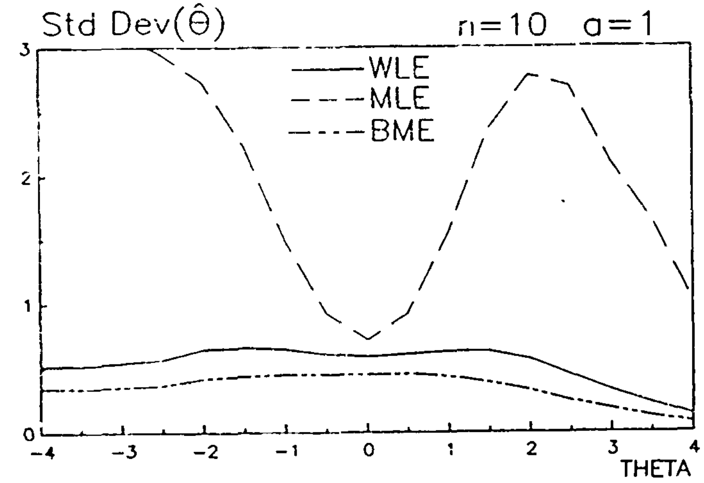


Figure 3. Standard Deviation of θ-estimates for conventional tests. SD(BME(θ)) was smallest, but SD(WLE(θ)) was also always small.

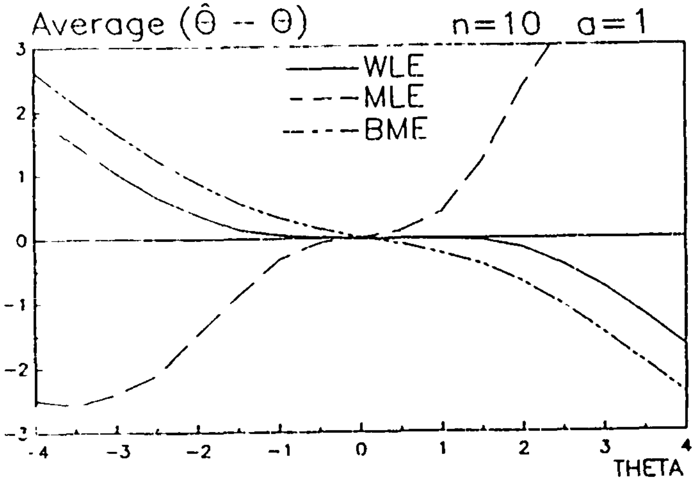


Figure 2. Bias of θ-estimates for conventional tests. WLE(θ) is less biased than MLE(θ) for all tests.

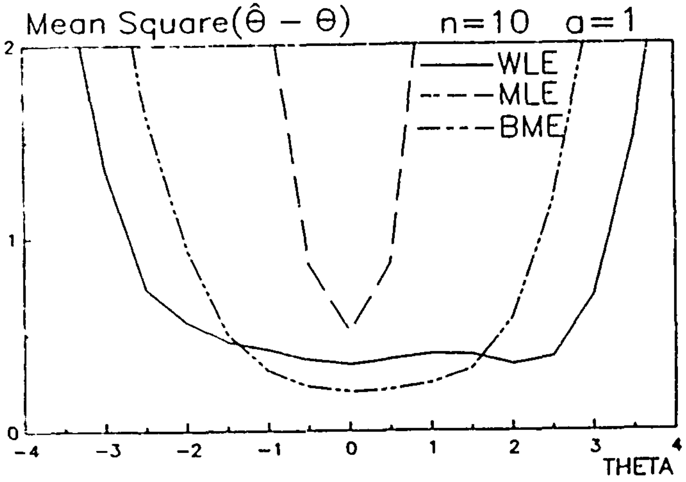
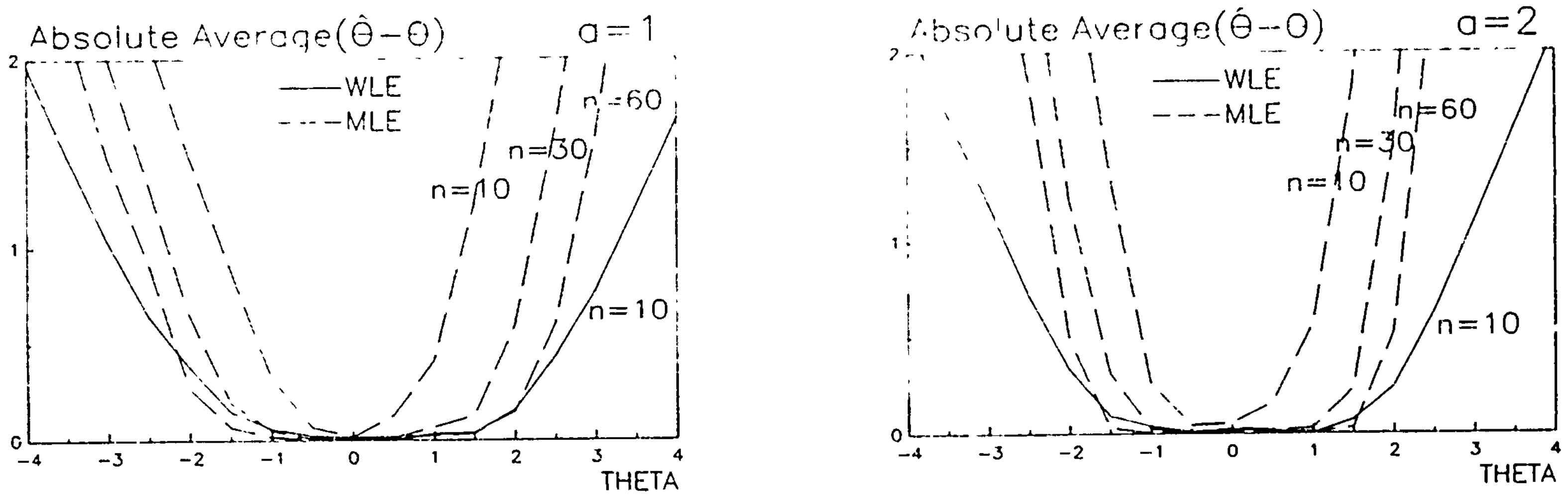
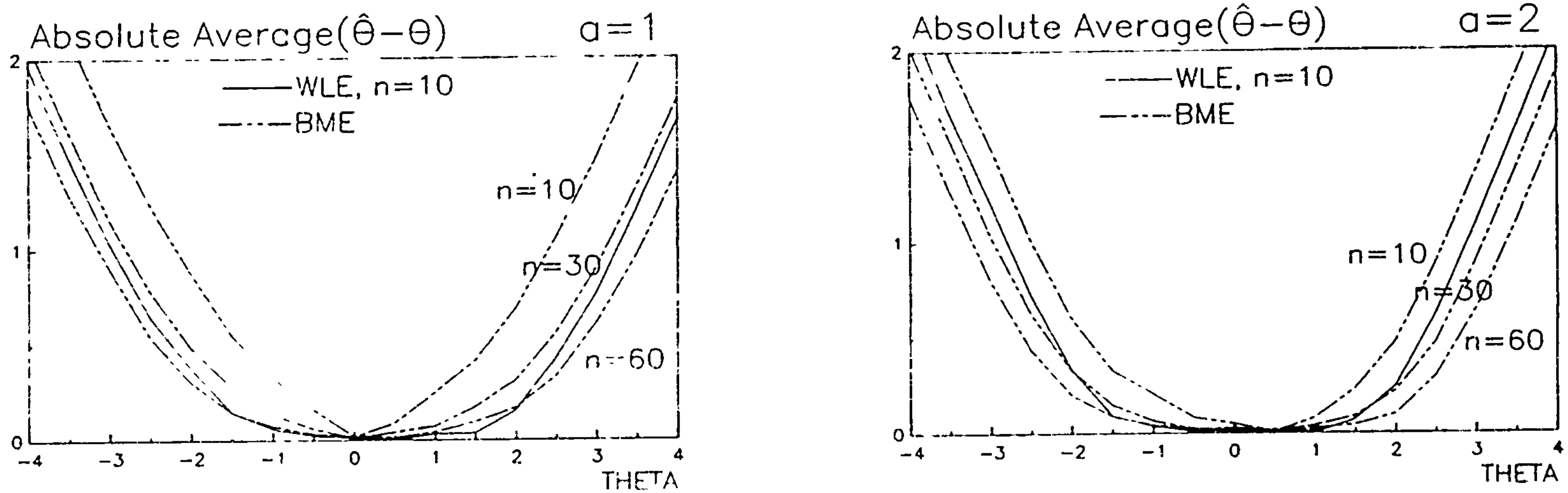


Figure 4. Mean squared error of θ-estimates for conventional tests. The range over which MSE(WLE(θ)) is small was always wider than for MLE(θ) or BME(θ).



Figures 5 and 6. Absolute bias of WLE(θ) and MLE(θ) on conventional tests. WLE(θ) with 10 items is less biased than MLE(θ) with 30 to 60 or more items.



Figures 7 and 8. Absolute bias of WLE(θ) and BME(θ) on conventional tests. WLE(θ) with 10 items is less biased than BME(θ) with 20 to 40 items.

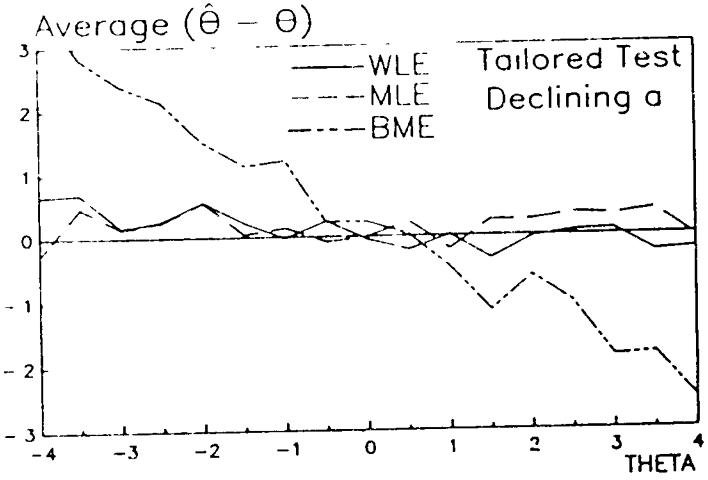
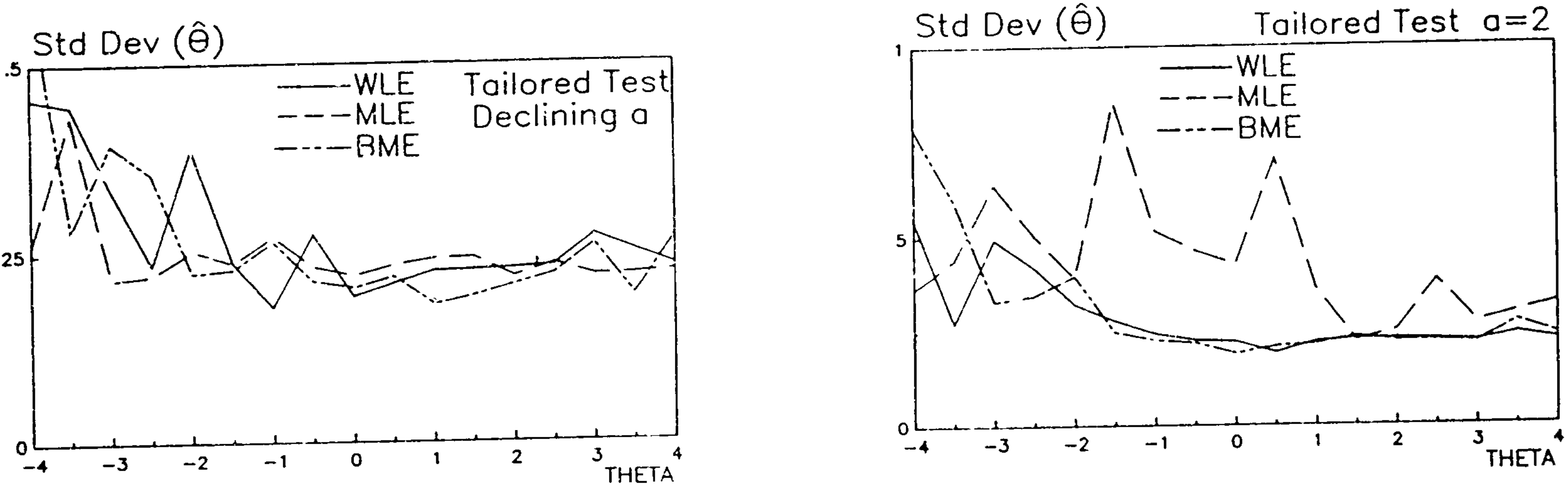


Figure 9. Bias of θ-estimates on tailored tests with simulated finite item bank. WLE(θ) is essentially unbiased. MLE(θ) is slightly biased, and BME(θ) is very biased.



Figures 10 and 11. Standard Deviation (SD) of θ-estimates on tailored tests with finite item bank(declining a-parameters), and infinite item bank(all a=2). With finite item bank SD of θ-estimates are about the same for all three estimators. Wtih infinite item bank MLE(θ) has high variance, exhibiting a conditional analogy of the attenuation paradox.

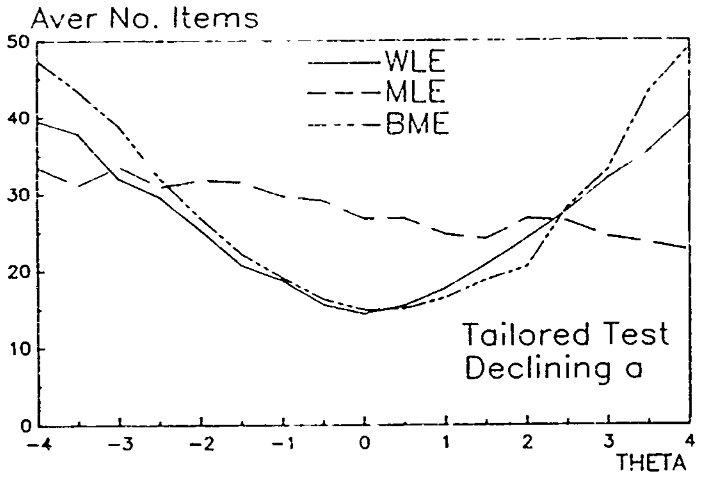


Figure 12. Average numbers of items administered for tailored tests. WLE(θ) used many fewer items than MLE(θ) at central values of θ.

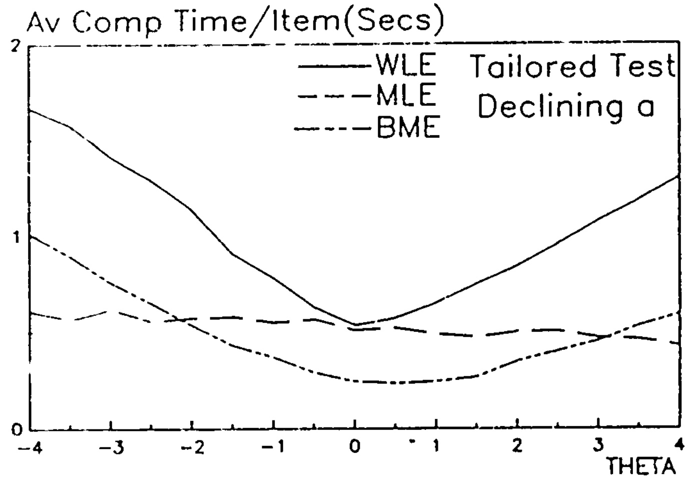


Figure 13. Average computation time (seconds) required between items administered by the tailored tests. WLE(θ) required two to three times as much computation tjme as either of the other two e‹timators.

LEADER BEHAVIOR AND THE PERFORMANCE OF FIRST TERM SOLDIERS

Leonard A. White, Ilene F. Gast and Michael G. Rumsey[1]
U.S. Army Research Institute for the Behavioral and Social Sciences

A large Army project is underway to validate new and current predictors of first term soldier performance. A major objective of this effort is to increase Army organizational effectiveness by improving the soldier job match. This will be accomplished by developing a set of selection and classification measures (predictors) and performance criteria and then empirically demonstrating relationships between the predictors and performance measures.

It is recognized that job performance is not only related to characteristics which are measurable and identifiable prior to enlistment, but is affected by experiences and developmental opportunities that occur throughout a soldier's life-cycle in the Army. The focus of the present research is on the performance-relevant consequences of a soldier's interaction with his or her superiors. Longitudinal research indicates that the quality of leader-subordinate work relationships are predictive of job success (Wakabayashi & Graen, 1984). Aspects of leader behavior such as providing rewards and recognition, disciplinary practices, and inspirational leadership have been related to subordinate effort and performance (e.g., Yukl, 1981).

However, past research on leadership and performance has generally omitted the influence of ability or the potential interactive effect between individual aptitudes and leadership on job proficiency and performance. Some investigations (e.g., Barnes, Potter, & Fiedler, 1983) have suggested that the prediction of job performance from general ability is moderated by leadership. Other researchers (Schmidt & Hunter, 1977) have argued that the relationship between general ability and performance is stable across time and situations for similar jobs.

To summarize, the model examined in this research assumes that job performance is influenced by a new incumbent's capabilities measured prior to enlistment and characteristics of the work environment. Within this framework the purpose of this research was twofold: (a) to examine relationships among dimensions of leader behavior and subordinate performance, and (b) to explore possible moderating effects of leadership on the correlation between general cognitive ability and job performance.

METHOD

Research participants were 696 first term soldiers in five military occupational specialties (MOS); 156 infantrymen (MOS 11B), 139 armor crewmen (MOS 19E), 125 radio teletype operators (MOS 31C), 141 light wheel vehicle mechanics (MOS 63B), and 135 medical care specialists (MOS 91A).

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Army Research Institute or the Department of the Army.

Of these soldiers, 88.5% were male and 11.5% were female; 28% were black, 3% were hispanic, 64% were white, and 5% other. Soldiers' report of work experience in their unit ranged from 2 months to 49 months (median=one year).

Instruments

The first step in this research was to develop measures of leader behavior and soldier performance on the job.

Supervisor behavior rating scales. Critical incidents workshops were conducted with 80 NCO in the five target MOS. These NCO generated a total of 474 examples of leader behaviors thought to influence soldier performance. Classification of the incidents by two of the authors and 31 NCO familiar with Army leadership requirements led to the identification of 9 categories of leader behavior (White, Gast, Sperling, & Rumsey, 1984). At least 5 and no more than 8 items were written to represent important leader behaviors in each category (e.g., Your supervisors are hard to find when you need them). These procedures resulted in a 60-item question naire. Responses to each item were made on a 5-point scale from very seldom or never (1) to very often or always (5).

Job performance rating scales. To develop these scales, critical incident workshops were conducted in which NCO provided examples of effective (as well as ineffective) soldier performance. The number of NCO and examples provided were as follows: MOS 11B, 51 NCO's, and 906 incidents; MOS 19E, 43 NCO's and 798 examples; MOS 31C, 45 NCO's and 830 incidents; MOS 63B, 49 NCO's and 882 incidents and; MOS 91A, 42 NCO's and 783 incidents. A variant of the behaviorally anchored rating procedure (Smith & Kendall, 1963) was used to develop behavior-based rating scales for each job. The resulting rating form for each job consisted of seven to ten 7-point behavior summary scales.

Army-wide performance rating scales. To prepare these scales, 77 NCO's and junior officers working in a wide variety of Army jobs generated 1,215 behavioral examples. The examples represent those aspects of soldier effectiveness that contribute, broadly speaking, to organizational effectiveness, such as following orders and regulations. The target criterion space for these scales went beyond job performance to include aspects of socialization and commitment to the organization. Eleven 7-point behavior summary rating scales were developed for each job.

Hands-on, task proficiency tests. For each of the jobs, 5-8 critical tasks were identified to represent the MOS-specific task domain. Multistep task proficiency tests were prepared for each task. Each step of a task was scored pass or fail. A score for each task was computed by calculating the proportion of steps passed and the task scores were averaged to yield an overall hands-on test score.

Job knowledge tests. Through job analysis, important knowledge areas were identified for each of the five jobs. With the help of subject matter experts, items were written to tap these knowledges. For each soldier, the percentage of correct items was the overall job knowledge test score.

Correlations of hands-on and job knowledge test scores, job performance ratings, and the Army-wide effectiveness rating with the leader behavior scales are presented in Table 2. Results are shown separately for each of the five jobs. A mean correlation ($\bar{r}$) across the five jobs was computed by weighting each correlation by its associated sample size (Hunter, Schmidt, & Jacks^n, 1982). The highest correlations were

Table 2

Correlations between Leadership Scales and Criterion Measures by Army Job

	Leadership Scale								
Job	1	2	3	4	5	6	7	8	Total Scale
				Hands-on Task Proficiency					
11B	.05	.03	.19*	.24*	.17*	.25*	.10	.11	.18*
19E	.14	.11	-.01	.14	.24*	.11	.18	.23*	.22*
31C	.02	.00	.15*	.18*	.02	.08	.03	.15	.09
63B	.12	-.05	.10	.06	.00	.05	.08	.12	.09
91A	-.05	-.14	-.12	.02	-.04	-.08	-.09	-.08	-.10
$\bar{r}$	.05	-.02	.08	.13*	.07	.09*	.05	.10*	.09*
				Job Knowledge					
11B	-.01	-.17*	.02	.13	.03	.15*	.09	.12	.03
19B	-.03	-.03	.05	.03	.06	-.13	-.02	-.03	-.02
31C	.17	.11	.12	.30*	.26*	.23*	.12	.17	.22*
63B	-.05	-.04	.05	.05	.20*	-.05	.01	-.06	-.01
91A	-.12	-.10	-.01	-.01	-.11	-.18*	-.22*	-.23*	-.13
$\bar{r}$	-.01	-.06	.04	.09*	.08	.00	.00	-.01	.01
				Job Performance Rating					
11B	.12	.01	.05	.23*	.08	.21*	.10	.03	.13
19E	.11	.04	.09	.16	.08	.05	.11	.21*	.13
31C	.21*	.01	-.04	.12	.20*	.17	.02	.07	.14
63B	.17	.08	.23*	.08	.20*	.07	.07	.08	.18*
91A	.08	.07	.05	.11	.00	.08	-.12	.01	.03
$\bar{r}$	.13*	.04	.08	.14*	.11*	.11*	.04	.08	.12*
				[illegible]					
11B	.17*	.06	.04	.23*	.12	.20*	.11	.07	.17*
19E	.12	.02	.07	.14	.22*	.10	.14	.15	.15*
31C	.41*	.19*	.12	.34*	.32*	.32*	.18*	.24*	.37*
63B	.20*	.13	.30*	.11	.19*	.06	.04	.08	.21*
91A	.17*	.05	.14	.20*	.07	.09	-.09	.07	.12
$\bar{r}$	.21*	.09*	.13*	.20*	.18*	.15*	.07	.12*	.20*

Note. Leadership scales: 1 (Support); 2 (Informing); 3 (Fairness); 4 (Participation); 5 (Performance Contingencies); 6 (Role Clarification); 7 (Results Orientation); 8 (Training & Development); 9 (Total).

*$p < .05$

General cognitive ability. The Armed Services Vocational Aptitude Battery (ASVAB) was administered to all participating soldiers prior to entering military service. The ASVAB, which consists of ten subtests, is used for selection and occupational classification. A composite measure of four ASVAB subtests, known as the Armed Forces Qualification Test (AFQT), was used as the measure of general cognitive ability.

Procedure

Raters were trained to use the behavior-based rating scales. After training, supervisors in groups of 3-15 evaluated their subordinates on the Army-wide and job performance rating scales. The mean number of supervisor raters/ratee ranged from 1.66-1.83 for the five MOS. Ratings were averaged across supervisor raters to form an overall job performance rating and an Army-wide effectiveness rating for each ratee.

The first term soldier (ratees) completed the supervisor behavior rating scales, and were also administered tests of job knowledge and hands-on, task proficiencies.

RESULTS

Principal components factor analysis was used to examine the dimensionality of the supervisor behavior rating scales. Varimax and promax solutions were computed and the interpretation restricted to factors appearing in both solutions. Comparison of the rotated structures yielded eight factors with eigenvalues greater than one. Items loading above .4 on one and only one factor were interpreted as measuring the factor. Items with weak loadings on all factors or similar loadings on two or more factors were not used to measure any factor. Factor score estimates were computed by unit weighting and summing individual's responses to the set of items representing each factor. Table 1 presents the intercorrelations among the estimated factor scores.

Table 1

Intercorrelations Among Leadership Scales and Summary Statistics.

			3	4	5	6	7	8	9	No. of Items	Scale Mean	Std. Dev
1. Support/Inspiration	.89	.64	.48	.64	.53	.72	.58	.68	.90	9	27.2	7.5
2. Informing		.78	.54	.49	.53	.54	.48	.50	.79	6	19.0	4.8
3. Fairness			.74	.47	.46	.40	.31	.36	.67	5	16.5	4.3
4. Participation				.70	.44	.60	.46	.56	.76	4	13.4	3.4
5. Performance Contingencies					.55	.47	.40	.39	.67	3	9.9	2.5
6. Role Clarification						.78	.55	.63	.80	4	12.9	3.1
7. Results Orientation							.56	.59	.66	3	9.4	2.2
8. Training and Development								.72	.77	5	14.7	3.9
9. Total									.94	39	123.1	24.7

Note. Internal consistency reliabilities are presented on the diagonal.

[a] N - 696

obtained between perceptions of leader behavior and the Army-wide effectiveness ratings. Within the set of Army-wide performance dimensions, strongest relationships were obtained between supportive and participative leadership and ratings of subordinate adherence to regulations and willingness to provide extra effort when needed. Statistically significant but low correlations between leader behaviors and job proficiency were evident in the two combat MOS.

Hierarchial regression analysis was used to estimate the relationships of cognitive ability (i.e. AFQT score), leadership climate, and their interaction to job proficiency and performance. The AFQT score was entered first in the regression to assess the contribution of mental ability at the time of enlistment to later job performance. Then, leadership and the ability X leadership interaction were entered to assess post-enlistment leader influences on performance and the utilization of ability on the job. In the regression model, leadership was represented by the sum of scores on the 8 leadership scales. The criterion variables were job knowledge, hands on task proficiency, and supervisor ratings of job performance and Army-wide effectiveness.

Of interest here, results of the regression analyses revealed no statistically significant increase in $\underline{R}^2$ due to inclusion of the ability X leadership interaction in the model. In each of the five jobs, the highest multiple correlations were obtained for prediction of job knowledge, with $\underline{R}=.30$, to .60, all $\underline{p}<.05$. This effect was primarly attributable to the influence of general ability on job knowledge. Leadership and cognitive ability had significant independent effects on task proficiency in the infantryman and armor crewman jobs with, respectively, $\underline{R}=.28$, $p<.05$, and $\underline{R}=.37$, $p<.05$. However, in MOS 91A and MOS 63B $\underline{R}^2$s for the prediction of task proficiency from the independent variables failed to reach significance. With respect to supervisory ratings of job performance, ability and leadership and their interaction accounted for less than 5% of the variance in this criterion. Leadership showed several significant correlations with Army-wide effectiveness ratings at the zero-order level, however the $\underline{R}^2$ for this criterion achieved significance only in the radio-teletype operator job, with $\underline{R}=.37$, $\underline{p}<.05$. Correlations between cognitive ability and the Army-wide effectiveness rating ranged from $\underline{r}= -.28$ to .03.

DISCUSSION

The present research explored relationships between leadership, cognitive ability, and the performance of first term enlisted soldiers. Results for the five Army jobs examined here support the conclusion that general ability and leadership behavior have independent effects on performance. However, each appears to contribute to effective soldiering in different ways. Leadership, as perceived by the subordinate, had the strongest effect on the motivation-related, dependability facets of performance measured by the behaviorally based rating scales. General cognitive ability contributed to performance by enabling enlistees to learn the facts and procedures required to perform their jobs.

No evidence was obtained indicating that relationships between general ability and job proficiency and performance are moderated by leadership influences. This finding supports conclusions by Schmidt and

Hunter (1977) that the validities of cognitive tests are similar across situations for the same job. Correlations between general cognitive ability and each criterion measure did vary somewhat across jobs, but almost all of the variation was attributable to sampling error.

The relationships between leadership and performance reported here should not be interpreted as indicating that leadership behavior "causes" performance. Leadership effects on performance may be understood in terms of exchange theory (Graen, 1976), which views the interaction between leader and subordinate as a reciprocal influence process that develops over time. Subordinates who are perceived as willing to work hard and support the mission will be evaluated more favorably by their superiors. In return for their support, these soldiers are likely to receive more individualized attention, information, and other resources from their supervisors; which, in turn, serves to reinforce and sustain subordinate effort.

The results reported here are largely exploratory. Future data collection and analysis will provide an opportunity to confirm the leadership factors and to examine potential moderating effects of leadership behavior on a broad range of soldier aptitudes and characteristics.

REFERENCES

Barnes, V., Potter, E. H., & Fiedler, F. E. (1983). Effect of interpersonal stress on prediction of academic performance. Journal of Applied Psychology, 68, 686-697.

Graen, G. (1976). Role-making processes within complex organizations. In M. Dunnette (ed.) Handbook of Industrial Organizational Psychology. Chicago: Rand McNally.

Hunter, J.E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating results across studies. Beverly Hills: Sage Publications.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

[illegible] Journal of Applied Psychology, 69, 603-614.

White, L. A., Gast, I. F., Sperling, H. M., & Rumsey, M. G. (1984, Oct). Influence of soldiers' experiences with supervisors on performance during the first tour. Paper presented at meeting of the Military Testing Association, Munich, Germany.

Yukl, G. A. (1981). Leadership in organizations. Englewood Cliffs, NJ: Prentice-Hall.

MEASURING MILITARY HEROISM

Jeffrey W. Anderson
U.S. Army Research Institute

To a large extent the drama of human history has been woven from the biographies of people who by the nature of their successes beyond the pattern of everyday life assumed a mythical or legendary stature and became national heroes (Downton, 1973). During periods of revolutions and wars the fate of entire peoples has seemed to hang upon the decisions of a few powerful people. Yet, Arthur M. Schlesinger, Charles de Gaulle, and others have remarked that the age of heroes is past (Jennings, 1960). Yet it is also possible that the character of our heroes, or what we recognize as heroism, has changed. We no longer regard our heroes as rare specimens of humanity. They are only somewhat different from ourselves generally due to the situation or circumstances surrounding the act of heroism. While gallant acts have given birth to legends of heroes, scientific research in the area of heroism has been very scarce. The present study is a preliminary attempt to examine four research questions. First, what are the ingredients of heroism? Second, do people still recognize and label certain behaviors as heroic? Third, have these heroic characteristics changed over time? And fourth, can we measure heroism before a crisis?

Heroism involves great bravery, daring courage, valor, gallantry, and intrepidity (Barnhart, 1967). It includes elements of extreme self-sacrificing courage, fulfilling a high purpose, and attaining a noble end (Webster, 1974). But these qualities do not readily lend themselves to measurement and research. In the literature on heroism there are three major foci. The first emphasizes the personality of the hero. It contends that the hero would have been socially recognized as a hero without regard for the circumstances surrounding his behavior. The second focus follows the arguments of Hegel and Spencer. Man is considered a mere product of social forces coincidentally combined in time and space to converge upon an individual that society labels as a hero. Finally, early social reformers and revolutionaries, emphasizing Darwinian concepts, contended that heroes were thrown up by some chance of the natural selection process. The social environment was a selection instrument to provide opportunities for these men to display their hereditary talents in heroic acts. The situation limited the hero, but did not dominate him.

In general, the literature of heroism tells us little about the type of person who is predisposed, with situation permitting, to become a military hero. In other words, while we all seem to know about heroism there is no operational definition of the concept. To attempt to provide an operational definition, we systematically analyzed the citations for heroism in which the recipient was given the highest military award, the Medal of Honor. These citations should also represent our society's general definition of heroic action. The Medal of Honor is presented to a service member who "distinguished himself conspicuously by gallantry and intrepidity at the risk of his life above and beyond the call of duty." As the highest military award for bravery, each branch of the armed forces has established a set of prescrip-

The views expressed in this paper are those of the author and do not necessarily reflect the view of the US Army Research Institute or the Department of the Army.

tive regulations that leave no margin of error or doubt. The deed of the recipient must be proved by incontestable evidence from at least two eyewitnesses, and it must be so outstanding that it clearly distinguishes his gallantry beyond the call of duty from lesser forms of bravery. As such, the citations from Medal of Honor winners are essentially narratives of critic incidents of heroism and well-suited to the development of an operational definition of military heroism (Flanagan, 1954).

Method

At the time of this study the Medal of Honor had been awarded 3369 ti spanning the years 1863 to 1978. From the citations given with these awa 337 were randomly selected for analysis.

A group of three judges, two male and one female, working independentl analyzed these citations to determine the critical behaviors exhibited by hero that were recognized by his fellow soldiers and rewarded by the military. The eight dimensions of behavior that were unanimously listed by al three judges were accepted as truly describing military heroism.

Another group of five judges, two female and three male, were then as to rate each of these dimensions concerning how well it described militar heroism as exemplified by a group of 254 randomly selected different citations. These additional citations were chosen in an effort to broaden th sample and thereby improve the generalizability of the dimensions and findings. In order to achieve these objectives ten percent of the remaining citations or 303 citations were randomly selected. Since many of these citations were verbatim duplicates of others, redundant citations were elkminated, giving 254 citations for further study. These judges had extensive military experience and represented the Army, Navy, and the Air Force. E judge was asked to rate on a 1 to 5 scale the extent to which the derived eight dimensions described the behavior of the award recipient in each citation (from not at all descriptive to totally descriptive).

Since the judges used the rating scales differently, each rating sco was converted to a standard score using the overall mean and variance of judge giving the rating. The interrater reliabilities, using Winer's technique and corrected for five judges, were calculated. Based on these reliabilities, the ratings from all five of the judges were combined and a mea rating was calculated for each performance dimension. The more accuratel dimension described the concept of heroism as defined by a specific cita the higher the judge's rating on that dimension. The dimensions were ra in order of their overall importance in describing military heroism.

Subsequently, ratings were categorized according to the conflict fro which the Medal of Honor citation was drawn. A comparison between confl of relative importance of each dimension to the concept of Military Hero was used to test the implied hypothesis that our societal definition of ism had changed over time.

Results

The first group of three judges were asked to independently derive sions of military heroism based on a randomly selected sample of 337 ci from the Medal of Honor. Dimensions that were unanimously selected by judges were used to describe each citation. All citations could be de using one or more of the dimensions presented in Table 1.

Table 1. Dimensions of Heroism in order of Importance with Interrater Reliabilities

Dimension		r
The hero is thoroughly devoted to accomplish his duty.	(Devotion to Duty)	.47
The hero sets a personal example of behavior for others.	(Personal Example)	.45
The hero risks his own life or places himself in danger.	(Accepting Danger)	.76
The hero rescues or saves another person.	(Saving Life)	.92
The hero overcomes his own injuries or illness.	(Overcoming Injury)	.93
The hero succeeds when the odds are overwhelmingly against him.	(Defeating Great Odds)	.71
The hero takes command or gives leadership when it is lacking.	(Taking Command)	.73
The hero seizes upon an opportunity.	(Seizing an Opportunity)	.57

Essentially, the military hero had to set an example of behavior before and after formal recognition by the organization. He persisted in the accomplishment of his duty and willingly accepted personal danger, subordinating his own life to the values of his cause. Given the opportunity to command (leadership) he did so, and when opportunity presented itself he seized upon it.

These dimensions did not, however, demo"strate any relative importance in describing heroism. A second set of judges, therefore, was asked to rate on a scale of 1 to 5 how important each dimension was in describing the heroism expressed by each of 254 different, randomly selected citations. After converting the ratings given by each judge to standard scores, the interrater reliabilities for each dimension, using Winer's technique and corrected for five judges, were calculated. These reliabilities are also presented in Table 1. While the first two dimensions show lower reliability than other dimensions, all are considered acceptable for untrained raters, especially in light of the unanimous agreement given to these two dimensions by the original three judges.

Based on these reliabilities, the judges' evaluations were combined. The results of this combination yielded the order of importance for the dimensions as shown in Table 1. In all cases, the five judges agreed that the dimensions previously constructed to describe heroism were adequate descriptors of heroism.

To test the implied hypothesis that military heroism has changed over time, each of the judges' ratings were grouped according to the time period of the citation from which the rating was derived and average ratings for each dimension were calculated for seven major conflict periods from the Civil War through the Vietnam conflict. A profile analysis of these average ratings by conflict period is shown in Figure 1.

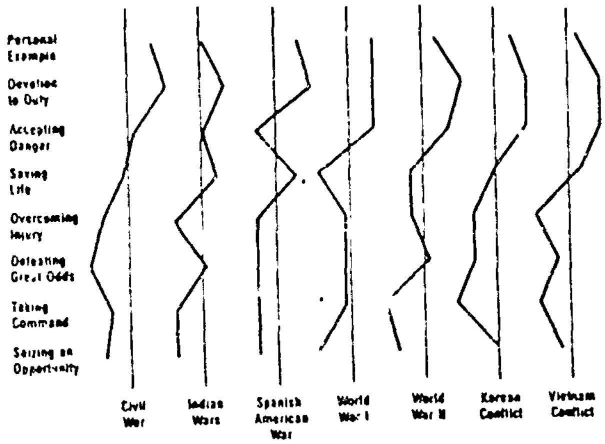


Figure 1 Profile Analysis of Ranks by Conflict

[illegible] is viewed as the mean, deviations to the [illegible] the citations for heroism [illegible] indicate that the dimensions [illegible] citations during that conflict. [illegible] often chosen by the [illegible] period. This analysis [illegible] military leaders to describe [illegible] in Vietnam and [illegible] stable over time.

[illegible] the first three dimensions are interrelated [illegible] matrix was computed. [illegible] are three of the four [illegible] these dimensions are [illegible] of the eight original dimensions [illegible] the eight descriptors [illegible] above [illegible]. Again the first three [illegible] loadings of .60 and above). [illegible] appears best for describing [illegible] measure of the individual's ability [illegible] situation. While this is certainly commendable [illegible], [illegible] seem to indicate by both their ratings and their [illegible] that this is not worthy of the term [illegible] only slightly above 1.0 and [illegible], therefore, [illegible] less reliable than the first factor.

[illegible]

Factor 1

[illegible]

Factor 2

[illegible]

Discussion

[illegible] that there are certain common elements [illegible] three judges to describe heroism [illegible] three dimensions describe military heroism [illegible] definition. The hero sets an example of [illegible] purpose or duty, which is [illegible] own life to complete the task.

[illegible] of military heroism could also describe [illegible] The Warrior Spirit encompasses all [illegible] qualities essential to successful soldiers [illegible] analysis of military history, in a [illegible] spirit as:

1) A selfless devotion to accomplishing a duty or perceived noble cause.

2) Leadership by personal example--especially applying high but achievable standards to himself and his unit.

3) A reasoned acceptance of risk (especially risk of his own life)--calm, confident and self-controlled in the face of mortal danger.

4) Decisiveness despite unreliable, incomplete and often inaccurate information (ability to separate the important from the trivial).

5) Being effective at communicating instructions so that every member of the unit knows and understands the leader's wishes.

6) Creating a team or cohesive unit that all work as one to achieve the noble cause or purpose and training that unit for combat.

The Warrior Spirit,then, appears to be a combination of native characteristics and training. The Fighter Studies (HumRRO, 1957-1958), examined effective combat soldiers during the Korean conflict. They found that a fighter (warrior) tended to (1) be more intelligent, (2) be a "doer", (3) have greater emotional stability, (4) have better health and vitality, and (5) have a greater fund of military knowledge. In 1979 and 1980 studies, Anderson found that successful combat leaders were more intelligent, task-oriented, had higher morale, and had more direct, job-related experience than their less successful peers.

An historical analysis by the Department of History, USMA (1984) found that successful combat leaders had: (1) terrain sense, (2) single-minded tenacity - moral courage (3) ferocious audacity - willing acceptance of reasoned risk, (4) physical confidence, and (5) practical, practiced judgement - common sense.

Based on these studies the hero may be selected in peacetime based on his intelligence, moral courage, character, mental and emotional health, physical well-being (medical and athletic), decision-making ability, common sense, and self-confidence.

Conclusions

This paper has presented a simple, empirical investigation to determine the operational definition of military heroism. It has corroborated this definition with findings from other, related research. Though current psychological literature has little information concerning the psychological profile of a hero, there are certainly indications from the analysis of a different body of literature that military heroism is readily recognizable for others and that the hero has certain measurable characteristics that distinguish him from common humanity. We have shown the underlying factors of heroism and in conjunction with related research propose that certain factors may be measured and used to predict military heroism. Since this study is the first of its kind, there are admitted imperfections, but it demonstrates conclusively that heroism has a psychological meaning for the average individual which may be scaled and reliably measured.

REFERENCES

Anderson, J. W. (1980). The prediction of combat effective leadership (Doctoral dissertation, University of Washington, 1980). Dissertation Abstracts International, 41, 1968b.

Barnhart, C. L. (Ed.) (1967). Comprehensive desk dictionary. Garden City, NY: Scott, Foresman, and Company.

Committee on Veterans Affairs. (1979). Medal of Honor recipients, 1863-1978. Washington, D.C.: Author.

Department of History. (1984). Leadership in combat: An historical appraisal. Unpublished manuscript, United States Military Academy, West Point, New York.

Downton, J. V., Jr. (1973). Rebel leadership. New York: The Free Press.

Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-358.

Human Resources Research Office. (1957). Fighter I: An analysis of combat fighters and non-fighters (HumRRO Technical Report No. 44). Washington, DC: Human Resources Research Office. (DTIC No. AD 158 148).

Jennings, E. E. (1960). An anatomy of leadership. New York: Harper and Brothers.

Webster's new collegiate dictionary. (1974). Springfield, MA: G. & C. Merriam Company.

ANALYSIS OF FUNCTIONAL TRAINING PROGRAMS AT THE NAVAL SUPPLY CENTERS

Neal R. CROWLEY
U. S. Naval Audit Service

In fiscal 1980-1981 $223,649,000 was misplaced, lost, or stolen from the Naval supply centers and management indicators tracking worker performance declined. Congressional concern resulted in several General Accounting Office (GAO) audits and numerous recommendations for improvement. In fiscal 1982 performance indicators improved and $48,974,000 in misplaced material was located.

Responding to strong Congressional and GAO criticism for poor inventory accuracy the Naval Supply Systems Command (NAVSUP) instituted several programs to improve the performance of warehouse workers. As part of the overall effort, NAVSUP started a functional training effort called Competency Based Certification (CBC). The CBC program increased training staffs, improved training facilities, provided money for more training, required the supply centers to develop job related training materials for their warehousing functions and to fully train all employees. CBC required supervisors to "certify" that employees could perform their jobs.

Scope of the CBC Program

Approximately 1,500 warehouse workers at seven supply centers received 79,776 hours of CBC training from 1 August 1982 until 31 December 1984. NAVSUP spent almost $3,500,000 during this time for contractor developed CBC training materials and $500,000 for classroom renovation, printing, etc. NAVSUP hired 45 new training employees to implement the program and issued a new $12,000,000 training contract because of management's belief in the success of the program.

CBC training is very expensive. Contractor costs alone exceed $5,000,000 so far for a per person cost of over $3,333. 79,776 hours of CBC training equals slightly over 53 hours per employee. NAVSUP paid an additional $970,000 for employee salaries while they were being trained and almost as much for the salaries of program administrators and instructors. Spread over the 27 months of the program studied, 53 hours of CBC training gave each employee an average of two hours per month.

Currently, more hours go into administering the program and developing training materials than go into employee training.

There are two training programs. One is based on contractor and supply center training materials. Another is physical distribution training. This consists primarily of greatly intensified training along traditional lines. Employees go to existing training courses ranging in size from one hour briefings to six month professional training courses. The supply centers conducted 226,429 hours of physical distribution training at a salary cost of $2,750,000. The cost of travel, per diem, tuition, materials, etc., for physical distribution training was approximately $500,000. Administrative costs were small.

This evaluation concentrates on the time spent on training and its effect on worker quality, timeliness, and productivity. Has the CBC program worked? To find out I analyzed the program at all seven supply centers. The basic question is whether the amount of training is related to supply center performance as measured by management indicators. If training changed performance, then the change must show up in performance and productivity indicators.

Although assessing the effectiveness of training is the main objective, I also evaluated the relative efficiency of the two training programs.

The primary hypothesis is that there is a strong, or at least moderate, statistical relationship between training and performance. A secondary hypothesis suggests that the relationships between quality, timeliness and productivity can be explained or depicted in a logical and statistically significant way.

Variables Used

I conducted a time series analysis looking for interrelationships between nine variables: quality, timeliness, both quality and timeliness (Q & T), work unit productivity, receipt and issue productivity, overtime, each type of CBC training, and a combination of all training. The quality variable measures how accurately warehouse workers perform their jobs. The timeliness variable measures how quickly workers process materials. The Q & T variable measures both accuracy and speed of worker performance. Two variables measure worker productivity and one measures overtime.

I used four types of analysis: first looking for a statistically significant change in performance; second evaluating the change by comparing the seven supply centers; and linear and multiple regression analysis for each supply center.

How I Computed the Variables

Three variables measure the accuracy and speed of warehouse worker performance. The means of five indicators make up the quality variable for worker performance. The means of six indicators make up the timeliness variable for worker performance. The composite mean for all eleven indicators makes up the Q & T variable. I took work unit and manhour statistics from official documents.

Threats to Internal Validity

Several factors threaten the internal validity of the analysis. Norfolk changed its receipt and issue testing procedures to stop counting previously recorded errors. NAVSUP program managers caught two supply centers fudging results.

Cheating may be widespread. Some data is missing. But, overall, the data used appears adequate.

Changes in procedures, policies, and equipment take place constantly at each supply center. The use of seven supply centers as comparison groups helped in isolating the probable cause of changes.

The turnover rate varies at each supply center and this undoubtedly influences performance. Unfortunately NAVSUP does not keep turnover rates by department or type of work.

Small Sample T Test

I used a standard small sample t test to evaluate the change in performance before and after training for the six non-training variables. The null hypothesis is that the difference between the means before and after training is due to chance. The alternative hypothesis is that a significant change took place. On the basis of a one-tailed test at the .01 level of significance, I would reject the null hypothesis if a t were greater than $t_{.99}$, which for 12+9-2 = 19 degrees of freedom is 2.539. On the basis of a one-tailed test at a .05 level of significance, I would reject the null hypothesis if t were greater than $t_{.95}$, which for 19 degrees of freedom is 1.729.

Overall, there is a significant difference between the means before and after training. I therefore rejected the null hypothesis. I evaluated the total change in performance for each supply center and conducted linear and multiple regression analysis to see if the change is related to training.

Cumulative Change in Performance

I computed the cumulative change in performance since the CBC program began for each supply center and then used linear regression analysis with the change in performance as the dependent variable and training as the independent variable. A minimum of F=6.61 is required for a confidence interval of 95%. F Test results have five degrees of freedom.

The correlation coefficients between CBC training and changes in performance are very low. CBC training did not come close to passing the F test for any performance indicator. The hypothesis requires that the more training, the greater the improvement in performance. The null hypothesis maintains that training and changes in performance are unrelated. This analysis fails to reject the null hypothesis for CBC training.

Physical distribution training shows a much stronger correlation with changes in performance than CBC training does. However, the only relationship with training that is statistically significant at the 95% probability level is receipt and issue productivity. Overall, the combination of CBC and physical distribution training shows a stronger correlation than either of the two alone, especially for timeliness and Q & T.

There is a statistically significant inverse relationship between overtime and quality. Even the corrected coeficient of determination is an impressive .79. The T test result is -4.8 giving it a probability of less than .005 that the result is due to chance.

Overtime appears to be more directly correlated with performance than training. As the amount of overtime declines the quality of performance improves. Inversely, as the amount of overtime increases quality declines. The relationship between overtime and timeliness is mixed.

Linear Regression for Each Supply Center

I conducted linear regression analyses and obtained the following results.

Jacksonville's, Norfolk's and Oakland's F test scores for CBC's correlation with receipt and issue productivity exceed the 5.59 required for a probability of 95%. No other CBC correlation is significant. There may be a causal relationship between changes in productivity and CBC training, that is, the more training, the better the performance for Jacksonville, Norfolk, and Oakland but not for Charleston, Pearl Harbor, Puget Sound, or San Diego.

Puget Sound shows a significant correlation between physical distribution training and quality. Jacksonville shows a significant relationship between physical distribution training and receipt and issue productivity.

Oakland shows a strong correlation with work unit productivity and Puget Sound shows a strong correlation with quality. Overall there was a slight improvement with CBC training removed except at Norfolk which shows a slight decline.

Overtime shows the strongest correlations with performance. Norfolk's overtime is related to timeliness, receipt and issue productivity, and work unit productivity with a probability of error of less than .05%. By far the strongest relationship (a probability of 99%) is at San Diego for timeliness, and Q & T.

General Trends

General trends are mixed. Overtime shows an overall negative correlation with performance but a mixed negative and positive correlation with productivity. Overtime work hours, result in higher production but overtime also increases the overall number of hours required to do a job and results in lower productivity when the relative portion produced during overtime falls below the relative portion produced during regular work time.

Both productivity measures show a mixed negative and positive correlation with all performance indicators and a positive correlation with timeliness indicators. Work unit productivity also shows a positive relationship in this area. CBC and physical distribution training show a weak overall inverse relationship with productivity, quality, and timeliness.

Multiple Regression

I conducted 57 multiple regression analyses for each supply center. Mixed results make interpretation difficult but do suggest some important relationships.

The dependent variables were quality, timeliness, Q & T, and the two measures of productivity. The strongest correlation overall exists between overtime and training as independent variables and performance measures as the dependent variables. Productivity (or the amount of work done) when used as an independent variable, is negatively correlated with quality and positively correlated with timeliness, and Q & T. Training is positively correlated with performance, except at Oakland where the relationship is negative for all training and Puget Sound where the relationship is nagative for CBC training.

Overtime shows a strong inverse relationship to performance indicators. In the case of quality this makes sense. It suggests the more hours people work the more mistakes they make. It is not clear why timeliness and productivity have inverse relationships with overtime. Next to overtime, all training shows the strongest correlation with various dependent variables. Overall, it is more effective than either CBC or physical distribution training alone. CBC's inverse relationships suggest a problem with program administration. In some cases CBC training was poorly planned and executed and adversely affected performance. The time devoted to CBC training may have detracted from performance and productivity.

Training's strongest influence is on productivity. Four CBC, three physical distribution, and six combined training tests exceed the 90% level of significance when used to test the relationship with productivity. One additional CBC, one physical distribution, and one combined test were negative at the same level.

No type of training correlates with timeliness. There is a strong relationship between productivity and timeliness. Overtime is also highly correlated with timeliness. Puget Sound appears to have a strong physical distribution training program. This could be responsible for improvements in performance. Other than Puget Sound, and discounting Oakland's inverse relationship, there is not a significant correlation between training and quality or timeliness of performance for any supply center.

At Charleston productivity is inversely related to quality and overtime is inversely related to work unit productivity. The only significant result fo Jacksonville is between training and productivity. Norfolk shows an inverse relationship between overtime and productivity and a positive relationship between overtime and timeliness. CBC training is also positively related to productivity and Q & T.

Oakland's the relationship between training and performance is generally negative, especially for CBC training. Overtime is positively related to quality. Experienced workers receive overtime and they make fewer mistakes than the overall workforce. Oakland's workload declined slightly over the last five years.

Pearl Harbor's overtime is inversely related to productivity. Workers may be less efficient when working overtime and do not produce as much as they would have in a comparable period during normal working hours. Nothing else is significant for Pearl Harbor.

At Puget Sound CBC training negatively impacted quality and timeliness. Physical distribution training positively impacted quality and productivity. Productivity is correlated with timeliness, and Q & T. The inverse correlation between overtime and timeliness, and Q & T is very strong.

San Diego is the only supply center that does not show some kind of relationship with either measure of productivity as a dependent variable.

Summary

The premise of CBC is that training will improve work quality, timeliness, and productivity. Proper training improves worker skills and attitudes, resulting in a more desirable on-the-job behavior, resulting in improved accuracy, timeliness and productivity. The analysis shows that a statistically significant change in performance occured, but was probably not caused by CBC training. At Puget Sound physical distribution training appears to have contributed to an improvement in performance and productivity.

The analysis presented in this paper does support some trends. Overall:

1. CBC training may not be related to performance;
2. physical distribution training is positively related to timeliness, Q & T, and receipt and issue productivity;
3. overtime is inversely related to quality.

Results differ for each supply center but nowhere does training appear to have significantly impacted warehouse worker performance.

At Jacksonville physical distribution training and CBC may be related to productivity. But, Jacksonville's workload increased while the number of workers remained the same resulting in an increased output per worker.

At Norfolk CBC may be positively related to productivity. Norfolk's workload declined while the number of workers remained stable resulting in an increase in productivity per worker.

At Oakland training is negatively related to all performance and overtime may be positively related to quality. Pearl Harbor and Charleston show no relationships at all with training

[illegible]

[illegible] training is [illegible] dissemination [illegible], or other forms [illegible] Activities and [illegible] practical [illegible] overall goal [illegible] learning objectives. [illegible] tuned to be [illegible] of instruction, and [illegible] intended final outcomes, [illegible] conducted.

[illegible] organization, presenta- [illegible] training satisfactorily [illegible] system is [illegible] Training Command (ISETC), [illegible] (CCR) project based on the [illegible], 1981) was begun in 1981. [illegible] training [illegible] for new technicians. Table 1 [illegible] of the CCR.

Table 1. Cyclic Curriculum Review

Job Analysis

[illegible] begins with a job analysis to compile an initial job/task inventory. [illegible] sources are used to gather task statements or information [illegible] the formulation of task statements. These include curricular [illegible] Navy medical or dental specialists' courses, occupational studies [illegible] analyses and surveys of tasks performed by the Naval [illegible] Personnel Command, reviews of enlisted training submitted by Navy [illegible] experts (SMEs), national technician training guidelines prepared [illegible] societies or professional organizations, and appropriate instructional materials from local, state, or federal agencies. The collected data [illegible] are examined and a job/task inventory is developed by dental [illegible]. The performance needs may differ depending upon the assignment [illegible] technician. Sea or shore duty; large, medium, or small [illegible]; and mission or function of the Navy facility will account for [illegible] performance needs. But formal school training must be [illegible] needs Navy-wide—regardless of where any particular technician [illegible]. Thus the validation process is necessary for further refinement of [illegible] inventory. Tasks may therefore be added by the validation [illegible] validation and clustering changes can be recommended.

[illegible] validation phase is composed of the following six steps:

- [illegible] and locations of SMEs are requested from the Commander, Naval [illegible] Command.
- [illegible] notified of their selection and the CCR purposes and procedures are explained to them.
- [illegible] inventory response form is developed to ascertain whether [illegible] should or should not perform each of the tasks and to provide for the listing of additional tasks, as recommended by the [illegible].
- Responses from SMEs are collected through interviews or mailed response forms.
- Response data are analyzed for unresolved discrepancies and any problems associated with clustering as well as determining which tasks should be added or deleted.
- [illegible] revised job/task inventory is prepared.

Task Training Survey

[illegible] validated job task inventory is the major input for conducting a formal [illegible] training survey. HSETC is provided extensive services by the Navy [illegible] Development and Analysis Center (NODAC). Cooperatively, personnel [illegible] both [illegible] design the task training survey instruments. Specific [illegible] for responding to a mark sense answer booklet are written to accompany the validated job/task inventory. One design is sent to either the entire [illegible] of the Navy medical or dental technicians, or to a sample [illegible] 200 if the specialty population exceeds 200 members. A [illegible] is sent to the [illegible] officers who utilize or direct the [illegible] technicians whose formal school curriculum is being studied.

[illegible] participating in the task training survey supply demographic [illegible] to respond to two questions for each task statement:

- "Do you perform the task?"
- "If yes, how difficult was it to learn this task in relation to the other tasks you mastered?"

The second question is answered by marking a 1 to 8 scale for response [illegible] from extremely low (difficulty) to extremely high (difficulty).

Commissioned officers are asked to respond to three questions in addition to providing data on their medical or dental specialty and the type and location of the facility to which they are presently assigned. Questions for commissioned officers are as follows:

- "Should the technician perform this task?"
- "If yes, how well were technicians prepared to perform this task prior to job entry?" [Responses are made for extremely low (preparation level) to extremely high (preparation level) on a scale ranging from 1 to 8.]
- "At what level is training recommended--familiarization only or hands-on/thorough knowledge?"

For the last question, the following explanation is provided (NODAC, 1982):

Familiarization means information should be provided on basic facts, components, capabilities, etc.

Hands-On/Thorough Knowledge means training should be provided which includes familiarization, but goes further with actual or simulated hands-on practice or in-depth knowledge requiring judgment or application of theory.

Responses from the enlisted personnel provide, by total and by subgroups, the percentage of technicians performing each task and data from which a task difficulty index can be derived. Data from the commissioned officers provide, by total and by subgroups:

- Percentages of responders who believe a task should or should not be performed.
- Information from which a task effectiveness index can be derived.
- The level to which each task should be trained (familiarization only or hands-on/thorough knowledge).

ANALYSIS OF TASK TRAINING SURVEY DATA

Computer printouts on task training survey response data prepared by NODAC are carefully reviewed and summarized by HSETC education specialists. Each task training survey solicits comments and suggested additional tasks from responders. These are also reviewed and summarized. Specific criteria are applied to these summaries to determine whether tasks have a high or low potential for formal school training and to classify selected tasks according to training priority. A task is considered having a high potential for training if 50 percent or more of the commissioned officers indicate that technicians should perform the task or if 50 percent or more of the technicians indicate that they do perform the task. Additionally, if a majority of any subgroup respond positively, the task is included in the high-potential group. Low potential for training is assigned to a task when less than 40 percent of respondents indicate that the task is or should be performed.

A [illegible] approach is taken when priority for training is determined. If [illegible] or more of the total sample of commissioned officers indicate a task should be trained to the hands-on/thorough knowledge level, that task is [illegible] priority training level.

[illegible] effectiveness and task [illegible] indexes based on task [illegible] are constructed and provided to formal schools [illegible]. [illegible] example of a report of training effectiveness and [illegible] tasks selected for training in the Cardiopulmonary [illegible].

[illegible] ABJECTIVES

[illegible], the analysis of response data [illegible] lists:

- [illegible] with a high potential for formal school training.
- [illegible] with a low potential for formal school training.

[illegible] the response data are inconclusive [illegible] tasks should or should not be selected for training, such as [illegible] percent [illegible] both the commissioned officers and [illegible] percent in [illegible], a third list of tasks [illegible]. A fourth list consisting of tasks [illegible] by responders is also prepared. [illegible] lists only, that is, tasks [illegible] for formal school training and [illegible] selected for formal school training, the four lists are sent to the [illegible] experts for final resolution. The [illegible] instructed to:

- Review all lists.
- [illegible] tasks from list three [illegible]) into list one or two.
- [illegible] appropriate tasks from list four into list one.
- [illegible] the level of training (familiarization only or hands-on/ [illegible] knowledge).

[illegible]

[illegible] inventory, job/task [illegible] job/task [illegible] inventory, job/task [illegible], task training survey, task training survey response [illegible], and [illegible] data resolution [illegible] in the following products:

- [illegible] for formal school training list.
- [illegible] selected for formal school training list.
- [illegible] tables.
- [illegible] tables.

[illegible] the specific [illegible] under review receive [illegible] lists and [illegible] on the CDR process. With [illegible] TSEC, program directors and lead instructors at the [illegible] requested to use the task lists to accomplish the following:

- [illegible] learning objectives that adequately support tasks selected [illegible] school training.
- [illegible] learning objectives that support tasks selected for training [illegible] revision.
- [illegible] learning objectives [illegible] not support tasks selected for formal school training.
- [illegible] selected tasks that are not supported by learning objectives.
- [illegible] adequate learning objectives with tasks selected for [illegible] training.

[illegible] for which training is not currently [illegible] of learning objectives that require development

Table 2
Training Effectiveness/Learning Difficulty Of Tasks Recommended For Training Cardiopulmonary Technicians

TRAINING EFFECTIVENESS INDEX

	Expected Mean Ranges	Actual Mean Ranges
1 Highly effective	5.51 8.00	5.53 7.00 (152)*
2 Effective	4.01 5.50	4.82 5.50 (54)
3 Least effective	1.00 4.00	3.67 3.92 (2)

The Training Effectiveness Index shows the relative effectiveness of training [illegible] perceived by commissioned officers responding to the survey. Tasks rated highly effective on the index were seen as more effectively trained in the formal school than were tasks rated less effective. This does not necessarily mean that training for tasks rated "Least effective" was actually ineffective.

LEARNING DIFFICULTY INDEX

	Expected Mean Ranges	Actual Mean Ranges
1 Most difficult	5.51 8.00	5.52 5.93 (4)*
2 Moderately difficult	4.01 5.50	4.03 5.50 (74)
3 Least difficult	1.00 4.00	2.03 4.00 (130)

The Learning Difficulty index shows the relative difficulty of learning the tasks, as perceived by the technicians responding to the survey. As with the effectiveness ratings, these are comparative only.

The rating scale on the survey was as follows

4.00 Below average	5.00 Above average
2.00 Very low	6.00 High
3.00 Low	7.00 Very high
1.00 Below average	8.00 Extremely high

*number of tasks in range is shown in ( )

or revision determine the type of curriculum revision mandated by CCR. Three types of revisions have been identified as follows (CNET, 1981):

- Type A--Changes in course length, objectives, and subject matter to such an extent that logistic support, personnel allocations, funds, and the like are affected. Require project plans and HSETC approval.
- Type B--Modifications within the established structure of the course, including major rescheduling of topics, time, or revision of instructional procedures. Require a Plan of Action and Milestones developed in conjunction with the formal school(s).
- Type C--Minor changes such as correction of clerical errors; insertion of titles or designations of films, publications, and equipment; minor adjustments in time allocations; and addition of learning activities. Require HSETC executive correspondence setting forth minor course changes and tentative completion dates.

REVISION PLANNING DOCUMENT

Cooperatively, the formal schools and HSETC determine the resource constraints that will affect effective and timely completion of the curriculum revision. The type of revision (A, B, or C), the number and experience level of instructors for the program under review, whether the program is offered at a single or multiple training site(s) and the type and number of curriculum support personnel at the training site(s), and the number of programs in the various stages of CCR process help determine resource constraints. This determination affects the level of responsibility assigned to the formal school(s) or assumed by HSETC for the design and development phases. HSETC education specialists and school personnel estimate completion dates for these two phases.

DESIGN PHASE

At the design phase, which is actually a redesign of the curriculum, HSETC and the schools consider task learning difficulty and training effectiveness data resulting from the analysis phase. The course review conducted prior to this phase provides a list of learning objectives to be deleted from the course and a list of learning objectives to be revised. Also provided is a list of selected tasks to be added to the course. After deleting learning objectives not supportive of tasks selected for training, the formal schools, with assistance from HSETC develop newly required objectives. This process includes revising existing learning objectives, where appropriate, or writing new learning objectives for selected tasks not currently provided for in the curriculum. Student assessment procedures are developed for all new or revised learning objectives. Both performance checklists and criterion-referenced test items are created or redesigned for evaluation of students. Finally, the sequence and structure of the entire curriculum are determined. These are tentative decisions on the time required for training students to master the tasks selected for training and the point or place in the training program where each learning objective will occur. Necessary adjustments in sequence and structure may follow a pilot phase of the revised curriculum.

DEVELOP PHASE

By reviewing appropriate previously used training materials and the revised curriculum outline, the formal schools with HSETC assistance:

- Specify learning strategies.

- Calculate resource requirements of time, manpower, and cost to implement revisions, and request resource approval from higher authority if Type A revision is required.
- Review existing materials.
- Develop revised instruction.
- Validate revised instruction.

As a result, program directors and instructors of the training program under review will have revised lesson topic guides and supporting materials available for students. These documents will be used in the pilot phase to validate instruction. A validation report is then submitted by the formal schools and a course approval letter is provided by HSETC. If subsequent additions or changes are deemed necessary, higher authority action on a request for resource allocation is obtained.

IMPLEMENT/CONTROL PHASES

The final two phases are combined to implement the revisions and evaluate the revised curriculum. The formal schools conduct the training of students using any adjustments to the curriculum found necessary during the pilot or validation phases of instruction, while HSETC monitors the outcome. The evaluation or control phase strategy uses a course evaluation plan with course review checklists. All evaluative data are analyzed and necessary changes are identified. If changes are required as a result of the evaluation summary report, the schools submit to HSETC a request to approve minor curriculum changes. At this point, the first cycle of the curriculum review is completed.

SUBSEQUENT CYCLES

Subsequent cycles will occur but will require far less time because of the work accomplished during the first cycle. Specifically, tasks lists requiring only minor revisions will be available. Learning objectives will be task based, and only new objectives necessary to support additional tasks selected for formal school training during subsequent cycles will need to be developed. Further, the ongoing evaluation plan and a minor curriculum changes procedure will keep each curriculum up-to-date. Finally, after the initial cycle, less critical specialty courses may be scheduled for a CCR on alternate cycles rather than every cycle.

SUMMARY

The Cyclical Curriculum Review procedure is based on a systems approach to curriculum development. It uses subject matter as well as process experts to select tasks for training based on state-of-the-art content that will produce highly competent medical and dental technicians. Through the CCR process, the effectiveness of these technicians in meeting the needs of the Navy is ensured.

REFERENCES

1. Gottesman, A.M. (1981, April). "Instructional Systems Development." Journal of Systems Management, 6-9.
2. Navy Occupational Development and Analysis Center (1982, May). Navy Occupational Task Analysis Program. Washington, D.C.: Navy Military Personnel Command.
3. Chief of Naval Education and Training (1981, September). Procedures for Instructional Systems Development. Pensacola, FL: Naval Education and Training Command.

Computer Aversion as a Source of Bias in Computerized Testing

Jo Anna Wood and Gordon F. Pitz
Southern Illinois University, Carbondale

Tests, conventional and computerized, are a pervasive aspect of our lives. They are used as part of placement and selection procedures for employment, training and educational opportunities. In all these instances, test scores are used to classify individuals on the basis of ability or aptitude, as measured by the test. When the testing procedure itself interferes with the measurement of relevant phenomena in a non-random way, that procedure will discriminate against certain test-takers, by incorporating an irrelevant attribute into the classification process.

A growing body of literature (e.g., Lawton & Gerschner, 1982; Naiman, 1982; Nickerson, 1981) suggests that a person's beliefs and attitudes about human-computer interactions will affect that person's ability to interact with a computer. The term computer aversion has been adopted by Meier (1984) to refer to such negative beliefs and attitudes. Computer aversion can hamper one's performance on computerized tasks such as data input, word-processing, or using databases. It seems likely that similar effects would be found when the task is a computerized test.

The research presented in this paper is an attempt to address systematically some shortcomings of published findings on computerized testing and computer aversion, as well to provide some validity data for a measure of computer aversion. The hypotheses to be tested concerned two dependent variables, state anxiety and test performance (number of errors). It was expected that computer aversion would negatively bias the results obtained in computerized testing. Further bias was expected when the testing program was "unfriendly" or difficult to use. In addition, this research was designed to determine whether computer aversion is different from two possibly related concepts, test anxiety and trait anxiety.

Methods

Subjects

Subjects (N=92) were recruited from non-college student adult populations in the Alton and Carbondale, Illinois areas. Subjects were recruited from populations of hospital in the two areas. Only English-speaking subjects were recruited to avoid confounding test results with language abilities.

Tests and other measures

Verbal and math questions were selected from published SAT tests (College Entrance Examination Board, 1983). Two tests were constructed, each composed of two math and three verbal subtests. The tests used a multiple choice format, and were timed. Each test was developed for both paper-and-pencil and computer administration.

Meier (personal communication, January 5, 1985) has developed a questionnaire to measure computer aversion among clinical psychologists (Computer Attitudes Scale). This scale was modified for use with health care personnel for use in the present study. The State-Trait Anxiety Inventory was used to obtain measures indicative of participants' levels of anxiety during testing (state anxiety), as well as to obtain a measure of general (trait) anxiety. The Achievement Anxiety Test (Alpert & Haber, 1960) was designed to measure subjects' perceptions of the extent to which anxiety is either facilitating or debilitating of test performance.

Computerized test administration and scoring were handled by computer programs written in Pascal and developed for this purpose. Computerized tests were administered on a 16-bit IBM-PC compatible machine.

Procedures

Subjects were randomly assigned to either a friendly or unfriendly program condition. Total allotted time was the same for both conditions. The unfriendly program forced subjects to work through each item at a fixed pace and required complex responses. These demands were expected to cause many errors. The friendly program allowed subjects to allocate their own time within each subtest, and used simple response procedures. Subjects were further randomly assigned to receive either the paper-and-pencil or computerized test first. Versions of the abilities tests were randomly assigned as either paper-and-pencil or computerized administration for each subject. Random assignment was handled through a computer program, with the only restriction being an equal number of subjects in all conditions.

Results

Analysis of Covariance (ANCOVA) was used to test the effects of program friendliness and test mode (i.e., computerized vs paper and pencil) on performance and state anxiety while controlling for the effects of computer aversion. An additional factor, order, was included to determine if the dependent measures were influenced by the order in which the test modes were used.

For verbal performance the significant effects of interest are interactions of Program (friendly vs. unfriendly) by Mode (computer vs. paper) (F (1,84) = 8.69, p=.004), and Order (computer first vs paper first) by Computer Aversion by Mode (F (1,84) = 5.65, p=.020). The former effect was also significant in the analysis of overall performance. The latter was not; overall performance included scores on math subtests that probably represented mostly random error.

The two-way interaction of program by mode is shown in Table 1. Both forms of the computer programs induced poorer performance than did the paper and pencil tests, but the effect was greater for the "unfriendly" version.

The significant three-way interaction indicates that one or more of the cells in the design matrix differed from others in terms of the relationship between computer aversion and performance. Separate regressions of performance on computer aversion were calculated for each cell. As shown in Table 2, computer aversion was significantly and positively related to performance on computerized tests only when those tests preceded the paper and pencil tests. Computer aversion was not related to performance on paper tests.

computerized test. 82% of the low computer aversive, but only 25% of the high computer aversives, had scores at or above the grand median.

The last comparison is most interesting because the "friendly" computer test most closely approximates computerizd testing procedures that would be used in selection and classroom testing applications. Although it has less overall effect on performance, this test appeared to be more biased than the "unfriendly" version. Insofar as this experiment represents computerized testing for selection purposes, it appears that computerized tests have a built-in bias against computer aversives.

If the results of the present study are indicative of what occurs in other computerized testing applications, then one of two things may occur. First, Computer Aversion may be related both to computerized test performance and to performance on some criterion measure (e.g., job performance, success in college). In this case, using computerized tests should result in more accurate predictions of the criterion. In the second scenario, Computer Aversion is related to computerized test performance, but not to performance on the criterion measure. If this is true, then Computer Aversion acts as a moderator variable; that is, it affects the relationship between computerized test performance and performance on the criterion, and does so differently for various groups of subjects.

References

Alpert, R. & Haber, R. (1960). Anxiety in academic achievement situations. Journal of Abnormal and Social Psychology, 61, 207-215.

College Entrance Examination Board (1983). 10 SATs Scholastic Aptitude Tests of the College Board, New York:CEEB.

Lawton, J. & Gerschner, V.T. (1982). A review of the literature on attitudes towards computers and computerized instruction. Journal of Research and Development in Education, 16, 50-55.

Meier, S.T. (1984). Computer Aversion. Unpublished manuscript, Missouri Institute of Psychiatry, St. Louis. MO.

Naiman, A. (1982). Women technophobia and computers. Classroom Computer News, 2, 23-24.

Nickerson, R. (1981). Why interactive computer systems are sometimes not used by people who might benefit from them. International Journal of Man-Machine Studies, 15. 469-483.

Table 1

Cell means and Standard Deviations for Math and Verbal Errors

		Friendly Program		Unfriendly Program	
COMPUTER TEST		Mean	S.D.	Mean	S.D.
Computer First	Math	14.48	(2.89)	16.74	(2.30)
	Verbal	14.43	(5.04)	16.26	(6.80)
Paper First	Math	15.13	(3.68)	16.26	(2.65)
	Verbal	12.87	(4.59)	19.22	(5.48)
PAPER TEST					
Computer First	Math	13.17	(3.76)	13.78	(4.06)
	Verbal	12.43	(5.67)	12.04	(6.67)
Paper First [1]	Math	14.48	(4.24)	14.48	(4.24)
	Verbal	13.83	(6.40)	13.83	(7.69)

[1] Cell means for these two conditions were calculated as a single group, since the two conditions were equivalent.

Table 2

Regression Coefficients for Verbal Errors on Computer Aversion

	Friendly		Unfriendly	
	b	beta	b	beta
COMPUTER TEST				
Computer First	.31	.49*	.52	.54*
Paper First	.08	.09	.08	.15
PAPER TEST				
Computer First	.20	.29	.17	.18
Paper First[1]	.17	.21	.17	.21

NOTES: (1) Starred items are significant at p<.01.
(2) b's are raw regression coefficients, while Beta's are standardized regression coefficients.

[1] These two groups were treated as single cell for regression

computerized test. 82% of the low computer aversive, but only 25% of the high computer aversives, had scored at or above the grand median.

The last comparison is most interesting because the "friendly" computer test most closely approximates computerizd testing procedures that would be used in selection and classroom testing applications. Although it has less overall effect on performance, this test appeared to be more biased than the "unfriendly" version. Insofar as this experiment represents computerized testing for selection purposes, it appears that computerized tests have a built-in bias against computer aversives.

If the results of the present study are indicative of what occurs in other computerized testing applications, then one of two things may occur. First, Computer Aversion may be related both to computerized test performance and to performance on some criterion measure (e.g., job performance, success in college). In this case, using computerizeo tests should result in more accurate predictions of the criterion. In the second scenario, Computer Aversion is related to computerized test performance, but not to performance on the criterion measure. If this is true, then Computer Aversion acts as a moderator variable; that is, it affects the relationship between computerized test performance and performance on the criterion, and does so differently for various groups of subjects.

References

Alpert, R. & Haber, R. (1960). Anxiety in academic achievement situations. Journal of Abnormal and Social Psychology, 61, 207-215.

College Entrance Examination Board (1983). 10 SATs Scholastic Aptitude Tests of the College Board, New York:CEEB.

Lawton, J. & Gerschner, V.T. (1982). A review of the literature on attitudes towards computers and computerized instruction. Journal of Research and Development in Education, 16, 50-55.

Meier, S.T. (1984). Computer Aversion. Unpublished manuscript, Missouri Institute of Psychiatry, St. Louis. MO.

Naiman, A. (1982). Women technophobia and computers. Classroom Computer News, 2, 23-24.

Nickerson, R. (1981). Why interactive computer systems are sometimes not used by people who might benefit from them. International Journal of Man-Machine Studies, 15. 469-483.

FIGURE 1: PAPER TEST SCORES
□ = Hi averse; △ = Low averse

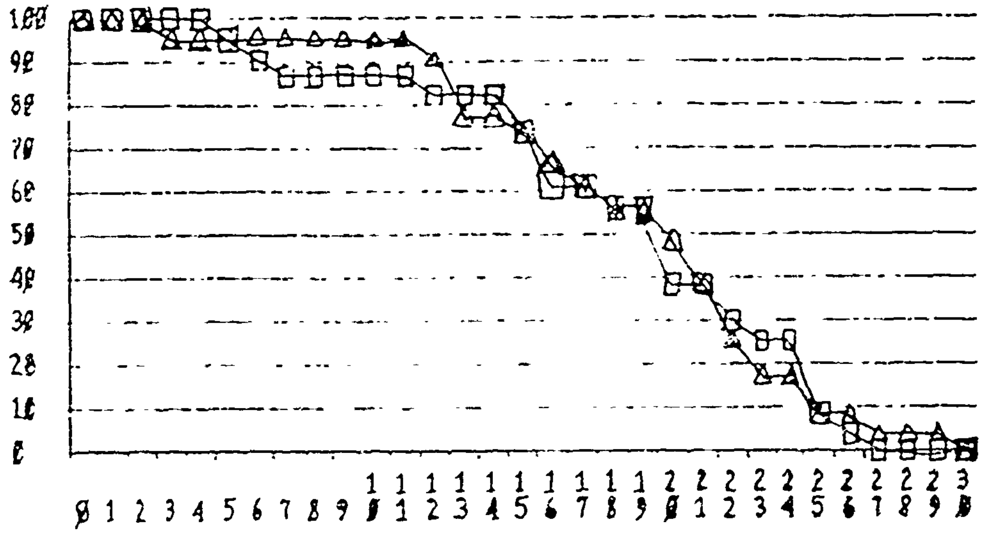


FIGURE 2: UNFRIENDLY COMP. TEST SCORES
□ = Hi averse; △ = Low averse

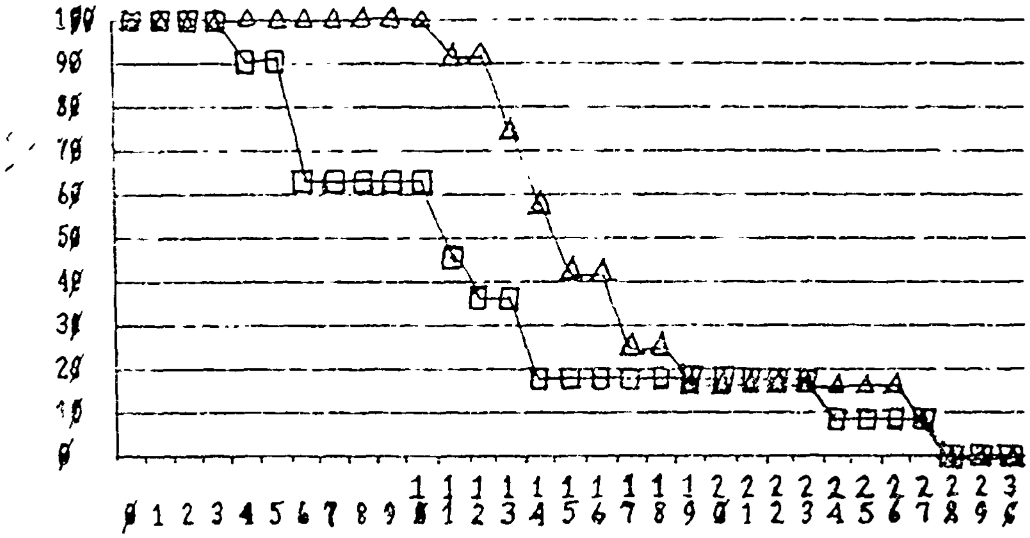
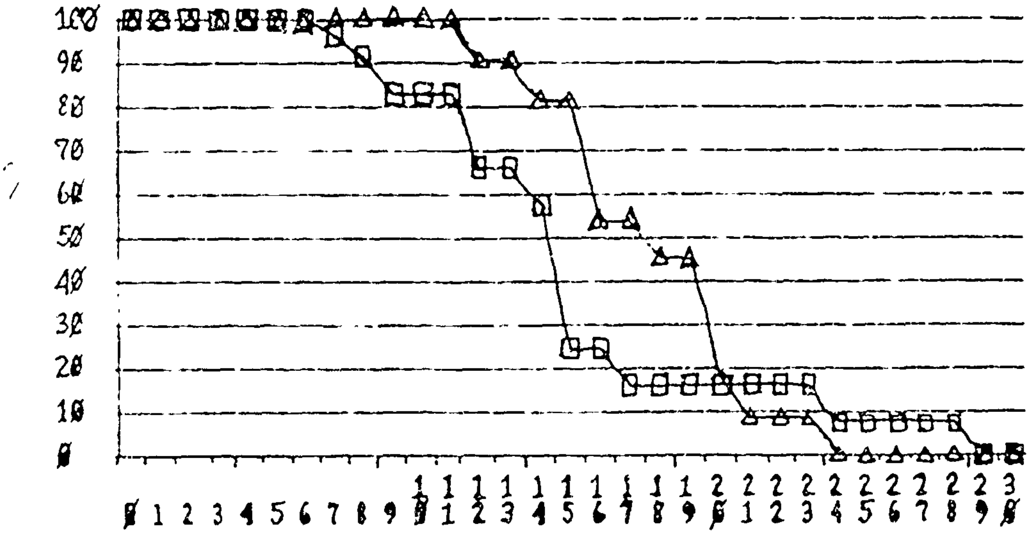


FIGURE 3: FRIENDLY COMP. TEST SCORES
□ = Hi averse; △ = Low averse



Orientation, surrogate travel, and gender differences in videogame strategy

Sharon Tkacz
Army Research Institute

Research has shown that individuals vary in their ability to process and use spatial information (Kosslyn, Brunn, Cave, & Wallach, 1983) They also differ in the frame of reference they use to form memory representations of space (Sholl & Egeth, 1980) Some use an egocentric system (e.g., right or left) whereas others use a topographic system (e.g., north or south). Developmental studies (Fick & Reiser, 1982) have suggested that individuals adopt the topographic, more sophisticated system as they mature.

Piaget has suggested that physical movement through the environment is how spatial reasoning skills are acquired. Goldin and Thorndyke (1981) support Piaget's arguments, demonstrating that navigation through space provides a unique kind of spatial knowledge, procedural knowledge, that cannot be acquired simply by reading maps.

The fact that movement leads to procedural knowledge acquisition has been applied to navigation training. Cohen (1980) has shown that the information derived from actual travel can be approximated by surrogate travel. In some cases, simulated movement may be even more effective as a training aid than actual navigation: if only relevant information is presented, irrelevant information cannot be distracting.

Much research (Wittig & Petersen, 1979) has demonstrated gender differences in spatial information processing, including the relationship of cognitive variables to sex-role identity. Since previous research has shown that individuals tend to conform to sex-role expectations, and expertise in computers and videogames is considered masculine, it is reasonable to predict that females may not perform as well as males. Whether lower female performance is due to a lack of cognitive ability or adherence to sex-role expectations has not been established.

The research described below investigated navigation through an artificial environment created by a microcomputer. The game required players to use a topographic reference system to indicate directions. In addition to dependent measures derived from the videogame, cognitive components assumed to underly game performance were assessed Cognitive components were also examined to see if females and males differ in the basic cognitive skills required by the game

Method

One hundred and ninety undergraduates served as participants They were administered several psychometric tests, and played a series of eight videogames requiring them to escape from a 5 x 5 x 5 cubic maze. Players moved from one room to the next through openings in the floors, walls, and ceilings by typing "n", "s", "e", "w", "u", or "d" for north, south, east, west, up, or down directions Two types of information available were current Position (P) and location of the escape room or Goal (G), each defined by x, y, z coordinates Combinations of these two types of information formed the four INFORMATION CONDITIONS: PG, P, G, or Q

In the PG condition, information on both the player's Position (P) and the Goal (G) was provided continuously on the screen. Subjects in the P or G condition had only one kind of information available. Those in the Q condition had no information displayed but were permitted to request both P and G coordinates. All participants played four PG games, in order to familiarize them with the game and the keyboard, and then four more games, in one of the four INFORMATION CONDITIONS.

Results and Discussion

Males showed superior spatial performance. (Table 1 shows means for all psychometric tests.) There were no gender differences on vocabulary or reasoning tests. These data indicate that males and females differ significantly in the skills they bring to the experiment that were expected to underly game performance.

Table 1. Gender differences in psychometric measures.

Variable	Female Mean	Male Mean	t	p
Abstract Orientation	101	113	3.4	.001
Map Orientation	8.5	10.2	2.7	.01
Figural Reasoning	22.6	22.4	.2	*
Mental Rotation	25.6	27.7	2.2	.05
Vocabulary	55.3	55.2	0	*

Game performance was described by ten dependent variables derived from individual key presses for each game. SCORE indicates the total time to complete one game. RESPONSE TIME indicates mean time between any two keypresses. Similarly, STATIONARY TIME indicates mean time spent in a room. EFFICIENCY is a ratio of the minimum distance to actual distance between starting Position and Goal room. REORIENTATION is the rate (the number of times per minute) that players changed the direction they were facing SURFACE RATE is a measure of time spent in surface rooms of the cube. Similarly, INTERIOR RATE is a measure of time spent in interior rooms. VISIBLE CRASH indicates how many times a player tried to go through a wall (visible on the screen) that did not have a door. Similarly, REAR CRASH indicates how many times a player tried to go through the wall behind them (not visible on the screen). Lastly, ERROR KEY is a measure of illegitimate key presses

An analysis of variance was performed on the ten dependent measures with two between-subjects variables (INFORMATION CONDITION, C and GENDER, G) and one within-subjects variable (PRACTICE, P) SCORE and REORIENTATION were the only dependent measures for which any GENDER (G) effect was obtained (see Table

Table 2. Analyses of variance on ten videogame measures.

Dependent measure	Source of variation	*F*	*p*
Score	C	76.3	.001
	G	11.6	.001
	P	15.3	.001
	CP	3.0	.001
	GP	2.9	.05
Response time	P	55.3	.001
Stationary time	C	18.1	.001
	P	30.0	.001
	CP	3.7	.001
Efficiency	C	64.8	.001
	P	2.9	.05
Reorientation	G	7.1	.01
Surface rate	C	6.7	.001
	P	40.9	.001
	CP	2.7	.01
Interior rate	C	27.2	.001
	CP	1.9	.05
Visible crash	C	11.6	.001
	P	9.6	.001
	G	4.1	.05
Rear crash	C	51.3	.001
	P	8.5	.001
	CP	5.1	.001
Error key	C	9.2	.001
	P	3.8	.01
	CP	2.5	.01

2). Although almost all measures improved with PRACTICE (P), the GENDER x PRACTICE interaction (GP) was significant only for SCORE, suggesting that, while their initial scores may be lower, females may show greater improvement. The effect of INFORMATION CONDITION (C) was significant for all variables except RESPONSE TIME and REORIENTATION, indicating that the rate of key pressing and the rate of turning is independent of the amount and type of information available. Finally, a PRACTICE x INFORMATION CONDITION interaction (CP) was obtained for several measures -- SCORE, STATIONARY TIME, SURFACE RATE, INTERIOR RATE, REAR CRASH, and ERROR KEY. The fact that this interaction was not obtained for all dependent variables indicates that improvement in performance is not simply a function of repeated practice. These game-derived performance measures may be indices of individual differences in information-processing capacity, and not subject to practice effects.

Means for these dependent measures are shown in Table 3. Here, SCORE and REAR CRASH were the only variables for which any gender difference was obtained, in only two INFORMATION CONDITIONS (G and Q). These results suggest performance is very similar for males and females.

Table 3. Results of t-tests on videogame measures.

Variable	INFORMATION CONDITION	Female Mean	Male Mean	t	p
Score	PG	82	61	1.197	*
	Q	142	99	1.656	*
	G	414	247	3.073	.01
	P	639	507	1.493	*
Response time	PG	4.4	4.0	.608	*
	Q	3.8	3.7	.182	*
	G	3.5	3.8	.651	*
	P	3.7	3.3	.987	*
Stationary time	PG	5.5	5.1	.457	*
	Q	9.8	8.8	.618	*
	G	4.7	4.8	.170	*
	P	5.2	4.6	1.075	*
Efficiency	PG	.60	.59	.212	*
	Q	.56	.63	1.246	*
	G	.21	.21	0	*
	P	.11	.11	0	*
Reorientation	PG	.07	1.16	1.726	*
	Q	.09	.47	1.212	*
	G	.10	.31	1.243	*
	P	.09	.20	.808	*
Surface rate	PG	8.28	8.52	.213	*
	Q	5.95	5.42	.581	*
	G	6.45	6.90	.623	*
	P	7.07	8.83	1.872	*
Interior rate	PG	3.59	3.68	.127	*
	Q	1.65	2.26	1.954	*
	G	1.33	1.72	1.047	*
	P	1.01	1.05	.119	*
Visible crash	PG	1.72	1.49	.572	*
	Q	.85	.88	.103	*
	G	3.84	2.38	1.692	*
	P	3.02	2.10	1.530	*
Rear crash	PG	.69	.70	.037	*
	Q	.48	.21	2.808	.01
	G	1.36	1.28	.293	*
	P	2.38	2.47	.224	*
Error key	PG	.11	.20	.976	*
	Q	.40	.52	.661	*
	G	.08	.12	.679	*
	P	.13	.22	.584	*

Data in Table 2 also indicate that performance varies across INFORMATION CONDITION (C) for every variable except RESPONSE TIME, suggesting that different strategies are employed in different INFORMATION CONDITIONS. Rather than indicating different difficulty levels of the same game, INFORMATION CONDITIONS may be qualitatively different games, from a problem-solving perspective. That is, a task that involves finding a goal without knowledge of your own position may not have much in common with a situation where your position is known.

This interpretation is supported by results of stepwise multiple regression analyses. SCORE was predicted from the psychometric measures, shown in Table 1, for males and females separately, and for males and females combined, for each INFORMATION CONDITION. Table 4 shows amount of variance in SCORE accounted for by the best combination of two psychometric predictors. The different INFORMATION CONDITIONS have different psychometric predictors, suggesting differences in cognitive components.

In contrast to the absence of gender differences in Table 3, data in Table 4 indicate that components of performance differ for females and males. This difference is particulary clear for condition Q. Cognitive correlates of female performance are vocabulary and reasoning, neither of which are spatial measures. Conversely, the best predictors of male performance are mental rotation and abstract orientation. Taken together with the data in Table 1, these results suggest that individuals may develop strategies that depend on their own skills, rather than strategies that are task dependent.

Table 4. Multiple regression analyses: predicting SCORE from psychometric measures.

INFORMATION CONDITION		Predictors	R-SQUARED
PG	females	reasoning, vocabulary	.453
	males	reasoning, abstract orientation	.556
	both	reasoning, vocabulary	.360
P	females	map & abstract orientation	.174
	males	map & abstract orientation	.320
	both	map & abstract orientation	.211
G	females	map & abstract orientation	.254
	males	map & abstract orientation	.177
	both	map orientation, reasoning	.186
Q	females	reasoning, vocabulary	.206
	males	abstract orientation, mental rotation	.409
	both	map orientation, mental rotation	.177

Summary & Conclusions

Differences in INFORMATION CONDITIONS demonstrate that this variable represents different task requirements. Thus, the dependent variables selected to describe game performance seem to do so adequately, since they reflect different aspects of the players' performance and strategy.

Regression analyses indicate that cognitive components underlying game performance are not the same for males and females Although components underlying videogame performance differ, suggesting gender-related strategy differences, actual game performance shows little variation attributable to gender.

In sum, given that videogame skills are well retained, fun and relatively easy to acquire, they have much potential as instructional tools. For example, games simulating navigation could provide a simple, cost-effective way of training spatial learning strategies and exercising navigational skills. Since individuals that have different cognitive skills demonstrated similar game performance, different strategies may be employed to achieve the same results. Future instructional paradigms should provide for flexibility in strategy development so that learners may make the best of their individual cognitive strengths. Further, the dependent measures employed here reflect complex, strategic behavior in a simulated environment. Measures such as these may have greater ecological validity than standard psychometric tests in predicting individual differences in complex, real world behavior.

References

Cohen, M. E. (1980). The effects of environmental interactions on the structure and process of cognitive mapping. Ph.D. Thesis, Temple University, Philadelphia.

Goldin, S. & Thorndyke, P. (1981). Spatial learning and reasoning skill. Santa Monica, CA: The Rand Corporation, R-2035-ARMY.

Jones, M. (1984). Videogames as psychological tests. Simulation and Gaming, 15(2), 131-157.

Kosslyn, S. M., Brunn, J. L., Cave, C. R., & Wallach, R. W. (1983) Components of mental representation. (Contract No. N00014-79-C-0982) Arlington, Virginia: Office of Naval Research.

Pick, H L., & Rieser, J. J. (1982) Children's cognitive mapping. In M. Potegal (Ed.), Spatial abilities Development and physiological foundations (pp. 107-128) New York: Academic Press.

Sholl, M.J., & Egeth, H.E (1980). Interpreting direction from graphic displays: Spatial frames of reference. In P A. Kolers, M. E. Wrolstad, & H. Bouma (Eds.), Processing of Visible Language (Vol 2) New York. Plenum Press

Wittig M A., & Petersen, A. C (Eds.). (1979) Sex related differences in cognitive functioning. New York Academic Press

The Navy's General Unrestricted Line Community: Career Management and Career Development Problems

Gerry Wilcove

Navy Personnel Research and Development Center
San Diego, CA 92152

Community Description and Background

In 1981, General Unrestricted Line Officers (i.e., General URLs) were given community status, and a community manager was selected. Given its short history, it is not surprising that the community is a relatively unknown entity to people both within and outside the Navy. The community is composed of approximately 3,000 officers, 80-percent of whom are women. A disproportionate number of officers are lieutenant or below, although it is expected that the number of officers selected for executive officer, the second in command, will triple from 1984 to 1987. The function of this community is to support the Navy's fighting forces by serving in shore billets as general managers and specialists in areas such as personnel management, financial administration, data processing and computer technology, organizational effectiveness, and communications. There are also limited numbers of General URLs in operations systems technology, naval systems engineering, political-military strategic planning, and weapons engineering. Women are eligible to receive some training on combat vessels and may even become surface warfare officers. However, duty on combat vessels is restricted by law to temporary duty under noncombat conditions.

Career Management and Career Development Problems

I will first discuss some of the problems that have made it difficult for the General URL community to manage the careers of its officers and for individual officers to develop their careers. I will then describe the policy and procedural changes that occurred in 1984 that ameliorated some of the problems. The problems that are discusssed were identified from. (1) 1982 questionnaire results from approximately 45-percent of the community, (2) open-ended comments from the same questionnaires (N=500), (3) 50 interviews with community members, and (4) conversations with Washington managers and policy-makers. The problems are not listed in any kind of rank order.

First, in the policy-making area, the community has been handicapped by the lack of freedom to control the yearly number of accessions. Instead, these numbers have been determined for the community in accordance with a "surge-tank" concept, i.e., in accordance with estimates of the number of billets, or assignments, that would be vacant X number of years in the future because of a lack of warfare specialists. Managing careers becomes difficult when officers from different commissioning years are subjected to different selection ratios at the same points in their careers. By the same token, a woman developing her career faces minimal or intense competition depending on her year of commissioning

A second problem for General URLs has been the fact that they have been reassigned upon the completion of a tour by officers from another community; i.e., Surface Warfare Officers (SWOs). Thus, there has been less freedom to groom those of demonstrated potential for high responsibility senior jobs. Individual officers have complained that SWOs do not know the duties and requirements of existing assignments or their career potential.

A third career management and development problem has been the lack of high level acceptance concerning the General URL's right to career enhancing billets. This obstacle has been manifested in two ways: (1) the tendency of the Navy to assign warfare specialists, instead of General URLs, to career-enhancing shore billets, when a direct competition occurs, and (2) the tendency to reserve certain billets for warfare specialists, even though General URLs may be capable of performing the work. Here, career managers are stymied in their attempts to develop officers for major shore commands because of restricted opportunities at lower levels. Conversely, officers attempting to develop their careers in particular directions have been forced to reformulate their career goals

A fourth problem has been a complaint about the quality of the billets available for General URLs. A LT perhaps best summarizes the feelings of dissatisfied General URLs when she states.

> "I feel that General URL jobs are overall the least desirable within the Navy. These jobs are poorly defined, usually not operational or competitive, and require no special education or background. Many are nonessential and have no clear career path associated with them. They can be filled by anything from a CWO2 to a LCDR and are often gapped for long periods of time" (i.e., left vacant).

A fifth problem is that the career path for General URLs has been an ambiguous one defined mainly in general terms; i.e., an individual should be sure to perform well in subspecialty and leadership positions. It has been difficult for career managers (i.e., the community manager, senior officers, assignment managers, and subspecialty managers) to offer advice when the "wickets" for career advancement were largely unknown, the criteria for promotion in competition with warfare officers were unformulated, and the career advancement potential of various subspecialties was a matter of conjecture.

This career advice problem showed up in survey results. For example, only 45-percent reported that they had been counseled on the "tickets" that have to be "punched" for career advancement, and only 25-percent said they had been counseled on the "blind alleys" that might destroy their careers.

The sixth problem concerns senior officers. The Navy has thus far concentrated on defining the career path for junior officers. Individuals who have completed a commander-commanding officer tour complain about the lack of a career pattern after that point and the lack of opportunities in the upper levels of the Navy hierarchy The latter includes policy-making positions in the "fifth wing of the Pentagon", senior administrative jobs at headquarter and staff commands (e.g., SURFPAC, RECRUITCOM, the Sixth Fleet), and commanding officer and executive officer billets at naval training commands and administrative commands. At the present time, there is no overall Pentagon group that formulates policy for the General URL community as there are for the other unrestricted line communities (SWOs, aviators, and submariners).

A seventh problem concerns training opportunities and pipelines, neither of which have been institutionalized. Survey results showed that only 25-percent had received training enroute to their new assignments. In addition, there is no department head school for General URLs, and General URLs do not attend the Prospective Executive Officer or Prospective Commanding Officer School unless they have been selected for major command.

The eighth problem is the subspecialty system that is difficult to manage and difficult to learn and influence by the individual officer. The system itself is composed of many administrative elements including a placement division, a division responsible for conducting a zero-based review of the entire subspecialty billet inventory, manpower claimants, sponsors, designator advisors, and assignment managers. Ninety-two percent of the survey sample viewed subspecialties as important for their careers Yet, there is confusion among officers on the administrative steps they need to take to obtain a subspecialty; also, when to obtain one. For example, 65-percent indicated that it was important to obtain a subspecialty early in their careers. And yet, a Navy policy makes subspecialty experience obsolete after five years.

The ninth problem centers on dual career couples. Approximately 75-percent of the married General URLs are married to military men. These couples want to colocate. This desire provides: (1) career management problems for the Navy which must be concerned with mission requirements and (2) career development problems for General URLs who may want to advance in their careers rather than simply "be employed." The Navy currently has a policy that every reasonable attempt must be made to colocate couples.

The tenth problem is philosophical; i.e., there are disagreements about what direction the evolution of career management systems should take. For example, some community members want billets that are reserved exclusively for General URLs or want separate assignment managers. In this way, it is argued that women will be able to compete more effectively with their male counterparts. On the other hand, complaints are voiced that such approaches are an example of a "separate but equal" status that removes women from the Navy's mainstream and prevents them from competing successfully with male warfare specialists.

Policy Initiatives in the Career Area

In November 1984, the Navy promulgated a series of policies designed to alleviate some of the career management and development problems faced by the General URL community. One of the changes was that General URLs, rather than SWOs, would serve as assignment managers, i.e., General URLs would assign their own community The exceptions to this policy are commanders who have screened for command and CAPTS. These two sets of individuals will still be assigned by SWOs. The "new accessions" desk will also be manned by a SWO. In brief, the policies seem to contain elements of both philosophical positions mentioned earlier.

A second change was the institution of a two-career track, the Leadership/Subspecialty Track, which is the existing one, and a new Specialty Track. The former emphasizes both leadership and subspecialty billets, culminating in commanding officer positions. The new track emphasizes the opportunity to stay in a subspecialty track, eventually becoming a program manager rather than a commanding officer. However, a small percentage of those in the second track will become commanding officers of shore installations specializing in activities such as computer operations. Approximately one-third of General URL LCDRs will be accepted into the Specialty Track.

It is hoped that the two-career track better defines the billets that are needed to advance in the Navy, that it gives General URLs more options to fulfill their career goals, and that it provides the Navy with the specialized skills it needs to meet its requirements.

A third policy was aimed at stabilizing accessions into the community. While the numbers established operate within broad parameters, there is at least some structure and community control over this important issue. The previous community manager characterized this policy as the single most important change within the community.

A fourth policy was aimed at freeing up additional numbers of challenging and career enhancing billets for the General URL community. That is, 1,800 billets were identified that were reserved for warfare specialists, but which seemed within the capabilities of many General URLS. The goal, which was reached, was to be able to reclassify 300 of these billets so that General URLs would be eligible for them.

Before reclassification, General URLs were blocked, to a large extent, from obtaining billets that represented the operational opportunities needed for career advancement. The policy change addressed the complaint, previously quoted, regarding the poor quality of billets available to General URLs.

A fifth policy clarified the definition of leadership positions below the level of executive/commanding officer. That is, criteria were established for division officer and department head billets, and billets were appropriately recoded. This policy further defines the career path, thereby helping the individual officer to formulate and plot career strategy.

A sixth policy further defined the career path for officers just entering the Navy. The problem addressed by the policy was that new accessions were obtaining limited subspecialty and leadership experience in those situations where the Navy's needs were defined as being preeminent. The new policy stated that, under such circumstances, the first assignment will be split-toured or be a 2-year rather than a 3-year obligation

A seventh policy was designed to ensure that General URLs, as they compete with warfaie specialists, receive their fair share of leadership positions at the lieutenant-commander (LCDR) level. Although specific billets were not reserved, the policy stated that 75-percent of the LCDR executive officer and commanding officer shore billets will be reserved for the General URL community. The policy also dictated that the same type of arrangement be implemented as the community matures and sufficient numbers of commanders are available.

Finally, the recommendation was made to increase the fields in which General URL officers are assigned so that they have the breadth of experience to fill a wider range of leadership, HQ (i.e., headquarter) and subspecialty billets at the 06 (i.e., CAPT) level. A lead and an assistant agency were designated to determine how best to implement this recommendation.

Questionnaires will be mailed shortly and interviews conducted to determine the community's reactions to the actual and recommended changes discussed in this paper.

Attitudes, Preferences and Career Intentions of ROTC and Non-ROTC Students

Melvin J. Kimmel
US Army Research Institute for the Behavioral and Social Sciences[1]

The Reserve Officer Training Corps (ROTC), the Army's main source of officer personnel, finds itself in a dilemma, for it is being tasked to recruit and retain record numbers of officer candidates in a shrinking and more competitive market (Hertzbach et al 1985). The White male college-bound population has been its main source of cadets. However, changing market conditions will necessitate going beyond this traditional pool to include women and the growing number of adolescent Blacks and Hispanics (McNeil, 1983). To accomplish its expanded mission, ROTC must have information on the backgrounds and attitudes of these people in order to develop effective recruiting and training programs.

Over the past fifteen years, the ROTC Advertising and Media Division has relied upon comparative studies of ROTC cadet and noncadet college students to develop its programs (e.g., Armstrong, Farrell and Card, 1979; Card et al, 1975; Hicks et al, 1979; and Montgomery et al, 1974). These research efforts have generally drawn similar conclusions: (1) the influence sources that ROTC cadets and noncadets use to make career decisions are similar; (2) the ROTC and military-related attitudes of noncadets have become more positive since the Vietnam era; and (3) cadets and noncadets differ markedly on values, ROTC and military-related attitudes, preferences, and intentions, although some of these differences are true for only certain ethnic and gender subgroups.

The present effort continues this line of research with a more recent sample of ROTC and non-ROTC students. Its focus is on variou^c socialization variables and military-related attitudes that may impact on one's decision to join ROTC and pursue a military career.

METHOD

Subjects. Usable data were gathered from 898 college students from 11 campuses with ROTC programs. The sample was composed of 427 first and second year ROTC cadets and 471 noncadet students. The frequencies for the ethnic and sex subgroups in ROTC and non-ROTC are presented below:

	White		Black		Hispanic	
	Male	Female	Male	Female	Male	Female
Cadet	211	71	51	39	30	15
Noncadet	130	123	60	63	60	35

The majority of the sample (54%) were enrolled in southern colleges; 29% came from schools in the northeastern and mid-Atlantic regions of the country; 11% were enrolled in midwestern colleges; and 6% were from western schools. These percentages accurately reflect the geographic distribution of males and females in our sample, but they are not as characteristic of the cadet-noncadet breakdown or the geographic distribution of the different ethnic groups. The greatest discrepancies are found in the eastern and

[1] The views expressed in this paper are those of the author and do not necessarily reflect the view of the US Army Research Institute or the Department of the Army.

southern college subsamples. Sixty-eight percent of the non-ROTC participants were from southern schools as compared to only 39% of the ROTC sample, while 18% of the non-ROTC sample and 41% of the cadets were enrolled in eastern colleges. An overwhelming majority of the Hispanics (86%) and Blacks (85%) in our sample came from southern colleges, while the White sample was more equally divided, with 34% from southern colleges and 42% from eastern schools.

Instrument and procedure. University staff members administered a slightly modified version of the 232-item "Career Attitude Survey" (Armstrong et al, 1979) during regularly scheduled class periods. The surveys were administered to ROTC cadets in MSI and MSII classes (The ROTC Basic Course) and to non-ROTC students enrolled in lower level required courses (e.g., English 101). The survey was composed of items on background characteristics, media preferences, education and career-related variables, and ROTC/Army knowledge and attitudes. The questionnaires took approximately 45 minutes to complete. All answer sheets were returned to a central location for coding, keypunching, 100% verification, and analysis. A 2 (Cadet-Noncadet) x 2 (Male-Female) x 3 (White-Black-Hispanic) factorial design was used to analyze main effects and interactions. The F-statistic was used for items associated with rating scales; and tne x^2 statistic to analyze categorical data.

RESULTS

Career and Education Influences. A number of variables may influence one's decision to pursue a military career. Among these are the military attitudes and experiences of family and friends. When asked to rate on 5-point scales how favorably their parents and friends perceived the status of an Army officer career, the mean ratings of ROTC cadets were significantly higher than the non-ROTC student ratings on both perceived parental attitudes ($\bar{x}$=3.92 vs $\bar{x}$=3.41, F=35.76, p <.001) and the perceived attitude of their friends (x=3.30 vs $\bar{x}$=3.06, F=13.92, p <01). Sex and ethnic differences were found in respondents' perception of their friends' attitude, but not with respect to their parents. Cadets and noncadet females perceived their friends' attitude as more positive than did the males ($\bar{x}$=3.42 vs $\bar{x}$=3.10, F=7.63, p <.01), and the Blacks' ratings ($\bar{x}$=3.88)were significantly higher than Whites ($\bar{x}$=3.18) or Hispanics ($\bar{x}$=3.23, F=10.94, p <.001).

When asked whether or not their parents, siblings, and friends had been in ROTC or the military, a significantly higher percentage of cadets than noncadets reported friends with ROTC experience (57% vs 50%, x^2=5.50, p <.05) and parents with military experience (69% vs 50%, x^2=19.25, p <.01). Ethnic differences also were found on ROTC and military experience variables. Fewer Hispanics (10%) than Whites (20%) or Blacks (14%) reported parents with ROTC experience (x^2=6.73, p <.05), while a greater percentage of Blacks reported siblings with ROTC experience (23% Blacks vs 13% each for Whites and Hispanics, x^2=11.97, p <.01). Blacks also had the highest percentages reporting military experience for siblings (28% Blacks vs 19% Whites and 15% Hispanics, x^2=11.97, p <.01) and friends (78% vs 69% and 68%, respectively, x^2=9.08, p <.05). Whites reported the highest percentage of parents with military experience (71% Whites vs 44% Hispanics and 38% Blacks, x^2=72.58 p <.001). The only significant sex difference that held for both cadets and noncadets was for friends with military experience, where the percentage of females was higher than males (73% vs 63%, x^2=3.86, p <.05). A higher percentage of males than females reported parents with military experience, but the difference was more marked for cadets (72% vs 45%) than noncadets (58% vs 52% x^2=6.79, p <.01).

The students were also asked directly about the sources that influenced their decision on whether or not to participate in college ROTC. ROTC influence sources were measured by asking participants to indicate which of 14 potential sources influenced their decision. The resulting percentages for each source is presented in Table 1. Cadets and noncadets agreed on four of the most frequently mentioned influencers: Family, friends, personal beliefs, and career goals. Media advertising and ROTC unit requirements were among the least influential of the sources for both groups. As might be expected, a larger percentage of cadets than noncadets were influenced by ROTC instructors and other military personnel. The noncadets more often mentioned personal beliefs, career goals, and ROTC obligated service. Some significant ethnic and sex differences were found that characterized both cadet and noncadet groups. Whites chose career goals more often than Blacks or Hispanics. Blacks did not base their decision to join ROTC on personal beliefs as much as Whites or Hispanics, but were more influenced by ROTC recruiters and media advertising than the other ethnic groups. In addition, a larger percentage of males than females mentioned economic conditions as a factor in their decision.

Other observed sex and ethnic differences in ROTC influence sources were significant for cadets, but not noncadets. Specifically, a greater percentage of ROTC females than males said they were influenced by friends, (51% vs 45%, x^2=11.84, p <.01), teachers (18% vs 9%, x^2=4.15, p <.05), and ROTC instructors (48% vs 25%, x^2=7.94, p <.05), while more ROTC males than females were influenced by military lifestyle (24% vs 13%, x^2=8.79, p <.01) and personal beliefs (34% vs 25%, x^2=4.15, p <.05). The only race discrepancy between cadets and noncadets occurred in percentages reporting educational goals as influencers (x^2=10.66, p <.01). Within the ROTC group, a greater percentage of Whites (30%) than Blacks (16%) or Hispanics (11%) indicated that this influenced their decision. Within the noncadet group, on the other hand, the ethnic groups responded similarly.

ROTC and Military Attitudes. When asked how they felt about serving in the military, cadets and noncadets responded very differently (x^2=122.02, p <.001). The percentages of cadets and noncadets stating that they would serve if needed were about the same (52% for cadets and 46% for noncadets). However, a significantly greater percentage of cadets than noncadets stated that they felt a duty to serve (29% vs 6%) , while a much higher percentage of noncadets (48%) than cadets (19%) indicated that they had not given much thought to military service. In general, females and Blacks in both groups were less committed to military service. Forty-nine percent of the females in our sample indicated that they had not give much thought to military service; 41% said they would serve if needed; and only 10% believed it was their duty to serve. In contrast, only 25% of the males indicated giving no thought to military service, as compared to 54% that would serve if needed and 21% who believed it was their duty to serve (x^2=59.10, p <.001). With respect to ethnic group differences (x^2=19.19, p <.001), only 11% of the Blacks saw military service as a duty as compared to 19% of the Whites and 18% of the Hispanics, while 47% of the Blacks as compared to only 30% of the Whites and 30% of the Hispanics said they had not given much thought to military service. Fewer Blacks than Whites or Hispanics also indicated a willingness to serve if needed (42% vs 30% and 31%, respectively).

The more positive attitudes of cadets was also reflected when participants were asked to rate the attractiveness of ten aspects of an ROTC program on five point scales (see Table 2). Although cadets rated all ten aspects significantly higher than noncadets, the rank orderings of the program elements were relatively similar. Among the program elements rated most attractive were a guaranteed job after college, the scholarship program, program quality, and program activities (e.g., course modules, social functions, etc.). Obligated duty requirements, ROTC cadets, and program requirements were seen as the least attractive aspect of the program by both cadets and noncadets. The only exception to this consistent pattern was in the ranking of ROTC instructors, which was ranked first by cadets, but only fifth by noncadets. In general, Blacks and Hispanics rated these elements more positively than did the White subgroups. This pattern was significant for four of the ten program elements: Program image, program environment, ROTC cadets, and obligated duty requirements. Two of the sex differences were significant (activities and ROTC instructors), with females rating these elements more positive than the males. These ethnic and sex patterns characterized both cadet and noncadet groups in all cases except for the female Hispanic's attitude toward ROTC instructors. Cadet female Hispanics rated their instructors significantly lower than the other cadets, whereas in the noncadet group female Hispanics rated ROTC instructors more positively.

Similar patterns of results emerged when respondents were asked to rate 16 aspects of Army life (see Table 2). As with ROTC attitudes, cadet ratings were consistently higher than the ratings of noncadets, but their rank orderings were similar. Job security, officer responsibilities, and officer pay and fringe benefits were rated among the most positive by both groups, while personal freedom, prejudice, and Army living conditions were among the most unattractive elements of Army life. The largest discrepancies between the cadet and noncadet rankings were on "required mobility and travel," which was ranked fourth by the noncadets but only eleventh by cadets, and "required discipline", ranked seventh by cadets and twelfth by noncadets. Blacks in both groups rated all aspects of Army life more positively than the other ethnic groups, and Whites differed significantly from Hispanics only with respect to "job security," which they rated higher, and "personal freedom," which they rated lower. The only significant sex differences that characterized both cadet and noncadet groups were in attractiveness ratings of required travel and personal freedom, which were rated higher by females than males, and in their feelings about Army training, which was seen as more attractive by males than females. While ethnic and sex patterns were generally the same in cadet and noncadet groups, there was one exception for the element, "officer responsibility" ($F=4.87$, $\underline{p} < .05$). Specifically, whereas the mean rating for ROTC males was higher than for ROTC females on officer responsibility ($\bar{x}=3.94$ vs $\bar{x}=3.83$, respectively), the reverse was true in the noncadet group, where the male means were lower ($\bar{x}=3.26$ vs $\bar{x}=3.36$).

DISCUSSION AND CONCLUSION

Results are remarkable similar to those reported in the earlier research efforts that compared cadet and noncadet characteristics. In comparison to noncadets, cadets as well as their family and friends continue to hold more positive attitudes toward ROTC and military service. Also consistent with the more recent efforts is our finding that noncadets are not as resistant to serving in the military as they had been during the Vietnam era. In fact, there appears to be even less resistance now than at the time of the Hicks et al (1979) and Armstrong et al (1979) surveys. In addition, the present research parallels the earlier finding that cadets and noncadets rely on similar influence sources to decide whether or not to participate in an ROTC program, although cadets are still more influenced by military personnel, and noncadets by personal beliefs.

While the observed gender differences also are relatively similar to the Armstrong et al (1979) findings, the consistencies are less clear regarding ethnic comparisons. Both studies report similiar ethnic group differences with respect to military service commitment, personal feelings about ROTC and the military, parental attitudes toward military service, and the military backgrounds of parents and friends. However, the two research efforts differ somewhat in ethnic group results for friends' attitude toward military service, parental ROTC experience, and the influence sources that students use to decide whether or not to participate in college ROTC. These discrepancies may indicate that the background experiences of college students have changed (at least, with respect to ethnic groups), or they may simply be the result of sampling error. Further research is needed to clarify this issue.

What are the implication of these findings to ROTC recruiting? First, the fact that noncadets are less resistant to entering military service than they were in the 1970's suggests that they may be more amenable to ROTC recruiting campaigns. Second, the results suggest that media advertising directed toward the potential recruit is not very effective. A better strategy might be to direct advertising programs toward what were found to be the major influence sources: Parents and friends. Finally, the large number of ethnic and sex differences observed in attitudes and potential influencers suggest that a multidimensional approach to ROTC recruiting is needed. Recruiting programs that are geared only to White males must be modified and expanded if ROTC is to attract the women and minorities that will make up a significant portion of the target pool in the 1990's.

References

Armstrong, T.R., Farrell, W. S. & Card, J. J. (1979). Subgroup differences in military-related perceptions and attitudes: Implication for ROTC recruitment. Research Report 1214. Alexandria, VA; US Army Research Institute for the Behavioral and Social Sciences.

Card, J. J., Goodstadt, B. E., Gross, D. E. & Shanner, W. M. (1975). Development of an ROTC/Army career commitment model. Washington, DC: American Institutes for Research.

Hertzbach, A., Gade, P. A. & Johnson, R. M. (1985). The 1984 ARI cadet decision-making survey: An overview of results. Working Paper 85-6. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Hicks, J. M., Collins, T. & Weldon, J. I. (1979). Youth aspirations and perceptions of ROTC/military: A comparison. Alexandria, VA; US Army Research Institute for the Behavioral and Social Sciences.

McNeil, I. (1983). Demographic Imperatives: Implications for Educational Policy. Washington, D.C.: American Council on Education.

Montgomery, J. R., McLaughlin, G. W., Pedigo, B. A., Mahan, B. T., and Associates (1974). Outlook on ROTC among high school, college, and ROTC students. Blacksburg, VA: Virginia Polytechnic Institute and State University.

Table 1. Percentage of Respondents Indicating Sources that Influenced Their Decision To Join or Not Join ROTC According to ROTC Membership, Ethnic Background and Gender

ROTC Influence Sources	ROTC Membership		Ethnic Background			Gender	
	Cadet	Noncadet	White	Black	Hispanic	Male	Female
Family	39	34	38	34	33	36	36
Friends	40	39	37	44	41	39	41
Personal Beliefs	32	45***	40	27	38**	37	36
Career Goals	29	38**	36	27	32*	34	32
Educ Goals	25	28	28	24	27	26	28
Military Lifestyle	20	30**	26	22	26	24	25
ROTC Instructors	32	5***	17	22	20	17	21
ROTC Recruiters	19	15	11	34	11***	16	18
Econ Conditions	18	13	16	11	21	18	11**
Teachers Counselors	12	12	11	15	10	11	15*
Military Personnel	16	7***	12	9	12	12	10
Obligated Service	3	11***	8	5	8	7	8
Media Ads	5	8	5	11	8**	6	8
ROTC Unit Requirements	3	6	5	3	3	5	4

* p<.05
** p<.01
*** p<.001

Table 2. Attractiveness of ROTC Program and Military Lifestyle According to ROTC Membership, Ethnic Background and Gender

ROTC PROGRAM Elements	ROTC Membership		Ethnic Background			Gender	
	Cadets	Noncadets	White	Black	Hispanic	Male	Female
Guaranteed Job	4.15	3.48***	3.76	3.79	3.80	3.79	3.80
Scholarship Program	4.08	3.45***	3.76	3.77	3.76	3.76	3.73
Instructors	4.23	3.09***	3.65	3.63	3.62	3.73	3.49**
Quality	3.98	3.27***	3.60	3.66	3.58	3.65	3.54
Activities	4.02	3.11***	3.52	3.60	3.58	3.66	3.36***
Environment	3.84	3.02***	3.32	3.61	3.48**	3.44	3.37
Image	3.67	3.01***	3.23	3.53	3.40**	3.33	3.32
Requirements	3.76	2.91***	3.28	3.44	3.32	3.35	3.26
ROTC Cadets	3.60	2.96***	3.17	3.46	3.38**	3.76	3.73
Obligated Army Duty	3.38	2.72***	2.95	3.26	3.04**	3.07	2.98
ARMY LIFESTYLE Elements							
Job Security	4.28	3.71***	3.99	4.17	3.71***	4.01	3.94
Responsibilities	3.91	3.31***	3.56	3.71	3.58	3.64	3.54
Pay/Fringe Benefits	3.92	3.26***	3.52	3.85	3.40***	3.52	3.67
Officer Quality	3.79	3.14***	3.39	3.67	3.37**	3.44	3.47
Army Goals	3.77	3.13***	3.34	3.71	3.40***	3.41	3.49
Recreation	3.64	3.11***	3.35	3.55	3.21*	3.38	3.36
Required Travel	3.50	3.15***	3.20	3.71	3.20***	3.19	3.52***
Relevance to Society	3.60	3.04***	3.31	3.46	3.09**	3.34	3.26
Discipline	3.63	2.86***	3.12	3.49	3.26*	3.28	3.14
Daily Activities	3.56	2.91***	3.11	3.53	3.22***	3.26	3.16
Training	3.60	2.85***	3.15	3.41	3.15*	3.34	3.01***
Personal Relations	3.38	2.92***	3.11	3.30	3.00*	3.14	3.14
Public Image	3.24	2.84***	2.91	3.42	2.95***	2.98	3.12
Living Arrangements	3.01	2.36***	2.49	3.06	2.82***	2.71	2.63
Prejudice	2.67	2.55***	2.59	2.79	2.44*	2.58	2.61
Personal Freedom	2.73	2.44***	2.43	2.90	2.70***	2.52	2.69*

* p<.05
** p<.01
*** p<.001

CAN MAINTENANCE DATA BE USED TO DEFINE TRAINING REQUIREMENTS?

Michael Wagner
Dynamics Research Corporation
and
Major Martin Costellic
Air Force Human Resources Laboratory

In the 1970's the Air Force shifted the emphasis of maintenance training from the schoolhouse to the workplace. The average length of formal technical training decreased from 48 weeks to 25 weeks, while there was a corresponding increase in the OJT burden.

This new approach to training was expected to accomplish several things. First, it would get maintenance personnel out into the field more quickly. Second, it would allow recruits to spend more time training on equipment on which they would ultimately work. Finally, the emphasis on OJT would tie training more closely to the actual maintenance tasks that are performed in the field.

When some of the responsibility for technical training was transferred from formal schools to maintenance units, supervisors were given few resources to accomplish the additional training. As a result, the supervisor had little time to conduct OJT. Clearly, supporting mechanisms are necessary to help OJT supervisors achieve their expanded training mission.

One mechanism designed to support the unit training needs is the Field Training Detachment (FTD). Operated by ATC, FTDs smooth the transition from formal training to the unit. After completing generalized technical courses, airmen are assigned to a base whereupon they are sent to an FTD. There they receive equipment - specific training on the actual equipment maintained at the base.

In order to determine training requirements for OJT (e.g., FTD) data is required that provides accurate and timely descriptions of the specific tasks which comprise a job (including equipment references), how time is allocated among these tasks, and how well these tasks are performed.

The purpose of the Training/Job Requirements System contract with AFHRL/ID was to assess the feasibility of using automated maintenance data collection (MDC) systems to develop accurate, objective and real-time job descriptions of maintenance jobs.

Maintenance-based Job Description Reports

Data on all maintenance activities performed on each aircraft are collected as part of the MDC system. With few modifications, this system can also keep track of personnel who perform the maintenance actions and could be used to develop historical summaries or descriptions of the maintenance actions that constitute a job. The principal objections to using data from MDC in the past to define jobs have been that the data are (1) incomplete and (2) inaccurate. However, with the advent of automated maintenance data collection systems such as the Centralized Data System (CDS) on the F-16 aircraft and the Automated Maintenance System (AMS) on C-5A aircraft, these objections are no longer valid. Specifically:

(1) Automated maintenance data collection systems ensure that data is complete and accurate. Work order generation and tracking ensures that maintenance actions are recorded; on-line editing produces accurate and complete entry of maintenance information; a centralized system contains daily data from all maintenance sites;

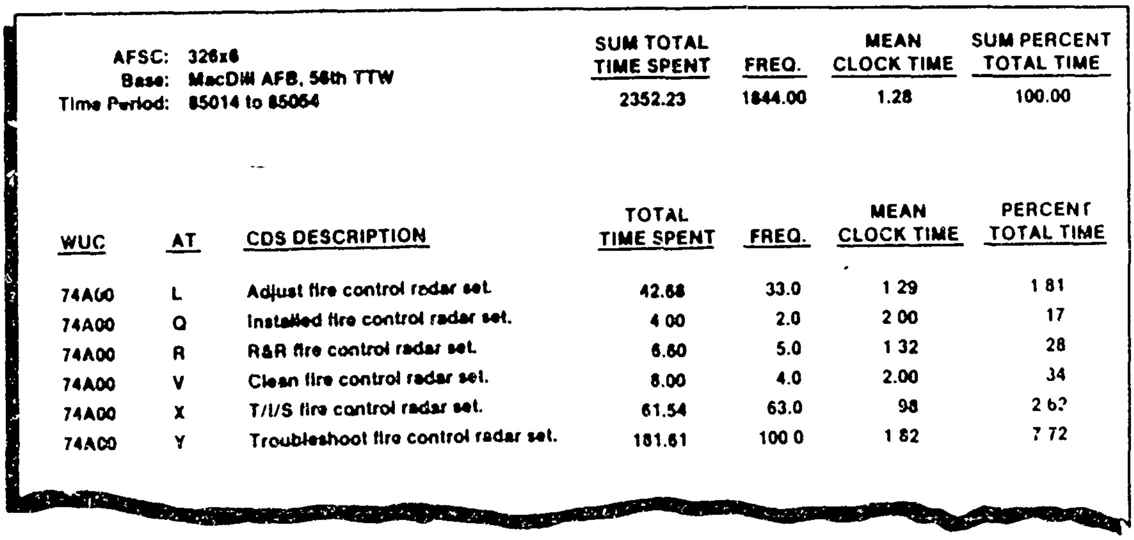
(2) Maintenance data consists of a set of standard task elements which refer to equipment involved (work unit codes) and actions taken (action taken codes). In addition, the clock time for specific maintenance actions, although influenced by the general workload level, provides accurate estimates of relative task times;

(3) Maintenance task data can be collected which identifies each crew member involved in a maintenance action. This data can be obtained without imposing an additional burden on maintenance personnel.

(4) Maintenance data on individuals can be aggregated over time to construct reports summarizing all maintenance actions performed by an individual. Reports can be constructed as a function of skill level, AMU, or base.

(5) The computer data base and terminals for collecting these data are already in place, at least for the F-16 aircraft.

A small portion of a maintenance job description report that can be generated using maintenance data is shown in Figure 1.



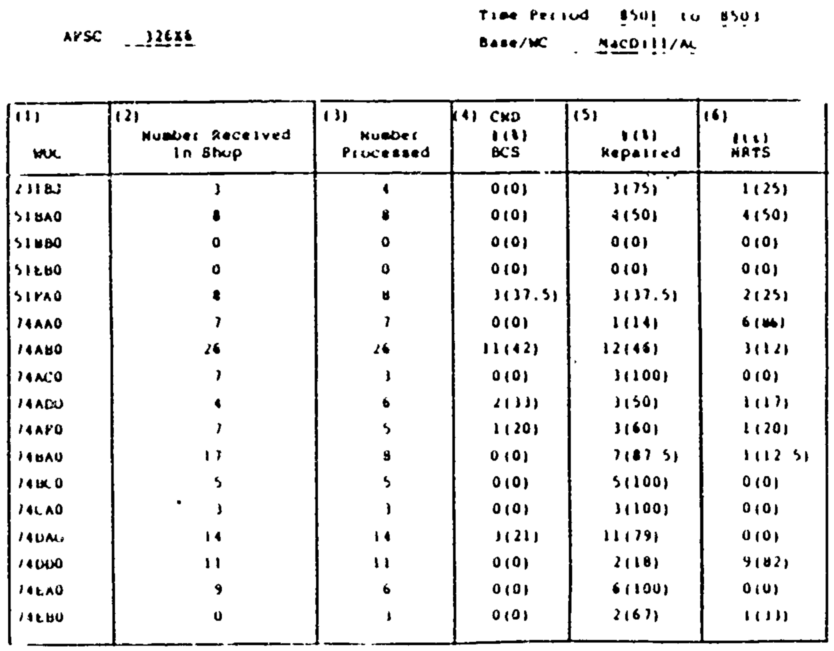
		SUM TOTAL TIME SPENT	FREQ.	MEAN CLOCK TIME	SUM PERCENT TOTAL TIME
AFSC:	326x6	2352.23	1844.00	1.28	100.00
Base:	MacDill AFB, 56th TTW				
Time Period:	85014 to 85064				

WUC	AT	CDS DESCRIPTION	TOTAL TIME SPENT	FREQ.	MEAN CLOCK TIME	PERCENT TOTAL TIME
74A00	L	Adjust fire control radar set.	42.68	33.0	1 29	1 81
74A00	Q	Installed fire control radar set.	4 00	2.0	2 00	17
74A00	R	R&R fire control radar set.	6.60	5.0	1 32	28
74A00	V	Clean fire control radar set.	8.00	4.0	2.00	34
74A00	X	T/I/S fire control radar set.	61.54	63.0	98	2 62
74A00	Y	Troubleshoot fire control radar set.	181.61	100 0	1 82	7 72

Figure 1. F-16 CDS Job Description Report Excerpt

This maintenance data-based report provides a listing of specific tasks performed and how time is allocated among these tasks.

A second type of report, shown in Figure 2, provides performance data. This report, called a WUC Removal Action Performance Report, identifies how often avionics LRUs removed at the flightline are later found to be serviceable when they are tested in the Avionics Intermediate Shop (AIS). This information can be generated for either a base or a work center and will identify which serviceable LRUs are frequently removed unnecessarily. This information can be used to decide if additional resources need to be allocated to the training of troubleshooting particular equipments.



AFSC 326X6

Time Period 8501 to 8503

Base/WC MacDill/AC

(1) WUC	(2) Number Received In Shop	(3) Number Processed	(4) CND #(%) BCS	(5) #(%) Repaired	(6) #(%) NRTS
231BJ	3	4	0(0)	3(75)	1(25)
51BA0	8	8	0(0)	4(50)	4(50)
51MB0	0	0	0(0)	0(0)	0(0)
51EB0	0	0	0(0)	0(0)	0(0)
51PA0	8	8	3(37.5)	3(37.5)	2(25)
74AA0	7	7	0(0)	1(14)	6(86)
74AB0	26	26	11(42)	12(46)	3(12)
74AC0	7	3	0(0)	3(100)	0(0)
74AD0	4	6	2(33)	3(50)	1(17)
74AF0	7	5	1(20)	3(60)	1(20)
74BA0	17	8	0(0)	7(87 5)	1(12 5)
74BC0	5	5	0(0)	5(100)	0(0)
74CA0	3	3	0(0)	3(100)	0(0)
74DA0	14	14	3(21)	11(79)	0(0)
74DD0	11	11	0(0)	2(18)	9(82)
74EA0	9	6	0(0)	6(100)	0(0)
74EB0	0	3	0(0)	2(67)	1(33)

Figure 2. Sample WUC Removal Action Performance Report

Users of these reports could include on-the-job training (OJT) supervisors; base training analysts; maintenance training managers at TAC; and analysts at ATC, MPC, and OMC involved in describing maintenance jobs, developing training standard , and evaluating training effectiveness.

Linkages to Current Job Descriptions and Training Standards

Job descriptive information derived from maintenance data support the definition of training requirements only if this information can be linked to current job descriptions and training standards. Only after such linkages are created can the existing job descriptions and training standards be evaluated and subsequently revised. Thus, in order to realize the full benefit of collecting maintenance data on individuals, linkages must be established for all the AFSs on which maintenance data is collected. Figure 3 provides an illustrative example of how linkages are established among maintenance tasks (F-16 CDS), occupational survey tasks (OS), and specialty training standards (STS).

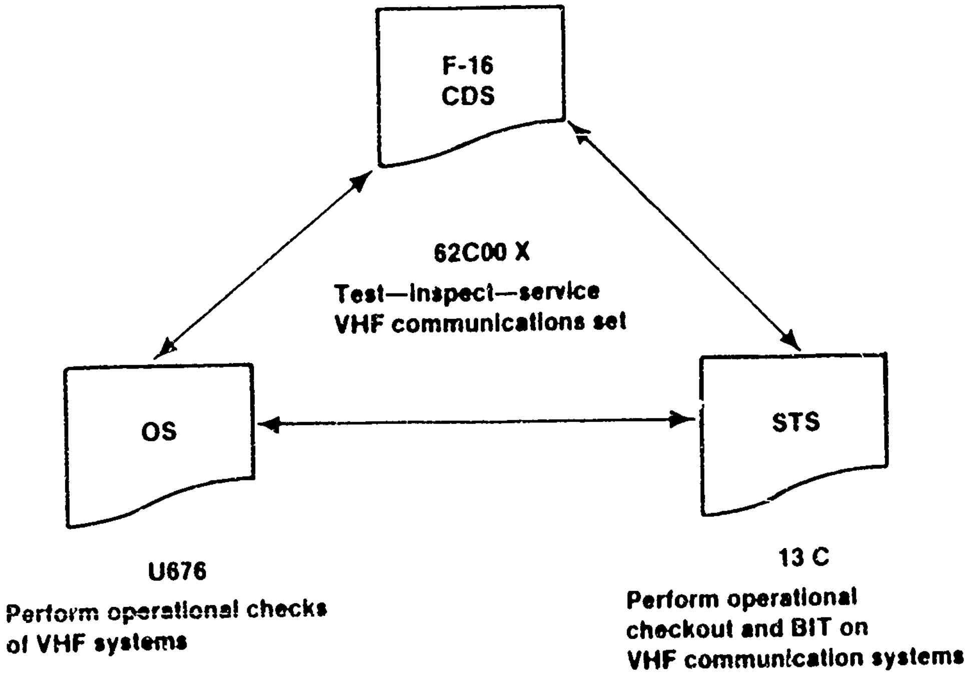


Figure 3. Task Linkages

Summary

Several capabilities were demonstrated in this study which have wide applicability. These capabilities include using automated maintenance data collection (MDC) systems to collect task data on maintenance personnel, linking these maintenance tasks to current AFS job descriptions (i.e., occupational survey reports) and specialty training standards, and developing reports which can be used to construct training requirements tailored to specific groups (i.e., bases) and to identify training deficiencies.

TRAINING EMPHASIS: THE BEST TASK FACTOR AVAILABLE

Lieutenant David L. Hardy, Lieutenant Colonel Charles D. Gorman,
and Dr. Walter E. Driskill
USAF Occupational Measurement Center

PURPOSE

The paper illustrates three important facets of the training emphasis (TE) factor. First, TE ratings provide the best data for establishing training priorities in comparison to the other task factors frequently cited as needed for curriculum decisions. Second, the reason training emphasis is the best task factor for establishing training priorities is the significant correlations that exist between the TE ratings and other task factors. Third, TE data are sufficient for instructional systems development (ISD) decisions. Following a brief historical summary of the training emphasis factor and a discussion of the just mentioned facets of TE data, a brief description of how TE technology is applied by the USAF Occupational Measurement Center (OMC) will be provided.

HISTORY AND BACKGROUND

The training emphasis task factor developed from a long, detailed, and complex research program. Christal (1970) first proposed the gathering of task factor information using subject-matter experts as the pool of raters from which samples could be drawn. Mial and Christal (1974) and Mead (1975) conducted the initial research using the policy-capturing approach to successfully predict training priorities. In addition, Mead's research was particularly hopeful in suggesting an intimate integration of ISD practices, occupational survey data, and curriculum design. In 1977, Stacy, Thompson, and Thomson reported that standard occupational survey techniques were reliable for the collection of task training factors.

The first major research specifically on training emphasis was reported by Ruck, Thompson, and Thomson (1978). This report established the ground work for use and analysis of TE data. Some of their recommendations were: that TE data should be collected and not predicted; that training emphasis ratings be collected while the routine collection of task delay tolerance and consequences of inadquate performance task factors should be discontinued, and that ratings be separately collected for each Air Force specialty.

The TE task factor research was released for operational use to the USAF Occupational Analysis Program, USAFOMC, Randolph AFB, Texas, in 1979. Driskill and Mitchell (1980) mentioned the exstensive gathering of TE data for use by technical training curriculum developers and training managers. In 1981, Staley and Weissmuller presented a paper on interrater reliability in the CODAP programs. They suggested training emphasis is a stable task factor in noncomplex specialties, but that more research needs to be done on the application of TE data to complex specialties. Jansen (1982 and further in 1985) tackled the issue of complex specialties and proposed a common rating policy (CRP) approach which is adequate for all but a few specialties.

TE vs OTHER TASK FACTORS

Before suggesting an integration of the TE task factor into a prominent position in ISD theory and practice, a discussion of the correlations between TE and a number of other task factors investigated by Goldman (1985), and Ruck, Thompson, and Stacy (in preparation) is necessary.

Goldman, in a different approach, both statistically and methodologically, showed the same conclusions alluded to by Ruck, et al (1978), that the TE task factor is the best task factor available and that other "training" factors are highly correlated with TE. The factors used in Goldman's research are identified in the Instructional Systems Development (ISD) Eight Factor Model. In addition to the TE factor, those factors included:

1. Percent of members performing
2. Average percent of time spent by members performing
3. Task Learning Difficulty (LD)
4. Consequences of Inadequate Performance (COIP)
5. Task Delay Tolerance (TDT)
6. Probability of Deficient Performance (PDP)
7. Immediacy of Performance (IP)
8. Relative Frequency (RF)

The conclusions reached by Goldman suggest that instead of nine separate factors, there is only one clearly defined training factor; and, in terms of predicting critical vs noncritical tasks, rather than nine factors, there is really only TE. Also, by collecting a minimum number of task factors, efficiency in data gathering and analysis is significantly enhanced. These conclusions are a result of the high correlations found between the TE task factor and the other factors as shown in Table 1.

Research conducted by Ruck, Thompson, and Stacy (in preparation), has direct impact on the use of task factors in the ISD program. Jansen (1985) states, "the utility, reliability, and validity of training emphasis ratings in terms of ISD theory have been demonstrated", by the aforementioned authors. The thrust of their research was the development of a task training emphasis scale and some training priority equations. The following task factors were used in their study:

1. Training Emphasis
2. Probable consequences of inadequate performance
3. Task delay tolerance
4. Learning difficulty
5. Percent members performing
6. Percent time spent
7. Task grade-level index

The results of Ruck, Thompson, and Stacy's analysis shows a very positive correlation between TE ratings and several other task factors (see Table 2). The conclusions reached by Ruck, et al, in relation to other task factors used are: that TE ratings are "construct valid" because they can be predicted by using ISD training factors; TE ratings are reliable since supervisors give their ratings independently and have high agreement with one another; and, the process used in their study for arranging task lists in order of training priority should be adopted by the training community to improve training. Further, they recommend that TE ratings be routinely collected and "that the consequences of inadequate performance and task delay tolerance factors be collected for those specialties for which they are of special interest, since recommended TE ratings include consideration of these factors."

TE AND ISD

The ISD program for the Air Force is operationally defined in AFP 50-58 (1978), then outlined in AFM 50-2 (1979). From all ISD sources, there are seven stated task factors; they are:

1. Percent members performing
2. Number of members performing
3. Consequences of inadequate performance
4. Task delay tolerance
5. Task learning difficulty
6. Frequency of performance
7. Training development time

From previous discussion in this paper, it has been established that training emphasis is the best task factor available, and other factors are, perhaps, redundant--mainly because TE ratings include consideration of most of the other task factor variables. The key point of Ruck, et al's, report is that training emphasis should be the most extensively used task factor when using the ISD approach to make training decisions. Indeed, their research provided several CODAP computer programs that allow for presentation of TE data in very effective, usable formats.

CURRENT APPLICATION OF TE TECHNOLOGY BY USAFOMC

Since the release of TE technology for operational use, USAFOMC routinely has collected TE ratings as a standard part of conducting occupational analyses. After the ratings have been entered into the computer, the CODAP program REXALL is used to identify and eliminate deviant raters. This process is repeated until an acceptable interrater reliability for a single rater is at least .20, and the interrater reliability for all raters is .90 or more. After achieving these minimum levels of acceptability, any remaining deviant raters are individually reviewed for retention or rejection, based on the correlation of their ratings with mean ratings for the total group. At this point, raters that have been eliminated are examined to determine if there are any systematic similarities among them. Similarities may suggest the presence

[illegible]

[illegible] Between Training Emphasis and [illegible]

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	65	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	71

[illegible]
= Training Emphasis
= Learning Difficulty
= Consequences of Inadequate Performance
= Task Delay Tolerance
[illegible] Performance
= [illegible] Performance
= [illegible] Frequency
= Percent of Members Performing
= Average Percent of Time Spent by Members Performing
[illegible] Table adapted from Tables 1 and 2 in [illegible]dman's (196[illegible]) Report

Table [illegible]

[illegible] Between Training Emphasis and [illegible] Task Factors

AFS	[illegible]	2	3	4	5	6	7
[illegible]X		[illegible]	-[illegible]	[illegible]	[illegible]	75	-66
[illegible]4X		44	-[illegible]	[illegible]	73	54	-71
[illegible]4X4		[illegible]	-[illegible]	[illegible]	77	71	-21
[illegible]X4		[illegible]	-[illegible]	-19	77	71	11
[illegible]X1		[illegible]	-79	-[illegible]	78	[illegible]	-47
[illegible]X4		[illegible]	-27	-20	68	80	-25
[illegible]		[illegible]	-[illegible]	-4[illegible]	[illegible]	77	-84
[illegible]X2		[illegible]	-1[illegible]	-[illegible]	47	78	-43
[illegible]X[illegible]		[illegible]	-[illegible]	[illegible]	48	82	-78
[illegible]X		41	-[illegible]	-[illegible]	79	65	-73
[illegible]X[illegible]		[illegible]	-[illegible]	-64	71	63	-61
[illegible]X		[illegible]	-	[illegible]	60	54	-65
[illegible]X1		13	-[illegible]4	-17	70	63	-36
[illegible]X2		[illegible]	-7[illegible]	-43	85	75	21
[illegible]X[illegible]		33	-[illegible]	-40	75	61	-67
[illegible]6X		28	-[illegible]	-04	48	43	-43
[illegible]X		[illegible]	[illegible]	-[illegible]	74	65	-06
[illegible]X		[illegible]	[illegible]	-[illegible]1	85	79	-77

[illegible]
Training Emphasis
[illegible] Consequence of Inadequate Performance
[illegible] Delay Tolerance
4 Learning Difficulty
[illegible] Percent Members Performing - First Job
6 [illegible] Time Spent [illegible]
[illegible]
[illegible] Table adapted from Appendix [illegible] in [illegible] et al. report
[illegible]
[illegible] and [illegible]

of multiple policies in the career ladder. Also, if more than 10 percent of the raters in the original sample are eliminated, a separate REXALL should be run, even though there are no apparent systematic similarities, to determine if another rating policy exist.

Once an acceptable REXALL has been achieved, the process of identifying deviant tasks begins. To do this, the mean of the task standard deviations (AVTSD) and the standard deviation of the standard deviations (SDSD) multiplied by 1.5 are added together. Tasks whose standard deviations exceed this value are defined as deviant tasks. These identified deviant tasks are then examined for systematic similarities. If the deviant tasks are random, then the data can be used with confidence. If the deviant tasks are systematically related, then the data are examined for multiple policies or the data will be used as they are with a caution to the user that within certain areas the raters disagreed as to the amount of emphasis that sould be placed on training.

SUMMARY

This paper briefly discussed the training emphasis task factor. To begin, background to the evolving technology of training emphasis was capsulized. This established the ligitmacy of the training emphasis ratings. After establishing some of the initial research results, a comparison of training emphasis to a number of other training-related task factors was made. This comparison showed TE ratings to be superior to the other task factors because of its predictive value and consistent incorporation of other task factors within it's ratings. Next, a short discussion pointed out the importance of training emphasis ratings in the ISD training development process. Finally, a brief description was given on how the USAFOMC is currently applying TE technology. The purpose of this paper is to urge the incorporation and integration of the training emphasis rating as the single most useful task factor available to the training community, and to those who are using the Instructional Systems Development approach for making training decisions.

REFERENCES

[illegible] Manual 50-2 Instructional System Development. Washington, D.C.: Department of the Air Force, [illegible] Mar 19[illegible].

[illegible] Pamphlet 50-[illegible] Handbook for Designers of Instructional Systems [illegible] Washington, D.C.: Department of the Air Force, 15 July 1978.

[illegible] Implications of Air Force Occupational Research for Curriculum [illegible] Smith and [illegible] (Eds.), Report of a Seminar. [illegible] and Techniques of Vocational Curriculum Development. Minneapolis, MN: University of Minnesota, 19[illegible]

[illegible], W. [illegible] and Mitchell, [illegible] The USAF Occupational Analysis Program. [illegible] Technology. Proceedings of the 22nd Annual Conference of the Military Testing Association, Toronto, Canada, October 1980.

[illegible], [illegible] A. Determining the Best Set of Training Factors. Proceedings of the [illegible] International Occupational Analysts Workshop, Randolph AFB TX, May [illegible]

[illegible]nsen, H. [illegible] Identification of Rating Policies in Training Emphasis Task Factor Data. Proceedings of the 24th Annual Conference of the Military Testing Association, San Antonio TX, November 1982.

[illegible]nsen, H. [illegible] Training Emphasis Task Factor Data: Methods of Analysis. (AFHRL-TR-84-50). Brooks AFB TX: Manpower and Personnel Division, Air Force Human Resources Laboratory, May 1985.

Mead, D. F. Determining Training Priorities for Job Tasks. Proceedings of the 17th Annual Conference of the Military Testing Association, Indianapolis, IN, September 1975.

Mead, R. E., and Christal, R. E. The Determination of Training Priority for Vocational Tasks. Proceedings, Psychology in the Air Force Symposium. USAF Academy, April 1971.

Ruck, H. W.; Thompson, N. A.; and Stacy, W. J. (In preparation). Development of a Task Training Emphasis Scale and Training Priority Evaluation. Brooks AFB TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Ruck, H. W., Thompson, N. A., and Thomson, D. C. The Collection and Prediction of Training Emphasis Ratings for Curriculum Development. Proceedings of the 20th Annual Conference of the Military Testing Association, Oklahoma City, OK, October-November 1978.

Staley, M. R., and Weissmuller, J. J. Interrater Reliability: The Development of an Automated Analysis Tool. Proceedings of the 23rd Annual Conference of the Military Testing Association, Washington, D.C., October 1981.

TRAINING EFFECTIVENESS ANALYSIS OF DRAGON SIMULATORS

Mr. Walter G. Butler

US Army TRADOC Systems Analysis Activity

The requirement to provide combat units with soldiers who are qualified and proficient in the use of the DRAGON antiarmor weapon system currently entails the use of training ammunition, ranges, and troop support facilities. The high cost and limited availability of these resources are constraints on the effectiveness of the institutional training program to produce such soldiers. In addition, current training devices limit the provision of a broad range of realistic target engagement conditions.

This paper presents the results of an evaluation of the cost and training effectiveness of three alternative DRAGON training programs. The programs were almost identical in structure, differing only in the training device(s) used to support the program. Three training devices were involved; the launch effects trainer (LET), the launch environment simulator (LES), and the simulated tank antiarmor gunnery system - DRAGON (STAGS-D).

The LET and LES were already being used for DRAGON training; the STAGS-D was in development. These devices were combined in the following mixes to produce the alternative training programs:

- LET and LES (base case)
- STAGS-D alone
- STAGS-D and LES

SYSTEM DESCRIPTIONS

DRAGON is a medium range wire-guided antitank weapon. It is fired by one man from the right shoulder with a bipod supporting the front end of the launcher. The two major components of the DRAGON are the round and the tracker. The round consists of a missile prepacked in its launcher. The tracker is mated to the round before firing. After firing, the empty launcher is discarded or destroyed and the tracker is used on the next round. The DRAGON can be fired at night by replacing the day tracker with the thermal night tracker (AN/TAS-5).

To engage a target with DRAGON, the gunner places the crosshairs in the tracker sight on the target and depresses both a thumb safety switch and a palm firing switch. There is a 0.6 second delay before the rocket motor fires to eject the missile from the launcher. The gunner continues to track the target until missile impact. During flight an infrared (IR) flare at the end of the missile allows the tracker to determine the relative position of the missile to the aim point. The tracker sends commands through the wire line to 30 pairs of thruster motors on the missile to correct deviations from the aim point. When corrections are needed, a single pair of thrusters will fire. These thrusters cannot be fired again.

[illegible] day or night conditions with [illegible] different ranges. Targets are [illegible] that can move on [illegible] right or left. Each target vehicle [illegible] sensitive camera can [illegible] stationary or moving at a prescribed [illegible]. The target vehicle can also rotate [illegible] target and return simulated fire on the gunner position. [illegible] electronically inserted in [illegible] to provide realistic firing conditions.

[illegible] DRAGON launcher, a performance [illegible] with an infrared source. A standard [illegible] to complete the system. [illegible] teach firing and tracking procedures and to provide [illegible] effects he would [illegible] when firing a live DRAGON. When conducting a LET engagement the [illegible] and then wait through the [illegible] with the live missile. The launcher assembly then fires [illegible] grenade cartridge at the rear of the launch tube. [illegible] noise and visual simulation and activates a weight shift [illegible] simulates the weight loss of a missile leaving the launcher. The [illegible] by the LET is not as loud as that of an actual firing, neither [illegible] heat or obscuration from smoke and debris. The gunner tracks [illegible] at an actual range of about [illegible] meters. The LET simulates [illegible] from [illegible] to [illegible] meters (in [illegible]-meter increments) by requiring [illegible] to track for different elapsed times before the device terminates [illegible].

[illegible] addition to DRAGON training. The device operates under [illegible] concept as the LET in that the gunner learns firing and tracking [illegible] and his performance is displayed on the monitoring set (same one used with the LET). The difference between LET and LES is that the LES [illegible] closer simulation of the noise, heat, and obscuration produced by a live DRAGON firing. The launcher has been modified to replace the [illegible] assembly with a chamber in which an explosive mixture of Methyl Acetylene Propadiene, Propane Propylene (MAPP) gas and oxygen is [illegible]. A separate control box is used to house the MAPP gas and oxygen [illegible] the controls needed to fill the launcher chamber with the proper mix of gases. When the LES is fired a glow plug ignites the gas [illegible] to produce noise, recoil, heat, and smoke that approximates that of [illegible] DRAGON firing. Further obscuration is produced by fragmentation of [illegible] on both ends of the launcher and by dust and debris raised by [illegible]. The weight loss sensation is simulated by the downward [illegible] of the blast force at the rear of the launcher. Because [illegible] level intensity, the Surgeon General of the Army has directed that no soldier will fire the LES more than five times per day or more than [illegible] per year.

METHODOLOGY

The sample of students for the study consisted of 150 soldiers who [illegible] in the DRAGON gunner course during one station unit training [illegible] at Fort Benning, GA. Approximately thirty soldiers were [illegible] classes conducted from 9 October through 15 November [illegible] sample. The soldiers were randomly assigned to one of [illegible] training programs under evaluation. Eighty-seven percent of the [illegible] were active Army enlistees with the remainder being in the National [illegible] were made between these types of soldiers in the [illegible] had been in service for less than ten weeks and had [illegible] training.

[illegible] of the Infantry Training [illegible]. These soldiers were qualified DRAGON [illegible] the DRAGON gunner course. They had [illegible] training program on the operation of the STAGS-D before [illegible].

[illegible] at Fort Benning was the base case. The [illegible] consisting of a combination of lecture and hands-[illegible] by two gunnery tables used to determine the [illegible] of each student. Both the LET and LES are used for [illegible]. The qualification tables require the soldier to engage a [illegible] under various conditions using the LET and the LES.

[illegible] alternative training programs were structured in the same [illegible] with appropriate substitution of training [illegible] replaced both the LET and LES with the STAGS-D, while [illegible] the LES and used the STAGS-D in place of the LET. The [illegible] the LET/LES POI, the STAGS pure POI, and the [illegible].

[illegible] data concerning soldier proficiency with an actual DRAGON [illegible] with each simulator used during training and soldier perceptions [illegible] training programs.

[illegible] of the soldiers with the DRAGON was assessed in live firing [illegible] at the end of each week. The live firings were conducted on South [illegible] range at Fort Benning using inert missiles. The target was a manned [illegible] target tank (MITT) with a driver and tank commander aboard. The [illegible] alternately right to left and left to right at 5 miles per [illegible] meters from the firing point.

[illegible] live missile firing was recorded on video tape by TRASANA personnel using the gunner aiming sensor (GAS) system. The GAS consists of a six-power [illegible] objective lens, a six-foot fiber optic bundle, a color video camera, [illegible] recorder, and a monitor. The objective lens, which produces the [illegible] the DRAGON tracker, is attached to a base plate bonded [illegible] the tracker. The fiber optic bundle carries the image from the lens [illegible] array of 10-micron fibers to the camera enclosed in a [illegible]. The camera transmits the image to a recorder located [illegible]. The image is stored on tape and simultaneously displayed [illegible] viewing of the engagement. The engagement can be [illegible] to assist in determining hit or miss and to distinguish between a [illegible] and a gunner error in case of a miss.

[illegible] sensor provides a reticle in the image transmitted to the [illegible]. The reticle can be [illegible] to the tracker reticle by adjustment mechanisms located on the bracket that attaches the lens to the base plate on the tracker. The alignment of the GAS and tracker reticles was checked before and after each shot and proved quite stable as only small corrections were necessary after several firings.

The presence of the reticle on the video tape of a live firing permitted a detailed analysis of gunner performance. The Data Sciences Division, National Range Operations Directorate at White Sands Missile Range, New Mexico, analyzed the tapes using manual and automated techniques to provide the magnitude of gunner aiming error at various points during the engagement, the times at which certain critical events occurred, and other data describing the engagment.

Data relating to soldier proficiency with the simulators were collected during the two qualification firing tables conducted at the end of each class. These data include the number of hits recorded on each of the qualification tables, the number of hits with each simulator, and the qualification status of each soldier as a result of his performance on the tables.

The perceptions of the soldiers concerning the simulators and training programs were obtained through surveys and written comments after the soldiers had fired the live DRAGON.

RESULTS

The video tapes of the live firings were analyzed to answer four questions about each soldier's performance during his DRAGON engagement. First, did the soldier achieve a successful launch by regaining control of the missile after undergoing the launch effects (heat, noise, shock, etc.)? Secondly, did the soldier keep the missile in the air long enough to reach the target; in this case, about eight seconds? Thirdly, did the soldier fly the missile through a rectangle containing, but slightly larger than, the target; that is, did he come "close" to a hit? Finally, did the soldier hit the target? A comparison of the results for the soldiers in each of the three training programs showed no statistical difference in performance among the groups. In fact, the largest difference between any two groups in any area was six percentage points. This analysis did reveal several points of interest concerning the engagements. Nearly half of all the target misses were caused by unsuccessful launches. Almost all soldiers who achieved a successful launch flew the missile "close" to the target. Many of the missiles that came close to the target but missed did so because of excessive tracker movement by the gunner during the last few seconds of missile flight. These observations provide the developers and providers of DRAGON gunner training with specific areas in which increased emphasis can produce significantly improved performance.

The video tapes also allowed examination of the maximum deflection due to launch effects of the tracker reticle from the ideal aim point on the target. Measurements were taken in both the horizontal and vertical planes. Again, no differences were observed between the groups. It was noted, however, that those soldiers who hit the target tended to have allowed a smaller deflection after launch in the horizontal plane than those soldiers who missed. This tendency did not hold in the vertical plane. DRAGON gunner training has long emphasized control of the tracker in the vertical plane to reduce the possibility of grounding the missile during or just after launch. The results noted above suggest horizontal control is at least as important in ultimately achieving a target hit.

These analysis results show that no differences could be found between the training effectiveness of the three training programs. Similarly, the twenty year life cycle costs of the programs were found to be nearly the same with the largest cost difference being ten percent.

The results of the qualification exercises using the various simulators were compared to the live-fire results to determine if any relationships could be found between soldier performance with the devices and performance with the DRAGON in terms of target hit or miss. The twenty qualification shots are divided into two tables (called Table III and Table IV) of ten shots each. The number of shots with each simulator is shown below for each training program:

	LET/LES	STAGS Pure	STAGS/LES
Table III	10, LET	10, STAGS	10, STAGS
Table IV	5, LET 5, LES	10, STAGS	5, STAGS 5, LES

The qualification standard used in previous DRAGON gunner courses was at least eight hits on each firing table. When this standard was applied to the results of the qualification firings of the soldiers involved in the study, thirty-one (62%) of the soldiers in the LET/LES program were deemed qualified while no one in the STAGS pure program and only two (4%) of the STAGS/LES soldiers were qualified. The qualification standard was determined to be inappropriate for the latter two programs, since all the programs had been shown to be equally effective. Considering only the soldiers in the LET/LES program, the existing qualification standard correctly predicted the hit/miss outcome of the live-fire engagement for 60% of the soldiers. By contrast, an alternative standard that ignored Table III and required at least nine hits on Table IV correctly classified the live-fire results of 68% of these soldiers. Still another standard that ignored the LET altogether and required at least four hits out of the five LES engagements on Table IV provided 70% correct classification.

For soldiers in the STAGS pure program, no relationships were found between performance on the STAGS and live-fire performance.

For soldiers in the STAGS/LES POI, a relationship was found that involved a requirement of at least one hit with each simulator and a total of at least four hits among the twenty qualification shots. This requirement provided 63% correct classification of live-fire performance.

The most notable result of this area of the analysis is that the LES seems to provide some relationship between performance with the simulator and performance with the DRAGON itself.

The compilation and analysis of the survey responses and written comments revealed the following soldier perceptions:

o all simulators, except possibly LES, needed more launch effects (noise, heat, smoke, etc.)

o STAGS was much more sensitive than DRAGON to gunner movement

o STAGS hurt confidence and morale

o Difficulties with STAGS and DRAGON seem easier

These perceptions indicate strengths and weaknesses of the STAGS simulator. Apparently the over-training provided by the STAGS in teaching steady tracking techniques caused the soldiers to feel they had good control over the actual DRAGON. The increased sensitivity of the STAGS that provided the over-training, however, also caused the soldiers to achieve far fewer hits during practice and qualification than did the soldiers who trained with the LET/LES. The soldiers hit 78% of all targets engaged with the LET or LES and 30% of all targets engaged with STAGS. This led to confidence and morale problems among the soldiers training on the STAGS.

SUMMARY

The analysis of the three DRAGON training programs produced the following results:

o the programs produced essentially the same levels of performance in the test soldiers

o the programs cost essentially the same

o performance with the LES provided information about performance with DRAGON

The following results concerning DRAGON gunner training in general were also observed:

o half of the live-fire misses were due to unsuccessful launches

o most of the remaining misses were due to tracking instability during the last few seconds of flight

o tracking control in the horizontal plane during launch was just as important as vertical control

Career Involvement of Recent West Point Graduates

Jerome Adams
United States Military Academy
West Point, NY 10996-5000

John D. Richards
Academy of Health Sciences
Fort Sam Houston, TX 78234-6100

In today's rapidly changing society, it is important to stay abreast of current developments, especially within the context of significant military attitudes which could affect the readiness of our Forces. Much has been written in recent years about career decision-making, planning strategies, career involvement, dual careers (Adams 1980; Davenport 1984; Hall & Hall, 1979), and overall commitment and adjustment to career and organization. Moreover, Moskos (1977) has noted changes from an institutional to an occupational organizational orientation among members of the military. Some suggest that today's military has less of the rational organizational devotion and dependency, and more of the entrepreneurial protean man (Hall, 1976; 1979). Hall (1979) also distinguishes between moral and calculative organizational commitment, with a shift being analogous to the phenomenon observed by Moskos (1977). In terms of personal decisions the concern with career planning and strategies utilized in making those decisions also has changed.

The present paper reports recent research pertaining to career attitudes and proclivities of male and female West Point graduates (Adams 1984, Adams & Yoder 1984). The data to be reported were imbedded within the context of larger, world-wide study of early adjustment experiences of US Military Academy graduates. Specifically, strategies of career planning, degree of career involvement, and overall commitment and adjustment to the Army were analyzed.

METHOD

Respondents. The participants were selected based upon stratifying by sex, work specialty and geographic assignment. The samples reflected the same proportions of branch specialty and location as the entire population of graduates. A ratio of three men were selected for each female respondent. The two year interviews were conducted with 116 members of the class of '80 and 104 members of the class of '81 at locations in CONUS, Korea, and Germany. A breakdown by gender and location is presented in Table 1.

Procedure. The two year protocol was the first to include extensive questions on social aspects of Army, Army life, and the interaction of an Army career and family life.

Interviews were usually conducted at the respondents' duty station, though on occasion commuting to a centralized location was required. Interviews were generally scheduled for 4-6 individuals, though, in practice they varied anywhere from 1 to 7. There were 5 interviewers in all, varying widely in number of interviews conducted, but all interviewing individually. Interviewers also varied as to sex, race, and military/civilian status. Interviews were usually 1 to 2 hours in duration, and were all tape recorded for later content analysis.

Results and Discussion

Several questions were asked about career involvement, officer role, and the Army lifestyle. There was a substantial amount of ambiguity about the meaning of the questions, both among respondents and among interviewers. The primary difficulty was the meaning of "career involvement." Did it mean attention on the job and career activities or did it mean intent to remain on active duty after the five-year obligation. In this analysis, responses which focused on the former alternative will be discussed in this section, while those addressing the five-year decision will be considered with a subsequent question on change in career commitment.

The most interesting results in responses to this question were in the differences between 1980 and 1981 graduates (Table 2). About 25% of the 1980 graduates indicated that they were more involved in their careers. The figure for 1981 graduates was almost 50%. There seems to have been a large shift in career interest during this one year time frame. The shift occurs for both males and females, although females are generally less likely to become more involved than males. A similar pattern is observed in the answers to questions on changes in adjustment. Class of '81 respondents are more likely to have experienced a positive change and males are more likely than females to have had a positive change. Again, interpretation of these results raises the question of statistical artifact or real difference. A significant proportion of the involvement change can be attributed to the radical change of males stationed overseas. Is this an accident or did the increased value of the dollar make everything seem rosier?

Answers to the questions on what factors influenced greater career involvement or adjustment to an officer's role shed little light on this issue (Table 3). Males in the class of '81 were much more likely to cite confidence, job satisfaction, and familiarity, but there seems to be no clue about why these factors should be more strongly felt on overseas assignments. They do, however, make sense as reasons for assuming a higher level of participation in an Army career.

Females from the class of '80 were more likely to express a negative change toward their officer roles than men. The majority of men who perceived a positive change simply felt more comfortable and familiar with their role. The majority of women who experienced a negative change were upset by different treatment for women and felt that men were just more likely to take the Army's "flack" than women. The latter comment probably says more about general attitude differences between male and female officers, as perceived by female respondents, than any other expressed in the interviews. Both groups perceive the same problems, for example both identify similar job dissatisfaction factors, but 1980 female graduates tend to be less tolerant of the negative aspects than men. Perhaps it is because many begin their careers from a more defensive perspective, as the first female West Point graduates. members of the class of '81 do not seem to have similar views.

Only a small number of individuals indicated a change in adjustment to the Army lifestyle. The most mentioned negative factors associated with the Army lifestyle are adjusting to time requirements and mandatory social functions. The most frequently cited positive factor is getting married. Six of ers were married in the previous year and this seemed to signi .antly contribute to their overall adjustment or at least it made them happier.

NOTE: This document represents the views of the author and not the official position of the U.S. Army, or any other governmental agency unless so designated by other authorized documents.

The interview tapes were transcribed by Dr. Richard J. Orend under contract DAAG 60 85 M 1799. The research program was supported by contract 13 ARI 85-28 from the Army Research Institute (Jerome Adams, principal investigator).

REFERENCES

Adams, J. Report on the Integration of Women with the Corps of Cadets at West Point -- Project Athena IV. West Point, NY: AG Printing Office 1980.

Adams, J. & Yoder, J.D. Women Entering Non traditional Roles: Conflicts between Sex Roles and Work Proceedings: The Academy of Management National Convention. Boston, MA: 1984.

Adams, J. Project Proteus: Early Career Preparation, Experiences, and Commitment of Female and Male West Point Graduates. Vol I, II, & III. West Point, NY: AG Printing Office 1980.

Davenport, L.C. Dual-Career Couples: Spouses Influence on Career Decisions. Unpublished Master's thesis San Diego State University, Summer 1984.

Hall, D.T. Careers in Organizations. Santa Monica, CA: Goodyear 1976.

Hall, D.T. Organizational Commitment: Theory, Research, and Management. Unpublished paper Evanston, IL: Northwestern University, April 1979.

Hall, F.S. & Hall D.T. The Dual Career Couple. Reading, MA: Addison-Wesley 1979.

Moskos, C. "From Institution to Occupation: Trends in Military Organization" Armed Forces & Society 1977, 44-50.

TABLE 1
Total Participants

	Class of 80		Class of 81	
	Number	% of Total	Number	% of Total
Male	88	75.9	85	81.7
Female	28	24.1	19	18.3
Total:	116		104	
Stationed in CONUS*	83	71.6	80	76.9
Stationed OVERSEAS**	33	28.4	24	23.1
Total:	116		104	
CONUS Males	67	80.7	68	85.0
CONUS Females	16	19.3	12	15.0
Total:	83		80	
OVERSEAS Males	21	63.6	17	70.8
OVERSEAS Females	12	36.4	7	29.2
Total:	33		24	

*Including Hawaii

**Germany and Korea

TABLE 2

Change in Extent of Career Involvement

			Class of 80		Class of 81	
				Total Sample		
Change in Career Involvement:		More Involved	29	25.0	49	47.1
		Less Involved	15	12.9	9	8.7
				By Gender		
Change in Career Involvement:	Males -	More Involved	22	25.0	41	48.2
		Less Involved	8	9.1	6	7.1
	Females -	More Involved	7	25.0	8	42.1
		Less Involved	7	25.0	3	15.8
				By Location		
Change in Career Involvement:	CONUS -	More Involve	22	26.5	34	42.5
		Less Involved	8	9.6	8	10.0
	OVERSEAS -	More Involved	7	21.2	15	62.5
		Less Involved	7	21.2	1	4.2

TABLE 3
Factors Influencing Career Involvement

	Class of 80			Class of 81		
Positive - More Involvement	Males #	Females #	Total #	Males #	Females #	Total #
1. Confidence, job knowledge	2	0	2	10	1	11
2. Job satisfaction, enjoy work	5	0	5	13	2	15
3. Security	2	0	2	1	0	1
4. Willing to put in time	2	0	2	3	1	4
5. Competitiveness	1	0	1	0	0	0
6. Think about it more	0	1	1	3	0	3
7. Married to another officer	0	1	1	0	0	0
8. Comfortable	0	1	1	0	0	0
9. Sense of duty involved with goals	0	1	1	0	0	0
10. Increased responsibility	0	0	0	2	2	4
11. Thinking about change in assignment	0	0	0	8	1	9
12. More in control	0	0	0	1	0	1
Negative - Less Involvement						
1. Frustration ("brick walls")	1	0	1	2	0	2
2. Lack job freedom	1	0	1	0	0	0
3. Current assignment	1	2	3	0	1	0
4. No other competition in life	1	0	1	0	0	0
5. Problems leave individual numb	1	0	1	0	0	0
6. Not in MOS	0	2	2	0	0	0
7. Dissatisfied with commander	0	1	1	0	0	0
8. Not enough time with troops	0	3	3	0	0	0
9. Poor commanders	0	3	3	0	0	0
10. Can't quit/move	0	2	2	0	0	0
11. Not good career for women	0	3	3	0	1	1
12 Senior officers to me are money oriented	0	1	1	0	0	0
13. Not motivated to keep up in field	0	0	0	1	0	1
14. Not as concerned about always doing well	0	0	0	0	1	1
15. Future in Army seems dim	0	0	0	2	0	2
16. Disillusioned by WP on role of officer	0	0	0	1	0	1

Mentoring in the United States Air Force:
The Mentor's Perspective

Captain Francis Lewandowski
9th Organizational Maintenance Squadron, Beale AFB, CA

Captain Benjamin L. Dilla
Air Force Institute of Technology, Wright-Patterson AFB, OH

As a follow-on to Roche's 1979 survey of business executives and Uecker's 1984 survey of Air Force officers exploring the mentoring concept, this study examined mentoring from the perspective of senior Air Force officers who have been both proteges and mentors. Respondents were 95 Air Force officers selected for the 1985 entering class at Air War College. Of these, 58% responded that they have had a mentor and 48% stated that they have been (or currently are) mentors for junior officers. Comparisons were made to the earlier data gathered by Roche and by Uecker with respect to military background, career factors, and effects of mentoring on the respondents. Furthermore, roles fulfilled by a mentor were contrasted from the perspectives of the protege versus the mentor. Results for this sample of senior officers verified the prevalence and perceived importance of mentoring in the United States Air Force.

Introduction

Mentoring has been defined as a relationship between a senior member and a junior member of an organization in which the senior member is influential in molding and shaping the career of the younger member (Uecker, 1984). The concept of mentoring has recently received considerable attention throughout the field of management. Trade journals, in particular, abound with articles ranging from cross-gender mentoring to reasons why one should (or should not) enter into a mentoring relationship.

Empirical work in the area is limited to two recent surveys. In a study by Heidrick and Struggles, Inc., and reported by Roche (1979), 1250 executives recently appointed to their positions were surveyed. This study found that "nearly two-thirds of the respondents reported having had a mentor or sponsor" (Roche, 1979, p. 14). The research also discovered that mentored executives earned more money at a younger age and had a higher degree of satisfaction with their jobs and their career progress. Furthermore, those in the mentored group were better educated and more likely to have formulated and followed a career plan.

A modified and expanded version of Roche's survey was recently applied in the U.S. Air Force; Uecker (1984) surveyed 252 officers attending Air Command and Staff College (ACSC) and Air War College (AWC) to examine the prevalence and effects of mentoring in the Air Force. Results of this survey were reported at last year's Military Testing Association conference (Uecker & Dilla, 1984). Mentoring was not found to be as prevalent in the military sample (42% versus Roche's 64%), but the effects were similar. Mentored officers were better educated and more likely to have formulated a career plan. They were more likely to have been promoted early (a parallel to Roche's finding of earning more money at an earlier age) and had greater job and career progress satisfaction (Uecker & Dilla, 1984).

An important extension of Roche's survey in Uecker's research involved examination of the roles or functions played by a mentor in helping to develop the subordinate. The survey asked respondents to categorize ten potential roles identified by Lea and Leibowitz (1983, as being primary, major, secondary roles, or roles not played by their mentors. Results showed that respondents saw their mentors primarily in the sense of role models, motivators, and advisors rather than in more directive roles such as sponsor and protector (Uecker & Dilla, 1984).

Since the prevalence of mentoring in the Air Force officer corps had been previously supported, the thrust of this project was to reexamine the prevalence of mentoring and to further investigate the phenomenon from the mentor's perspective. To accomplish this, a sample of high-potential senior officers was surveyed to find out if they had had mentors, if they had become mentors for others, and, in general, to examine their point of view concerning the mentoring process within the Air Force. The survey, based on those of Roche (1979) and Uecker (1984), also attempted to estimate the perceived effect the mentor had on the career of his protege and on the Air Force.

Methodology

Sample

The sample for this study needed to be drawn from a population of Air Force officers senior enough to have had the opportunity to be mentors as well as to have had mentors. Air Force policy precluded sending questionnaires to general officers who would most closely parallel Roche's (1979) sample in meeting this criteria. Authorization to survey designees for the 1985 entering class of AWC was granted. These officers, 112 in number, were lieutenant colonels and colonels (O-5 and O-6) coming out of a variety of leadership and staff positions including squadron commanders, directors at air division level, and system program directors. Furthermore, they had been identified, by the fact of their selection for AWC, as having high potential for further advancement.

Procedure

Surveys were mailed to officers at their duty addresses several months before their departure for AWC. Participation was voluntary, and respondents were assured of anonymity. They were asked to mark their responses directly on the survey instrument and return it in a postage-paid return envelope. Because of the frequent negative connotation of mentoring in the Air Force, the survey cover letter defined mentoring and carried an endorsement by the Dean of the School of Systems and Logistics (a USAF O-6).

Measures

The survey instrument was based on Uecker's (1984) questionnaire used the previous year with AWC and ACSC students. The format of the survey was expanded so that the officers would respond to items separately from the perspective of being a protege and from being a mentor. Eighteen items focused on the officer as protege, fourteen as a mentor; fourteen items were concerned with desirable characteristics of a mentor, and ten items addressed personal background, promotion history, and current satisfaction. Space was provided to list important characteristics of a protege and other open-ended comments and suggestions at the end of the survey.

Results

Respondent Profile

Of the 112 officers sent surveys, a total of 95 (85%) responded. The majority received their commission from ROTC (58%) vice OTS (27%) or a service academy (15%). The median response for age at commissioning was 22 years, although ages ranged from 20 to 33. Most had an advanced degree (93%) and had received at least one below-the-promotion-zone (BPZ) promotion (87%). The largest group identified with no single major command (36%) although the three largest commands were well-represented--SAC (21%), TAC (16%), and MAC (12%).

Mentoring Experience

Of the 95 respondents, 58 reported having had a mentor who took a personal interest in them and guided or helped mold their careers. This 61% rate of mentoring is very close to the 64% reported by Roche (1979) for the private sector. Using a normal approximation to the binomial distribution and comparing the percentages, no significant difference was found. The median response for the number of mentors was two, with a range from one to six. Most (23 of the 58) said their mentor had first exhibited an interest in them only fairly recently, after the tenth year of service. The majority (35) indicated that they still had a relationship with their mentor, although most of these (24) described the relationship as "friendly" rather than "close". The largest group (26) reported having had a general officer as a mentor; the next largest group (12) said their mentor was their immediate supervisor. When asked how much influence their mentor had exerted, most said it was substantial (25) or moderate (21); only a few chose the extremes of extraordinary (5) or little (7) influence.

In the second section of the survey, 46 officers (48% of the sample) stated that they had served as a mentor to another individual. Two-thirds of these officers (30 of the 46) reported that they currently had one or two proteges. When asked how long their longest mentoring relationship had lasted, the modal response was a tie between two and three years. This finding may be an indication that these officers had only recently moved into positions of command and leadership where they could be mentors, or it may be a result of the very mobile Air Force way of life, such that relationships do not extend much beyond one assignment. When asked how much influence they had exerted over their proteges, again there was a fairly equal split between substantial (22) and moderate (24) influence with none reporting extraordinary or little influence.

Effects of Mentoring

This study failed to find any significant differences between mentored and unmentored groups with regard to formulation of a career plan, job satisfaction, or early promotions as had been found in the previous research. However, the small sample size of this study may have been a limiting factor. A significant difference was found between groups with respect to career progress satisfaction (t=-2.85; p .01), with the mentored group reporting greater satisfaction.

For the 46 officers who had served as mentors, there were no significant differences with respect to early promotions or career progress satisfaction; however, they had significantly higher job satisfaction than their contemporaries who had not been mentors ($t=-2.25$; $p<.05$).

Roles of the Mentor

This study examined roles of the mentor from the perspective of both proteges and mentors. This allowed for a contrast between the two perspectives as well as a comparison with Uecker's (1984) results for proteges from a similar sample. For each of the ten roles identified by Lea and Leibowitz (1983), respondents were asked to indicate if the role was "most important" (assigned a scale value of 3), major (2), secondary (1), or a role not played (0). Mean responses for each role for this study's group of 46 mentors, 58 proteges, and Uecker's (1984) group of 106 proteges are presented in Table 1. Rankings of the ten roles within each group are also presented for ease of comparison.

Table 1
Roles of the Mentor

	Average Rating[1] (& Rank within Group)		
Role	AWC Mentors (n=46)	AWC Proteges (n=58)	AWC & ACSC Proteges[2] (n=106)
Advisor	2.158 (1)	1.706 (5)	1.853 (2)
Counselor	1.932 (2)	1.708 (4)	1.598 (5)
Motivator	1.798 (3)	1.734 (2)	1.800 (3)
Role Model	1.711 (4)	1.840 (1)	1.924 (1)
Guide	1.675 (5)	1.321 (8)	1.500 (7)
Teacher	1.611 (6)	1.344 (7)	1.441 (8)
Communicator	1.561 (7)	1.093 (9)	1.505 (6)
Supporter	1.500 (8)	1.511 (6)	1.613 (4)
Sponsor	1.343 (9)	1.716 (3)	1.426 (9)
Protector	1.095 (10)	0.713 (10)	0.964 (10)

[1] Ratings and assigned scale values were:
Most Important = 3; Primary = 2; Secondary = 1 ;
Not Played = 0.

[2] Data adapted from Uecker, 1984.

The largest difference in ranks occurred for the controversial role of "sponsor". This term, which often carries a negative connotation in the Air Force, was rated relatively low by Uecker's (1984) sample and by the smaller group in this study which rated the roles from the perspective of being a mentor. Yet, when rated from the perspective of being a protege, the respondents of this study gave it the third highest mean rating; in fact, it received the largest percentage of "most important" responses for this sample.

Other notable differences included the advisor role, which received the highest mean rating for the group of mentors (also the highest absolute rating in any of the groups), but emerged fifth when rated from the protege's perspective in this sample (second in Uecker's [1984] sample). In the opposite direction were the results for "role model", rated highest by both groups of proteges but only fourth among the roles rated by the mentors.

Discussion

This study found mentoring in the Air Force to be as prevalent in the Air Force as in Roche's (1979) private sector study. Despite this similarity, there were distinctions in the nature of these relationships. In Roche's study, two-thirds of the respondents reported that a mentoring relationship began for them in the first five years of their career, while the majority of officers stated their relationships did not commence until after ten years of service. Also, relationships in the Air Force seem to be shorter and more varied. Only one-third of Roche's respondents reported having had two or more mentors (despite their earlier start), while 73% of the AWC officers (64% in Uecker's [1984] study) reported two or more mentors.

Roche (1979) found that 62% of the 1250 executives had proteges. In this study, only 43% of the officers indicated they had proteges; interestingly, most of these had previously had a mentor, so the phenomenon seems to be largely self-perpetuating.

With regard to the effects of mentoring, officers who had mentors had greater career progress satisfaction, while those who served as mentors had higher job satisfaction. These results are consistent with previous findings (Roche, 1979; Uecker, 1984) and with the effects predicted from career development theories discussed elsewhere (Dilla, 1985).

Contrary to Roche's (1979) findings, no difference was found for this sample with regard to formulation of a career plan. It should be noted that Uecker found an relationship for his total sample but not for the AWC respondents alone. This difference from Roche's results may be due to the centralized reassignment processing within the Air Force. Respondent comments indicated that "needs of the Air Force" and unforeseen career opportunities sometimes dictated career plan changes.

Regarding the roles of the mentor, proteges most often described their mentor as a role model, although the mentors assigned less importance to this function. This difference seems natural since it would be difficult for the mentor to tell to what extent the protege is observing and striving to emulate his behavior. The most important roles from the mentors' perspective were those of advisor, counselor, and motivator, roles that are more active but not highly directive. Proteges were in agreement with the relative importance of the motivator role but seemed less willing to admit that their mentors had to "counsel" them. This difference may be due to some of the negative connotations to the term itself.

There was clear agreement across all groups that the mentor does not serve as a protector to the protege, or at least does so very infrequently. It is noteworthy that the mentors assigned greater importance to this role than either group of proteges; further study of the occurrence of this function may be merited. Respondents also tended to play down the importance of the sponsor role, with the exception of the AWC proteges. It appears that the more senior group of AWC officers perceived that their mentors had provided growth opportunities for them to a greater extent than did Uecker's

(1984) combined sample of AWC and ACSC students. However, the low rating by the same group of officers when viewing the roles as mentors is puzzling. Further research with more detailed definitions of the terms and larger sample sizes may be the only way to resolve or clarify these differences.

Comments at the end of the surveys indicated that many officers understood and supported the use of mentoring in the Air Force; however, many harsh, negative comments revealed stereotypes and misconceptions even at this senior officer level. The Air Force should publicize the reasons for, and potential benefits of, the informal mentoring process. Furthermore, research on mentoring should continue and be expanded to both a broader scope and higher level of officers in order to better understand the dynamics and effects of the process.

References

Dilla, B.L. (1985). Mentoring: Cause and effect of good leadership. Proceedings of the Association of Human Resource Management and Organizational Behavior, pp. 630-635.

Lea, D., & Leibowitz, Z.B. (1983). A mentor: Would you know one if you saw one? Supervisory Management, 28, 32-35.

Roche, G.R. (1979). Much ado about mentors. Harvard Business Review, 57(1), 14-28.

Uecker, M.E. (1984). Mentoring and leadership development in the officer corps of the United States Air Force. Unpublished master's thesis, 84S-30, Air Force Institute of Technology, Wright-Patterson AFB, OH.

Uecker, M.E., & Dilla, B.L. (1984). Mentoring as a leadership development tool in the United States Air Force. Proceedings of the 26th Annual Conference of the Military Testing Association, pp. 423-428.

Preparing to Evaluate Career Path Changes: Pre-Change Database Development

Robert E. Chatfield
Robert F. Morrison

Navy Personnel Research and Development Center

The purpose of this study was to establish, if feasible, a pre-change database that the Navy could use to assess the effectiveness of major revisions to the Surface Warfare Officer (SWO) career path.

Background

The Surface Warfare community has traditionally stressed that its officers be "generalists" with diversified work experience perceived to be career enhancing. However, recent program and career assignment policy changes have implemented changes that increase specialization prior to the Executive Officer (XO) tour. The origin of these significant changes to the Surface Warfare Officer (SWO) career path can be traced to the 1981 Surface Warfare Commanders Conference (SWCC) where the issue of fleet readiness and its relationship to SWO technical competence and experience was discussed. The SWCC concluded that the generalist approach to officer development actually undermined and degraded readiness by not allowing officers to gain the requisite technical experience required to manage operate today's complex shipboard systems. The following FY81 statistics were cited to support their contention:

--Only 41% of engineering department heads had prior engineering experience as a division officer.

--Only 38% of operations department heads had prior operations experience as a division officer.

--Only 48% of weapons/combat systems department heads had prior weapons/combat systems as a division officer.

In every instance, fewer than half of the officers had any work experience in the type of department that they were being asked to manage.

As a direct result of the SWCC concern, major revisions were made to SWO training programs and career policies over the next two year period. The goal of the revisions, promulgated as NAVOP 105/83, was to develop more technically experienced officers at the key department head level, which would in turn be expected to promote an increase in the operational readiness of ships. As a department head, the SWO must manage the maintenance, operation, and employment of complex systems/platforms.

The new career path is tailored to provide them with the opportunity to gain specialized technical expertise in the first 10-12 years of their Navy career without the expectation of being a generalist. Specifically, the new career path is structured fc specialization in one of the three major surface warfare areas: engineering, operations, and weapons combat systems.

Advocates of the revised SWO career path insist that this shift toward specialization was technology-driven and inevitable. However, critics argue that these newly created "specialists will encounter difficulties at the XO CO level where familiarity with the entire ship is critical. Only time, and a well planned evaluation of the impact of the revisions will prove which of these contrasting views is indeed correct.

Objectives

The overall objectives of the study were:

1. Identify existing measures of ship department readiness and performance.

2. Build a pre-change (prior to implementation of the new career path database consisting of FY82-FY84 data for the most promising measures.

3. Assess the consistency across measures and the stability of individual measures over the three year time period.

4. Identify measures acceptable for inclusion in a pre change database that the Navy could use 3-5 years hence to evaluate the impact of the career path revisions on fleet readiness.

As the project was of short duration (9 months) and funding was limited, it was not possible to design and implement procedures for evaluating career path effects. Therefore, the study was limited to investigating the appropriateness of using existing measures of ship department readiness and performance to assess the impact of the changes.

Method

A multiple measure, or "performance profile," approach was taken under the assumption that no single performance measure could acceptably serve as an evaluative standard for the new career path. It was theorized that by focusing on essentially a battery of readiness indices, the impact of the new policies could be inferred. An analysis of inter-measure consistency and year-to-year stability of individual measures would indicate if the multiple measure approach was appropriate.

Two different classes of data were used to build a preliminary, pre-change database. The first class, yielding "ship data," involved identifying existing measures of departmental readiness and performance and collecting data for those satisfying acceptability criteria. The second class provided "personnel data" that might describe the performance of department heads as reflected in the personnel records of individual officers.

In order to determine which ship data measures of readiness were appropriate for inclusion in the pre-change database, and to provide a means of cross-measure comparison, nine evaluative criteria were developed. For example, numeric data were required for quantitative summarization. Additionally, scores could not be inflated such that improvement potential and variability suffered. In general, measures had to be objectively scored, focus on department head performance, be available in a format facilitating rapid computer analysis, and be well accepted by the SWO community. These criteria were developed to address issues of reliability and validity while also reflecting project-related constraints.

Over fifty interviews were conducted with senior level officers in the Surface Warfare community to identify existing measures of departmental readiness. Information collected about measures included: measurement cycle; scoring procedures; fleet reputation, etc. The interviews yielded 21 "candidate" measures for possible inclusion in the pre-change database. These measures were then individually scored on the nine evaluative criteria. A measure was eliminated from consideration if its standing on any criterion was completely unacceptable.

Six of the 21 measures proved to be acceptable on all criteria, however, data for only three were actually collected and analyzed. The three remaining measures were not included because a substantial amount of clerical effort would have been required to extract relevant information and transcribe it to a format facilitating computer analysis. The three ship data measures included in the pre-change database were:

1. Propulsion Examining Board (PEB) Assessments (both the Lighting Off Examination (LOE) and Operational Propulsion Plant Examination (OPPE).

2. Nuclear Weapons Technical Inspections (Nuclear Weapons Acceptance Inspection (NWAI); Nuclear Technical Proficiency Inspection (NTPI); and Defense Nuclear Security Inspection (DNSI).

3. Departmental Excellence Awards (presented in conjunction with the Battle Efficiency Awards cycle).

Two types of personnel data were collected: 1) The proportion of department heads failing to complete their full tour during the FY82-FY84 period; and 2) The annual number of department heads involved in a Detachment for Cause (DFC) action. The former is often informally indicative of poor job performance while the latter is the formal administrative action for relieving an officer of his duties. The assumption is that the incidence of both of these occurrences would decrease if the new career path was having its intended effect.

Results

Ship data were evaluated by several different methods to determine if their inclusion in the pre-change database was appropriate. First, each measure was examined for year-to-year consistency. Substantial fluctuation or variability in pass rates across the three year time period was indicative of poor consistency and "reliability." While perfect consistency from year-to-year was not expected, 10%-15% variation was the maximum acceptable. Second, measures were assessed for trends in rising or falling pass rates across the three year time period. Finally, cross-measure comparisons were made to evaluate the extent to which the various assessments of departmental readiness were in agreement.

The generally poor stability of ship data performance measures is reflected in Figure 1 which presents LOE and OPPE pass rates for several ship types. Substantial year-to-year fluctuations in pass rate percentages are evidenced in the variability columns. For the LOE, variability was unacceptably high for both ship types presented. That is, yearly pass rate differences ranging from 20%-50% do not provide the stable performance baseline required to evaluate effects of the new career path. While OPPE variability was marginally acceptable for DD/DDG (11%/16%), it was completely unacceptable for LSD/LST (50%/89%). No consistent trends in pass rates were evident for either measure. Pass rates appeared to rise and fall in a random, unsystematic manner.

FIGURE 1

Examples of LOE and OPPE Pass Rate Instability

	PACIFIC			ATLANTIC			VARIABILITY (HIGH-LOW)	
SHIP TYPE	1982	1983	1984	1983	1984	1985	PAC	LANT
LOE Exam								
FF/FFG	100	67	88	100	88	80	33	20
LHA/LPH/LKA/LPD	86	60	60	33	50	0	26	50
OPPE Exam								
DD/DDG	76	65	68	75	71	88	11	16
LSD/LST	100	67	50	100	50	13	50	87

All numbers expressed as percentages.

In contrast to PEB inspections, NWTIs displayed rather good stability. A summary of NWTI pass rates for four ship types is provided in Figure 2. Several of the variability percentages fall below the 10%-15% tolerance previously discussed. Pass rates were exceptionally stable for Auxilliary class ships where the variability was only 4% in the Pacific fleet and 7% in the Atlantic fleet across the three year period.

FIGURE 2

Examples of NWTI Pass Rate Stability

	PACIFIC			ATLANTIC			VARIABILITY (HIGH-LOW)	
SHIP TYPE	1982	1983	1984	1983	1984	1985	PAC	LANT
CG/CGN	83	100	100	89	100	89	17	11
DD/DDG	85	88	74	85	81	84	14	4
FF/FFG	81	92	91	93	90	75	11	18
AUXILLIARY	72	71	75	81	74	79	4	7

All numbers expressed as percentages.

Inter-measure comparisons made on a year-to-year basis indicated that there was little consistency among readiness indices. For example, there was essentially no relationship between ship types' pass rates on the OPPE and on the NWTI. Although all measures ostensibly assess the operational readiness of ship departments, inter-measure consistency was lacking.

The final ship data measure, departmental excellence awards, displayed both excellent (less than 5% variability) and poor (mc e than 35% variability) stability. There appeared to be some type of interaction between ship type and award category as awards were often stable for one ship type but highly volatile for another. Because of this inconsistency, departmental excellence awards do not provide the stable performance baseline desired.

Analysis of personnel data measures revealed that neither was appropriate for inclusion in such a pre-change database. The percentage of department heads leaving the job early was quite small (maximum of 10.1%) and no year-to-year trends were evident. Figure 3 provides a summary of performance-related DFCs by major department for the three year period. Once again, very few department heads are represented in this measure as there were a total of only 59 such actions. Variability of DFCs year-by-year was greater than desirable and no trends were identified. For example, while DFCs were steadily on the rise in the Atlantic fleet, a similar trend was not evident for the Pacific fleet.

FIGURE 3

Performance-Related DFL Actions for Major Department Heads in FY82-FY84

	ENG		OPS		W/CS		TOTAL	
YEAR	PAC	LANT	PAC	LANT	PAC	LANT	PAC	LANT
1982	5	3	0	2	3	0	8	5
1983	4	7	4	5	2	0	10	12
1984	1	12	3	2	3	3	7	17
TOTAL	10	22	7	9	8	3	25	34

Conclusions

Analysis of ship data performance measures indicated that such indices were not stable enough to establish a performance baseline for assessment of career path influences. The NWTI was the only measure to display any promise for such an application. In addition, consistency among multiple measures of readiness and performance was not evident. Results from one type of evaluation were not in agreement with those from another on a year-by year basis.

Personnel data measures proved to be equally unacceptable for evaluating the impact of the new career path on readiness. In both instances, the number of department heads represented was very small and stability was lacking.

In conducting analysis of these existing measures of departmental readiness and performance it became evident that there are a variety of factors beyond the scope of the department heads' control that can moderate results. While a number of situational (e.g., experience level of department personnel, CO influence) factors are essentially random and other miscellaneous factors (e.g., age of ship) can be statistically controlled for in analysis, there remain a plethora of systemic factors whose influence cannot readily be identified. For example, such factors as the fleet's operations tempo, changing evaluation standards, budgetary limitations, etc. can dramatically affect readiness assessments. The influence of these systemic factors undoubtedly contributed to the poor stability of individual measures and lack of inter-measure consistency.

In summary, multiple measures of departmental readiness and performance evaluated in the present study should not be used to assess the revised SWO career path. In light of the poor stability documented for existing measures of ship department readiness and performance, it is recommended that the Navy conduct a critical review of these measures in an effort to improve their reliability and validity.

USAF OFFICER PROFESSIONAL MILITARY EDUCATION[1]

by

JOHN M. BELL, First Lieutenant, USAF
Directorate of Military Occupational Structures
National Defence Headquarters, Ottawa, Ontario, Canada

and

JOSEPH S. TARTELL
Occupational Analysis Division
USAF Occupational Measurement Center, Randolph AFB, Texas

INTRODUCTION

To ensure that USAF officers will have the skills to carry out their leadership and managerial responsibilities, the Air Force provides a variety of precommissioning and postcommissioning Professional Military Education (PME) courses that can be taken at specific career points. To determine whether these courses are truly responsive to the needs of USAF personnel, the Air War College requested that the USAF Occupational Measurement Center (USAFOMC) conduct an occupational survey that would help validate or redesign the curricula of officer PME courses. Specifically, USAFOMC was asked to determine leadership, management, and communicative tasks performed by company and field grade officers, and to determine the need of, or benefit from, the various PME schools and courses. Precommissioning PME sources are the USAF Academy, AFROTC units, and Officer Training School (OTS). Postcommissioning PME sources include Squadron Officer School (SOS) for junior officers, Air Command and Staff College (ACSC) for middle level officers, and Air War College (AWC) for senior officers.

METHODOLOGY

To meet the request, USAFOMC developed and used five separate survey instruments to collect data. The first of these was a task list containing 347 leadership, management, and communicative tasks under 14 duty headings. These tasks were to be rated by survey respondents on a 9-point scale according to the relative amount of time spent on each task, compared to the time spent on each of the other leadership, management, and communicative tasks they performed.

Second, data on the difficulty of the same leadership, management, and communicative tasks discussed above were collected via a task difficulty booklet. Difficulty was defined as "the amount of time needed to learn to do each task satisfactorily." Respondents were asked to rate each task on a 9-point scale according to its relative difficulty, compared to the other tasks

Third, using an inventory with the same leadership, management, and communicative tasks discussed above, respondents were asked to rate each task on a 10-point scale according to its need in Air Force educational programs.

A fourth set of data was collected via a survey containing a list of 275 topics from the curricula of officer PME courses. For each of these topics, respondents were asked to rate the extent to which knowledge of, or skill in, each topic was necessary to perform their present job (need-in-job).

Using the same list of topics, respondents to a fifth booklet were asked to rate the extent to which knowledge of, or skill in, each topic was necessary to function as a career officer (need-in-career). An 8-point scale was used to rate the topics in both of these surveys.

These surveys were validated and approved by representatives of the various PME schools. Random samples were selected for administration of the surveys between June 1983 and April 1984. Representative samples across the surveys were achieved.[2]

RESULTS

The analysis of the task performance data showed a pattern of increasing involvement in leadership, management, and communicative tasks as officers increased in rank from lieutenant to colonel. Supporting this pattern were data that showed the percentages of officers who had supervisory responsibilities increased from 38 percent among lieutenants to 95 percent among colonels. Additionally, the percentage of total job time spent on the tasks in the survey increased from 56 percent to 81 percent, from lieutenant to colonel, respectively. Related to these was the organizational assignment pattern, which showed the manner in which the percentage of officers assigned to organizational levels as rank increased. This pattern of increasing involvement is not surprising, but it does illustrate the changing nature of most officers' responsibilities. Further, it provides some rationale for a continuing multiphased professional development program.

The data showed a great amount of diversity in the tasks performed across career fields and ranks.[3] At the same time, there was a substantial amount of similarity in certain tasks performed, particularly in some tasks dealing with communication skills and motivating others. Further, the differences between ranks in relative time spent on tasks in each duty were very small.

The analysis of the education emphasis data revealed a low reliability of raters. In short, insufficient agreement on the amount of education required to perform tasks existed among officers in general and across career fields and ranks. This finding reemphasized the diversity of opinions and needs concerning USAF officer PME.

The reliabilities of task difficulty data were quite high.[4] Fifty-six tasks received high difficulty ratings from the total group of raters, and most were communicative tasks, such as drafting or writing relatively high level documents (officer effectiveness reports, plans, staff papers, and reports). Other highly rated tasks involved skills such as determining resources; administering disciplinary actions to civilians; and ordering, persuading, or influencing those superior in rank or position. Fifty-eight tasks received low difficulty ratings. Many of these, also, dealt with communicating and motivating, but were of much lower-level activities, such as drafting or writing short note replies and reading professional publications. Other lowly rated tasks include providing informal feedback, attending training sessions, and maintaining appearance standards.

Analysis of officers' self-perceived need of various PME topics in their jobs and in their careers showed a great deal of diversity within most of the career fields and ranks. In spite of this diversity, it was possible to create a rank order listing of topics from each of these groups displaying the relative need of these topics in the job or in the career. In general, the data showed those PME topics which officers believed were needed in their jobs were topics they also needed in their careers; conversely, they generally felt those topics not needed in their job were not needed in their careers. Only minor differences were seen in perceptions across these two studies. Communication topics (eg, effective listening, active writing, logical thinking) were rated as most needed by officers in their jobs and careers; senior officers saw the needs as about equal, while junior officers perceived them as greater in their job than over their career. Topics on the military environment, national security, and military employment generally received the lowest relative need ratings, both in terms of in the present job and over the career. Rated lowest were topics on other services policies and doctrines, economic theories and systems, and foreign relations.[5]

Background data, collected in all five surveys, were used to assess, among other things, officers' job satisfaction and perceptions of benefits from PME. Of the 10,607 responses to the 5 parts of the project, 10,177 were used in assessing background data responses. (The difference here reflects the elimination of duplicate responses of officers who were asked to complete more than one kind of survey booklet, since background questions across booklets were identical.)

While the job satisfaction indicators were high, perceptions of benefits from USAF PME were, at best, mixed. Those who participated in precommissioning PME at the USAF Academy, indicated the highest degree of benefit, while the extent of benefit from ROTC and OTS PME was much lower. Officers indicated very low benefit from SOS by correspondence, and lower benefits in general from PME by correspondence or seminar, than through residence programs.

COMPARISON TO PREVIOUS RESEARCH

Results of this survey were compared to those of the Officer Professional Military Education Curriculum Validation Project produced by USAFOMC in August 1980 and the Survey of Air Force Officer Management Activities and Evaluation of Professional Military Education Requirements produced by the Air Force Human Resources Laboratory in December 1969.

While the scope and complexity of PME studies increased over the years, certain results were consistent. Analyses of task involvement over time, for example, showed a consistent increase in the level and number of leadership, management, and communicative tasks with every increase in rank. Differences in task performance across utilization fields were also observed across studies; operations personnel, for example, were consistently lower performers (compared to other fields) at lower ranks, increasing the scope of their leadership, management, and communicative responsibilities as they increased in rank to the point of equality with other fields at the grade of colonel.

Task difficulty data (collected in this and the 1980 studies) generally agreed on the tasks officers consider most and least difficult. Heading the lists were drafting or writing high-level official correspondence, conducting high-level investigations, and determining resources. Low on both lists were attending training sessions, maintaining personal appearance standards, and drafting or writing low-level correspondence.

The perceptions of need of particular curriculum topics were analyzed differently in this study than that data in the two previous surveys, but two similar findings were evident. First, there was a great amount of diversity within sub-groups analyzed as to what PME topics officers perceived they needed. Second, there was a general overall consistency as to what were the most and least needed topics within sub-groups. While previous studies used average ratings and this one used a rank order method to assess need, agreements on self-perceived needs across time were striking: topics dealing with oral and written communication, leadership, and principles of management were consistently viewed as among the most needed, while those dealing with Army and Navy doctrine, international relations, and warfare were perceived as least needed.

A comparison of the perception of benefits from the various methods of PME between the two most recent studies showed few differences. Both studies found the perception of benefits from SOS by correspondence to be small, from residence programs to be greather than for correspondence or seminar programs, and from USAF Academy PME to be of greater benefit than ROTC or OTS-OCS PME.

CONCLUSIONS

Data from this project provide a good view of what leadership, management, and communicative tasks officers perform; the relative difficulty of tasks in the judgement of Air Force officers; and rank order listings of curriculum topics which officers perceive they need in their jobs and their careers. Based on the data, several conclusions about Air Force officer PME may be drawn.

First, there are differences in the leadership, management, and communicative task performance of USAF officers across career fields and ranks. Further, there is a lack of agreement among officers as to what PME topics are required in their jobs and careers.

Second, there is a progression in terms of the number and complexity of tasks performed by officers as they advance in rank. Based on this, there is sufficient rationale for the continuation of a multiphased PME program, although it is doubtful a program designed purely on differences in rank can be of equal value to all officers.

Third, the success of a PME program must be judged on its usefulness to the most people. While concentrating on the PME-related tasks most officers at any given phase perform, PME must also educate officers as to the need and importance of a broader understanding of military-related topics to their roles as professionals.

NOTES

1. This paper is based on Professional Military Education -- Officer, AFPT 40-XXX-522 (USAF Occupational Measurement Center, October 1984). Copies of the report are available upon request from USAFOMC/OMY, Randolph AFB, Texas 78150-5000.

2. The return rates for these inventories was as follows: Task List - 2,016 usable/3,638 administered: 55 percent return rate; Task Difficulty - 312/584: Education Emphasis - 316/598: 53 percent: Need-in-Job - 4,100/6,742: 61 percent; and Need-in-Career - 3,863/6,384: 61 percent.

3. The career fields, in descending order of the average number of LMC tasks performed (and the number) were: logistics (146); legal (141); security police (140); personnel resources (139); chaplains (132); comptroller (119); scientific and engineering (102); cartography, geodesy, and intelligence (102); and operations (95).

4. A series of related computer programs, called the Comprehensive Occupational Data Analysis Programs (CODAP), was applied to the data to assist in the analysis. Reliability of individual raters was .30 and reliability of raters as a group was .99, both of which were well within the range of reliabilities accepted in USAF occupational research.

5. Correlations of rankings of need-in-job by rank ranged from a high of .99 between majors and lieutenant colonels to a low of .91 between lieutenants and colonels. Lowest correlations were between rankings by direct-commission officers and academy graduates (.75) and between cartography-geodesy intelligence specialties and other utilization fields. Correlations of rankings of need-in-career by ranks, commissioning sources, and utilization fields were slightly lower, but comparable.

[illegible] Time Measurement of Speed of Information [illegible] Intelligence

[illegible] Saccuzzo[1] and Gerald E. Larson and Bernard Rimland[2]
San Diego State University | Navy Personnel Research and Development Center, San Diego

Abstract

[illegible] auditory reaction time and microcomputer generated measures of speed of information processing were administered to [illegible] college students between [illegible] years of age. In addition, each subject was given a battery of tests designed to provide independent evaluations of right and left hemispheric functioning. Criterion measures included a verbal (Vocabulary) and nonverbal (Block Design) measure of intelligence. Results revealed a general processing speed factor in addition to task specific sources of variability. Morover, the processing speed tasks loaded on the same second order factor as did the traditional measures of intelligence and aptitude. The findings support the theoretical view that processing speed may be a general factor in individual differences in performance on complex intellectual tasks. An important objective for future work in this area is to separate and evaluate the common and specific sources of variability on processing speed tasks, such as those used in this study, which contain little or no intellectual content and involve little or no complex problem solving.

[1] Reprint requests to the Psychology Clinic, San Diego State University, San Diego, California 92182. Funding for this work was provided by a faculty fellowship from the American Society for Engineering Education and by Future Technologies for Manpower and Personnel (PE62763, RF63-521-806).* The authors wish to express their appreciation to Steven H. Paul for his assistance in the statistical analyses.

[2] Now Retired

* The views expressed in this paper are those of the authors, are not official, and do not necessarily reflect the views of the Navy Department.

[illegible] of intelligence tests that [illegible] reached an asymptote more than 60 years ago in predicting [illegible] performance. Consequently, there has been a search for new approaches to the measurement of individual differences in ability and intelligence. One such approach, referred to as chronometric measurement, attempts to evaluate an individual's speed of information processing. The potential of chronometric measurement in military settings, however, is unclear due to a limited data base and seemingly conflicting evidence. Though the preponderance of studies have revealed a small but significant correlation between processing speed tasks and those involving complex cognitive behavior, the meaning of such findings is a matter of controversy (see Hunt, 1980; [illegible]).

In the present study, we obtained visual, auditory, and reaction time measures of processing speed in the same subjects. We also explored the feasibility of using microcomputer controlled tasks to evaluate processing speed, as suggested by [illegible], Rimland, and Bryson (1977). Finally, we attempted to probe more deeply into the meaning of performance on a processing speed task in terms of how it might relate to the functioning of the right and left cerebral hemispheres.

Method

The subjects were 96 San Diego State University students from an introductory course in psychology. All were between 18 and 22 years of age. Fifty-three were male, forty-three female. Eighty were Caucasian; the remaining 16 were black, Hispanic, and Asian. The sample was representative of the total population of San Diego State students in the 18-22 age range.

Each subject was tested on a set visual, auditory, and reaction time processing speed tasks. Subjects were also given the Vocabulary and Block Design subtests of the Wechsler Adult Intelligence Scale, Revised (WAIS-R) and the Cognitive Laterality Battery.

Two types of visual tasks were used, with the order of presentation alternating between subjects. Those of the first type involved tachistoscopic presentations in a paradigm that followed the general procedures used by Saccuzzo et al. (1979). A microcomputer presented battery of five tests for estimating visual processing speed, as described by Larson and Rimland (1984), provided the second set of experimental tasks. Each microcomputer task was presented twice to each subject, once via a TRS-80 and once via an Apple [illegible] in a counterbalanced design. Auditory speed of processing was measured by the Repetition Test of Tallal and Piercy (1973). A reaction time task, as described by Larson and Rimland (1984), was presented on the TRS-80 computer and keyboard. Each subject's median reaction time (based on [illegible] trials) was used as the index for the one (RT1), three (RT2), and five (RT3) choice tasks. The Cognitive Laterality Battery (CLB), developed by Gordon (1983) to evaluate individual differences in hemispheric asymmetries, was administered to subjects in small groups (n = 16 subjects) after all processing tasks had been completed.

Results

Table 1 presents the results of a Schmid-Leiman hierarchical factor analysis in which all factors are orthogonalized and first-order factors are residualized (i.e., their correlated variance is absorbed into the second-order general factor). As the analysis shows, the processing speed tasks and the criterion variables loaded together on a second order factor, which was labeled General Mental Speed. Four first order factors, Reaction Time, Auditory Speed, Psychometric Intelligence, and Visual Speed, also emerged.

Table [illegible]

[illegible] Analysis of Criterion Variables and Measures of Processing Speed

[illegible]	Factor 1 [illegible] General Mental Speed	Factor 2 Reaction Time	Factor 3 Auditory Speed	Factor 4 Psychometric Intelligence	Factor 5 Visual Speed
AUDLO	[illegible]	[illegible]	64	02	01
AUDST	5[illegible]	- [illegible]	67	02	01
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	32
[illegible]	[illegible]	[illegible]	[illegible]	- 11	38
TRAV	[illegible]	09	- [illegible]	11	33
APAV	1[illegible]	08	09	- 03	43
RT1	[illegible]	61	- 02	- 03	03
RT2	61	65	08	00	- 06
RT3	5[illegible]	61	- 06	03	03
HSGPA	30	- 01	- 07	56	- 07
FRGPA	40	- 15	12	44	06
SAT	47	08	06	59	- 11
VOCAB	[illegible]	- 12	- 07	60	02
BD	50	19	- 05	37	10
Total Variance	1%	9%	7%	10%	4%

* Salient factor loadings underlined. Correlations in the original matrix were reflected so that good performance has been positively correlated with all other variables. AUDLO and AUDST = Auditory Processing Speed for long and short interstimulus intervals respectively. ISIc1 and ISIc2 = The Critical Interstimulus Interval, trials one and two, respectively (a tachistoscopically determined nonverbal measure of speed of information processing). TRAV and APAV = TRS-80 and Apple II microcomputer derived scores for speed of visual processing. RT1, RT2, and RT3 = One, Three, and Five choice reaction times. HSGPA and FRGPA = High School and Freshman Grade Point Average. SAT = Scholastic Aptitude Test (total score). VOCAB and BD = Scores on the Vocabulary and Block Design Subtests of the Wechsler Adult Intelligence Scale, Revised.

Table [illegible]

Correlations and First Unrotated Factor for Measures of Processing Speed and <u>Hemispheric Functioning</u>+

	Left Hemisphere	Right Hemisphere	Factor Loadings ++
AUDLG	[illegible]**	04	<u>66</u>
AUDST	[illegible]1*	02	<u>54</u>
[illegible]	[illegible]	[illegible]*	03
[illegible]	[illegible]	[illegible]	21
TRAV	04	22*	<u>45</u>
APAV	05	23*	<u>36</u>
RT1	[illegible]	24**	<u>70</u>
RT2	15	07	<u>77</u>
RT3	14	08	<u>65</u>
Factor Loadings ++	<u>[illegible]6</u>	<u>30</u>	
Mean	[illegible]	034	
SD	56	52	

+ Correlations have been reflected

* p < .05

** p < .01

++ Factor loadings for first unrotated factor matrix for Principal Components factor analysis (total variance accounted for equals 25.4 percent). Salient loadings are underlined.

Right and Left hemisphere functioning was measured by the Gordon Cognitive Laterality Battery. AUDLG and AUDST = Auditory Processing Speed for long and short interstimulus intervals, respectively. ISIc1 and ISIc2 = The Critical Interstimulus Interval, trials one and two, respectively (a tachistoscopically determined nonverbal measure of speed of information processing). TRAV and APAV = TRS-80 and Apple-II microcomputer derived scores for speed of visual processing. RT1, RT2, and RT3 = One, three, and five choice reaction times.

Table [illegible] presents the correlations and first unrotated factor for the measures of processing speed and those for left and right hemispheric functioning. Examination of the table reveals a different pattern of correlations between the processing speed measures and the measures of hemispheric functioning. Specifically, auditory processing was related more strongly to the left hemisphere, visual processing to the right.

Discussion

A set of processing speed tasks loaded on the same second order factor, derived from a hierarchical analysis, as did a set of traditional measures of intelligence and aptitude. This finding reveals that performance on complex cognitive tasks such as Block Design and the SAT is related to performance on tasks that have little or no knowledge content and require no complex problem solving strategy. The findings thus support the theoretical view that processing speed may be a source of individual differences in performance on complex intellectual tasks.

The data further reveal two major components to the variability in measures of speed of processing: a general component that accounts for roughly half the variance and task-specific components that account for the other half. These task specific components account for a fairly substantial share of the variability and cannot be ignored when considering the relationship between measures of processing speed and measures involving complex cognitive processing. The task specific variance, moreover, may be of intrinsic interest in that any given task may be related to important processes. For instance, the pattern of correlations revealed that the visual tasks tended to be related to the right hemisphere related tasks on the Gordon Battery, the auditory tasks, by contrast, were related to the Gordon Battery left hemisphere tasks. Hypothetically, it might be feasible to construct a set of processing speed tasks both to measure general intelligence, through obtaining a composite score in which individual differences due to task-specific abilities average out, and to measure more specific group factors that are relatively independent of each other and can help predict job performance. Furture research in this area should be directed toward separating the common variance from the task specific and to determine more precisely what each of these components measures. We conclude that:

(1) Processing speed tasks, which contain little or no intellectual content and involve little or no complex problem solving skills, share common variance with conventional psychometric tests that do involve complex reasoning and problem solving skills.

(2) Scores on processing speed tasks are multi-dimensional. Though a general processing speed factor emerges from a hierarchical analysis of a set of visual, auditory, and reaction time tasks, more specific factors emerge as well. Consequently, it appears to be an oversimplification to ask whether any particular processing speed task is related to intelligence. Rather, more specific questions are needed.

(3) Measurement of Spearman's "g" factor of mental ability by means of speed of processing tasks will most likely depend on using a battery of such tasks having sufficient diversity in specific task features to permit the averaging out of task specific variance so that the composite score will predominately reflect the general ability factor that is common to all of the tasks.

References

Cory, C. H., Rimland, B., & Bryson, R. H. (1977). Using computerized tests to measure new dimensions of abilities: An exploratory study. Applied Psychological Measurement, 1, 101-117.

Gordon, H. W. (1983). Cognitive Laterality Battery. Western Psychiatric Institute and Clinic, University of Pittsburg, School of Medicine.

Hunt, E. (1980). Intelligence as an information-processing concept. British Journal of Psychology, 71, 449-474.

Larson, G. E., & Rimland, B. (1984). Cognitive speed and performance in basic electricity and electronics (BE & E) school. NPRDC TR85-3, October.

Saccuzzo, D. P., Kerr, M., Marcus, A., & Brown, R. (1979). Visual information-processing in mental retardation. Journal of Abnormal Psychology, 88, 341-345.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. Psychomentrika, 22, 53-61.

Tallal, P., & Piercy, M. (1973). The Repetition Test. Neuropsychologia, 11, 389-398.

Vernon, P. A. (1981). Reaction time and intelligence in the mentally retarded. Intelligence, 5, 345-355.

Wechsler, D. (1981). Manual for Wechsler Adult Intelligence Scale-Revised. San Diego: Harcourt, Brace & Janovich.

REPRESENTATION AND MODIFICATION OF HUMAN PERFORMANCE FACTORS IN TARGET SELECTION

Dwight J. Goehring
U.S. Army Research Institute Field Unit
Presidio of Monterey, California

I. Abstraction is Fundamental in Science

Abstracting or modelling of phenomena is a primary method of science for increasing knowledge. This approach is vital where the phenomenon under investigation cannot be subjected to direct scientific manipulation, such as astronomical processes or force-on-force combat. However, the quality of conclusions Obtained from a particular abstraction, or model, is critically dependent upon how completely and accurately the essential components and their interrelationships are represented.

Models which simulate the process of combat have long been important to military planners in the formulation of policy, strategy, and tactics (Brewer and Shubik, 1979). The use of modern computers has enabled the representation of an unprecedented level of detail in combat modelling. For example, extant Army combat simulation models consist of hundreds of thousands of lines of high-level-language computer code. This code represents the combat process in extreme detail, down to the field of vision and magnification specifications of the specific optics selected by every relevant weapon system at a given point in time.

Despite such great detail incorporated in many models, larger questions remain about the overall adequacy of the representation of the combat process. Characteristics of optical or weapon systems can be precisely determined using experimental methodologies. Data from such methods serve to determine with high confidence the values of many of the very large number of parameters present in modern combat simulation models. One class of phenomena which is not, however, so easily characterized is human behavior. As human performance factors (HPF) are an integral part of virtually all systems comprising the combat process, the nature of the representation of these factors in combat models has direct implications for the validity of any results.

II. Human Performance Factors in Models

There are several issues pertaining to the representation of HPF in the modelling of combat. Some of these can be illustrated by taking a simple behavior such as the likelihood that a single soldier will detect a particular type of vehicle under some situation specifying distance, atmospheric conditions, lighting, the soldier's physical and psychological states, and so forth.

One issue concerns the accuracy of estimating the likelihood of detection. With conventional experimentation and statistical methodology, an estimate of known certainty can be obtained. Similar methods can yield information on the variability between individual soldiers in detection likelihood. However, a serious difficulty arises with this approach when the detection parameters are sought for different sets of situation contexts. Human performance can vary not only as a function of each of the myriad of single factors but also based upon any possible interaction of the factors. Both combinatorial explosion and experimental impracticality become problems immediately. Thus, to rigorously establish model parameters of relevant contexts for HPF of even a simple task is simply not possible. Therefore (1) HPF must be represented in combat simulation models in simplified form, and (2) at some point assumptions about such aspects as the value limits on HPF variables and the nature of interactions among variables become necessary.

When examples of representations of HPF in combat simulation models are considered, both the form and detail are found to vary widely. Miller and Bonder (1982) conducted an investigation of the HPF treatment in nine combat simulation models. Fifteen combat processes (e.g. communications, maneuver control, and mobility, countermobility, and survivability) were identified which contained 110 "human factor process phenomena" (e.g. decisions to request fire support, route selection of forces moving on the ground, increase in exposure due to encountering a minefield or obstacles). They observed that a phenomenon could be represented in a model in one of five basic forms: (1) assuming that the phenomenon is unvarying and always correctly performed, (2) representing the factor by a human player in real time while the model is running, (3) directly inputting the effect of the factor prior to model execution, (4) representing the phenomenon explicitly in computer code within the model, and (5) representing the phenomenon implicitly, that is, inclusion only as a logical necessity of some subsuming, larger-scale process present in the model. Among the nine models, the number of explicit or user input representations of human factor process phenomena ranged from 14 to 77. Large-scale models, such as theater-level simulations typically are based upon the results of lower-level models, necessarily implicitly incorporating whatever representations of and assumptions about HPF are contained in the models upon which they are based.

A second issue regarding the representation of HPF in combat simulation models centers upon the extent to which the assumptions made are empirically based. While detailed elucidation by experimentation is generally impossible, other sources of information exist, each with unique shortcomings. For example, Dupuy (1985) and others have proposed greater use of historical data. Of course, any historical account is to some extent both subjective and highly dependent upon specific circumstances. Another source of information is the judgments of experts in the domain in question. Such persons may include combat veterans and behavioral scientists. Problems are that experts frequently disagree and even achievement of consensus does not guarantee accuracy. Another source of relevant information is the data collected from the battalion-size training exercises conducted at the National Training Center (NTC) in recent years (Fobes, 1984). This data, however, is pertinent to only a subset of the HPF of interest and the analysis of data does present some challenges (Whitmarsh, 1985).

A final issue concerning HPF representation is establishing their relative importance in determining model outcomes. Perhaps it is too obvious to state that the HPF which most impact outcomes demand greater attention in terms of both the accuracy and completeness of their representation in combat models.

III. Representation of Target Selection Behavior

The Training Performance Analysis Tool (TPAT) was developed as a part of a larger research program (Banks, 1985) to assist in the analysis and interpretation of data contained in the NTC database. As training resources are always constrained, they require allocation on the basis of expected return. The current work examines how HPF are modelled in TPAT with the purpose of identifying where training intervention is most likely to have the greatest impact. Typically, computer-based combat simulation models are of sufficient inherent complexity that determination of effects of variations in a particular component cannot be assessed by examination. Instead, the model, with changes incorporated, must be executed and the results evaluated. The scope of the current work is limited to the HPF area of target selection, which is relatively richly represented in TPAT.

The design philosophy of TPAT (Weaver and Briesemer, In press) was based upon three principles. First, parameters in the program are primarily based on empirical results obtained from real-time casualty assessment (RTCA) experiments (Clark et al, 1974) conducted by the Combat Developments Experimentation Command (CDEC). Second, the level of detail in the simulation is restricted to only what is necessary to accurately simulate outcomes and the outcomes are limited to those both observable and recorded in the NTC database. Third, Monte Carlo simulation techniques are employed to enable the study of outcome variability. TPAT is simple enough to facilitate understanding and manipulation of HPF without undue time and expense. Further, it may be possible to eventually validate the findings of this investigation using the NTC database.

As implemented, TPAT incorporates three behavioral tendencies of defending forces--friendly forces in the current scenario--which both conflict with Army doctrine and run counter to common sense. They are, however, empirically based (Clark et al, 1974). First, defending tanks and antitank tube-launched optically-tracked wire-guided missiles (TOWS) lack target preferences concerning target vehicle types. Doctrine states that higher threat targets should be engaged first. In the scenario this implies that the long-range opposing force antitank weapons (SAGGERS) and tanks (KTANKS) should be preferred to unarmed armored personnel carriers (APCS) as targets. Defenders presumed interest in self-preservation would also suggest that the APCS, at least at moderate to long ranges, should be targets of secondary importance because they are of relatively little threat.

A second empirically-based behavioral tendency concerning target selection incorporated into TPAT is that newly visible targets are preferred targets. Since new targets would, in general, be less likely to have acquired a target, be more distant, and, therefore, be less threatening than previously visible targets, the same doctrinal principle applies and again is violated. The behavior also appears not to be in the best interests of self-survival.

Finally, TPAT includes the behavioral tendency of defending tanks to prefer previously engaged targets. This preference leads to previously killed targets being frequently engaged. This target perseveration also violates the doctrinal principle of firing at highest-threat targets first. Also, as with the other behavior tendencies, it promotes neither the survival interests of the individual players nor of the defenders collectively.

These three target selection phenomena are included in TPAT because they describe actual behavior of defenders in RTCA experiments. Whether the phenomena exist in NTC training exercises must await the collection and analysis of appropriate data. Rather than speculate about possible psychological bases of these behaviors (e.g. information overload (Miller, 1956), preference for novelty, need for closure, "shooting gallery mentality," and so forth), it is more important to understand that they are empirically based and to realize that each could potentially be modified to some degree through training. TPAT itself can be used to provide estimates of the effects alterations in these three target selection behaviors have upon simulated battle outcomes. The results should reveal their relative importance on outcomes and, therefore, suggest where the largest training payoff lies.

IV. Modification of Modelled Target Selection Behaviors

TPAT was run employing a digitized portion of NTC terrain with an approximate battalion-sized force (18 tanks and 18 TOWS) in defense and an approximate regiment-size attacking force (30 KTANKS, 30 SAGGERS, and 30 APCS) using rapid approach tactics. To assess the effects of the three behavioral assumptions upon model output, each of the behaviors was modified. The preference for newly visible targets and target perseveration where both reduced by 100 percent. Preference for target type was manipulated by reducing desirability of APCS to zero when there are other target candidates and when the range of APCS was greater than 500 meters. Target selection assumptions were modified for defending tanks and TOWS.

The effects of modifying each of the three target selection tendencies of the defending force are summarized in Table 1. The model scenario was executed fifty times for each target selection modification. Casualty ratios are shown for both the offensive and defensive forces broken down into inflicted and sustained components. The values for casualties can be interpreted as the mean number of vehicle kills per vehicle per trial. The casualties sustained values can be interpreted as the mean probability across trials and across vehicle types of a vehicle being killed.

The results in Table 1 indicate that eliminating either the perseveration on last target or the preference for newly-visible targets has virtually no effect on casualty values in comparison to outcomes from the original model. However, the conclusion is <u>not</u> warranted based upon the the evidence at hand that they have no effect upon performance outcome. Using other scenarios or alternative terrain may reveal they too affect performance. At present, it is appropriate to conclude only that these behaviors have no impact within the

Table 1.

Effects of Changes in Target Selection Processes on Combat Outcomes

MEASURE OF PERFORMANCE	ORIGINAL MODEL	NO PERSEVERATION ON LAST TARGET	NO PREFERENCE FOR NEWLY-VISIBLE TARGET	REDUCED PREFERENCE OF APCs AS TARGETS
Offensive Casualty Ratio (inflicted sustained)	2 .97	.2) .97	[illegible] .99	[illegible] .98
Defensive Casualty Ratio (inflicted sustained)	2.43 .5	2.48 .5	.48 .49	2.45 .34

[1] The standard error of estimate across trials is less than .022 for all numbers shown.

current scenario conte.t. B, contrast, reducing the preference for APCS as targets produces strikingly different performance. Because defenders concentrate firepower on KTANKS and SAGGERS, these high-threat attackers are attrited at a relatively higher rate. This leads to a reduction in the casualties the attackers are able to inflict. In turn, this produces the result that the likelihood of defender being killed falls from .50 to .34. The values represent a decrease in per vehicle risk of approximately a third.

The finding that reducing the attractiveness for APCS as targets produces fewer defender losses provides unambiguous support for the doctrinal principle that targets of greater threat should be engaged first. Also of interest is the supposition that target selection behavior is probably amenable to modification through appropriately designed and conducted training. While target discriminability probably has inherent limits, especially at e:treme range, it is plausible to expect some improvement with training.

A series of TPAT runs was conducted to explore how much effect upon outcomes might be expected from varying degrees of reduction in the preference for unarmed APCS as targets. In addition to the preference level of APCS in the original model and the previous 100 percent reduction, their attractiveness was reduced by 25, 50 and 75 percent. Fifty runs under each condition were performed. Figure 1 presents the results displayed as mean probabilities of defensive vehicles becoming casualties. The specific values are less meaningful than the overall pattern of the relationship. Finding a nonlinear relationship underscores the need for actually making model changes and executing the model to determine effects. In addition, the direction of the nonlinearity observed suggests that small amounts of reduction in the preference of APCS as targets yield little reduction in risk--a 25 percent reduction from the original model produces no risk reduction. The relationship suggests that only when there is 50 percent or greater reduction in preference are substantial benefits realized. This finding suggests that while relevant training may reduce risks to defenders, incomplete or ineffective training may produce no benefit.

V. Conclusions

Clearly, assumptions made about HPF can affect the outcome of simulated combat. This finding has training as well as analysis implications. Investigation of HPF effects in combat models can be important to the identification of training needs, to the planning of training, and to the development of training criteria goals. Model developers may be aware of the

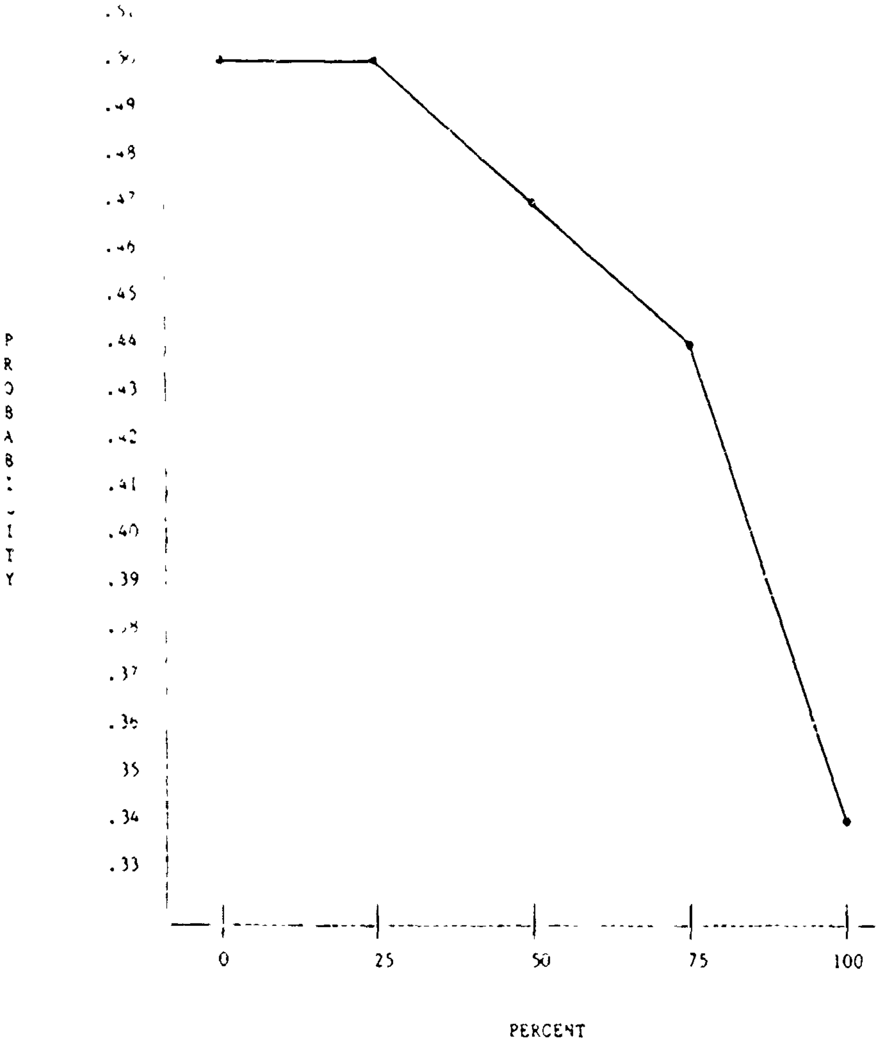


Figure 1. Effect of percent reduction in defender target preference for APCs upon mean probability of a defending vehicle becoming a casualty.

HPF assumptions and simplifications that have been made and the consequential limitations upon model outcomes. It is also vitally important for policy decision makers, the model users, to know of these assumptions and caveats. Therefore, the value of models is enhanced when these considerations are explicitly documented.

Further systematic research on the HPF representation in combat simulation models is necessary. Several of the specific needs are (1) development of strategies for gathering relevant empirical data, (2) determination of most important HPF, and (3) identification of methods for better incorporating the effects of HPF into combat simulation model findings.

VI. References

Banks, J.H. (1985). An overview of ARI's research program on the National Training Center. In the Proceedings of the Military Testing Association, San Diego, CA.

Brewer, G.D. & Shubik, M. (1979). The War Game A Critique of Military Problem Solving. Cambridge, MA Harvard University Press.

Clark, I.B., Marchi, R.P., Sargent, J.D. Vanarsdall, D. & Weaver, W.B. (1974). Final Report, TETAM Extended Analysis, Vol. I. BDM CARAF FR-74-075.

Dupuy, T.N. (1985). Criticisms of combat models cite unreliable results. Army, 35(3), 16-18.

Fobes, J.L. (1984). National Training Center Data Handbook. ARI Research Product 84-17, Alexandria, VA U.S. Army Research Institute.

Miller, G.A. (1956). The magical number seven, plus or minus two Some limits on our capacity for processing information. Psychological Review, 63, 81-97.

Miller, G.J. & Border, S. (1982). Human Factors Representations for Combat Models. ARI Technical Report 571, Alexandria, VA U.S. Army Research Institute.

Weaver, W.B. & Griesemer, H.A. (in press). NTC Training Performance Analysis Tool. Research Note, Alexandria, VA U.S. Army Research Institute.

Whitmarsh, P.J. (1985). Types and quality of National Training Center data. In the Proceedings of the Military Testing Association, San Diego, CA.

QUALITY OF CANADIAN FORCES OTHER-RANK RECRUITS: TRENDS 1975-1983

Major K.W.J. Wenek

Canadian Forces Personnel Applied Research Unit
Willowdale, Ontario, Canada

Introduction

Depending on whether a country chooses man its armed forces on the basis of voluntary or obligatory service, the strategic forecasting of human-resource supply and the planning of human-resource needs will require different emphases. Conscript forces, which are more or less ensured a reasonable supply of recruits and a representative share of the available youth talent, are primarily concerned with personnel screening and allocation functions. All-volunteer forces (AVFs), on the other hand, which are particularly susceptible to the effects of free-market economics on the quantity and quality of the available manpower pool, must devote considerable attention to the business of attracting applicants before practical screening and allocation policies can be implemented. Moreover, because AVFs are dependent on the willingness of young people to serve in the military, special attention must be given to any qualitative biases in the composition of the target recruiting pool and recruit intake introduced by the market forces which directly affect enlistment propensity. (For a more comprehensive discussion of these issues, see Cotton, Crook & Pinch, 1978; and Pinch, 1982.)

While there are a number of dimensions on which qualitative comparisons among recruit cohorts can be made, military staffs are typically concerned with the trainability of recruits, that is, their potential to acquire those job skills which are a necessary, but not a sufficient, condition for operational effectiveness. In the Canadian Forces (CF), the general indices of quality and trainability most commonly invoked are entry scores on the General Classification (GC) test (a group-administered intelligence test) and educational attainment. Although the relationship between these indices is not entirely orthogonal, GC test scores can be essentially interpreted as measures of learning potential, while educational attainment levels provide measures of acquired learning. Trends in the quality of Other Rank (OR) recruits over the 1975-83 period will be examined in these terms.

Trends in OR Recruit Quality: 1975-1983

As illustrated in Table 1, the average GC-test scores of OR recruits and the distribution of these scores have been remarkably consistent over the 1975-1983 period. Furthermore, because the GC test is a standardized test, it seems reasonable to conclude that recruits

The views and opinions expressed in this paper are those of the author and not necessarily those of the Department of National Defence.

enrolled in 1983 were, in an absolute sense, neither more nor less able or trainable than those enrolled in previous years.

Table 1

Raw GC Score Means and Standard Deviations for Anglo and Franco CF Enrollees (1975-83)

Intake Year	Anglo Enrollees		Franco Enrollees	
	Mean	SD	Mean	SD
1975	42	9.8	37	8.9
1976	44	9.5	40	8.6
1977	44	9.3	40	8.6
1978	45	9.4	40	8.3
1979	44	9.2	40	8.4
1980	44	9.4	39	8.3
1981	43	9.3	38	8.4
1982	43	9.4	39	8.5
1983	44	9.2	40	8.3

A different pattern emerges when educational attainment data are considered. As shown in Table 2 (Tivendell & Gaudet, 1985), the distribution of educational attainment levels among CF enrollees over the 1975-83 period has shifted towards the better-quality end of the spectrum, with higher proportions of enrollees possessing some, or complete, post-secondary and university education and lower proportions of enrollees

Table 2

Percentages of CF Enrollees at Various Education Levels Compared to Population Percentages

Year	Education Level							
	Some/Complete Elementary		Some/Complete Secondary		Some/Complete Post Secondary		Some/Complete University	
	CF	Cdn Pop	CF	Cdn Pop	CF	Cdn Pop	CF	Cdn Pop
1975	7.0%	5.5%	85.9%	60.3%	3.3%	25.2%	3.7%	3.5%
1976	4.9%		87.4%		4.2%		3.5%	
1977	5.0%		85.9%		4.3%		3.9%	
1978	3.7%		87.7%		4.1%		4.5%	
1979	3.3%	4.9%	87.7%	67.2%	5.0%	19.7%	4.1%	3.4%
1980	4.1%		87.8%		4.7%		3.5%	
1981	4.3%		88.1%		4.4%		3.2%	
1982	3.3		83.1%		7.7%		5.8%	
1983	2.6%	4.1%	81.7%	64.3%	9.3%	[illegible]	6.4%	3.1%

with some, or complete, elementary and secondary education evident. In comparison to benchmark figures for the Canadian population, CF enrolments from the elementary-education group have shifted over time from over-representation to the more favourable position of under-representation. Similarly, in the university-education group, the CF has maintained a very favourable recruiting position, and has, in fact, improved the extent of over-representation. It is in the middle groups, however, which are the sources of the bulk of its recruits, that the CF has had less success, continuing to be over-represented at the secondary education level and substantially under-represented at the post-secondary level.

Data obtained from sources (Tierney, 1981; Park, 1982) which deal exclusively with the educational attainment, and which break out educational attainment figures in terms of program completion and non-completion, provide a more useful, if less complete, picture. While these data confirm the general shift towards higher proportions of better educated recruits and lower proportions of the less educated, it also reveals certain weaknesses in the CF's recruiting posture. Recruit intakes continue to be over-represented by non-high school graduates and under-represented by graduates of post-secondary schools and universities. These trends are reason for concern in view of the training demands imposed by the proliferation of complex technology in all areas of military functioning and in light of the strong relationship between high school completion and early-career survival (Sinaiko & Scheflen, 1982).

Table 3

Percentages of CF OR Enrollees at Various Education Levels
Compared to Population Percentages

Year	Education Level					
	Non-High School Grad		High School Grad		Post Sec & Univ	
	CF	Cdn Pop	CF	Cdn Pop	CF	Cdn Pop
1975	8[illegible]		17%		2%	
1976		44%		20%		36%
1980	[illegible]		2[illegible]%		11%	
1981		42		21%		37%
1982	48		-----------52%-----------			
1986*		39%		19%		42%

* Projected

Discussion

The information on trends in educational attainment presented above prompt at least two major questions. Why has the CF been relatively unsuccessful in attracting a representative share of technical/community college students? Do the apparent general gains in the educational quality of recruits represent real and sustainable gains.

surveys of applicants to the CF ([illegible], 198[illegible]) and of high school students planning to pursue post-secondary education (Urban Dimensions Group, 1982) have indicated that interest in the CF as a career option generally decreases with higher levels of education and that, among technically oriented students in particular, the CF is relatively unattractive as a training or employment opportunity. For those high school students who have the potential and desire to acquire technical skills after graduation, the CF has very little to offer in the way of accelerated training programs or incentive programs, and it is in this sense that the CF has been displaced as a major technical training institution by the numerous technical institutes and community colleges which sprang into being during the 1960s and 1970s (Pinch, 1982; McMenemy, 1984). With respect to those who already hold post-secondary qualifications, the CF does not have a lateral-entry program for skilled applicants nor does it yet have a career structure which specifically rewards technical-skill specialization (although a program to address this latter area is under development). Overall, therefore, it must be concluded that, unless the CF's technical training programs, career structures, and benefit packages can be made more attractive to aspiring or qualified technicians and technologists, and can be made more competitive in comparison to civilian educational and career opportunities, this important source of personnel is not likely to be adequately tapped for some time to come.

The question of apparent versus real gains in educational quality among CF recruits is somewhat more difficult to answer directly. As a point of departure though, it should be noted that the pass/fail criteria used in schools and which, to a large extent, affect the rate and extent of school advancement, can vary considerably over time, particularly when institutional survival is threatened by declining enrolments. This elasticity of standards and the resulting difficulties in making absolute comparisons in achievement should also be viewed in the light of recent public controversy in the United States and Canada over the declining quality of education. Research on trends in scholastic achievement scores (standardized, hence comparable, measures) strongly indicates a decline in the academic competence of youth (Waters & Laurence, 1982). What these observations collectively suggest is that neither the absolute nor relative quality of education in recent years is what it used to be.

With respect to the sustainability of, or improvement upon, the present state of recruit educational quality, future developments will depend on a number of factors, many of which have contributed to the past few years of good fortune. Since the onset of a national economic recession in 1981, the national unemployment rates in the work force have ranged between 10% and 13%, with the rates running somewhat higher, at 16% to 20%, for the youth population (Statistics Canada, 1985). Consequently, the effects of the bleak demographic prophecies of the late 1970s have been temporarily postponed. Over the same time period, labour-market conditions have had a substantial impact on the voluntary attrition of CF other ranks. As reported by Jenkins and Lyon (1984), attrition fell off rapidly from 1979 to 1983 as unemployment rose (r=-.74), and annual recruiting quotas fell accordingly. The combined effects of an increasing supply, brought about by high unemployment in the 15-24 year-old segment of the population, and of a decreasing demand, brought about by declining attrition, can best be seen in the applicant and enrolment figures for this period and the corresponding dramatic improvement in CF selection ratios (Table 4).

Table 4

Total Number of CF Other Rank Applicants and Enrollees for Fiscal Year 1979/80 to 1983/84

Fiscal Year	Applicants	Enrollees	Selection Ratio
1979/80	28,468	10,316	1:2.8
1980/81	35,243	11,365	1:3.0
1981/82	39,626	12,261	1:3.2
1982/83	40,148	6,674	1:6.0
1983/84	29,738	4,038	1:7.4

It would be extremely surprising if, under these conditions, recruit quality were anything but high. Yet because current CF recruit quality is largely an artifact of a recent national economic downturn, it is sobering to think what might happen if there is an upswing in the economy. On the supply side, declining unemployment in the youth population and the intractable effects of demographics will rapidly shrink the eligible personnel pool. On the demand side, a surge in voluntary attrition, as external employment opportunities materialize, will probably drive up recruiting quotas to or beyond pre-recession levels. The first impact would be on quality. Numbers could also be a problem.

Without forward-looking recruiting strategies and personnel policies, particularly those designed and preplaced with a view to coping with a manpower-shortage contingency, the CF runs a real risk of suffering a strategic self-inflicted wound. The policy areas which require urgent review in order to forestall this eventuality include the following:

a. redefining the eligible manpower pool, which would minimally entail the removal of age barriers, the expansion of women's roles, and the development of lateral-entry programs for skilled applicants,

b. improving the fit between military training and civilian education, which could mean increased reliance on "off-the-shelf" skills and the establishment of CF-sponsored programs at technical colleges;

c. developing the recruitment potential of the military-socialization infrastructure, which basically means providing more support to cadet and Reserve organizations; and, finally,

4. *exploring the possibility of a national service scheme*, which could vary in application from the institution of a national public-service draft, which includes military service as an option, to a comprehensive incentive-and-benefits program for military service.

REFERENCES

Cotton, C.A., Crook, R.K., & Pinch, F.C. (1978). Canada's professional military: the limits of civilianization. *Armed Forces and Society*, *4*, 365-90.

McMenemy, J.P. (1984, April). *Recruiting, selecting, and training the next generation of technicians.* Paper presented at the NATO DRG Panel VIII Workshop on Applications of Systems Ergonomics to Weapons Systems Development, Shrivenham, England.

Eagles, L.C., & Lyon, C.D.F. (1984). *Canadian Forces attrition/retention study* (Working Paper 84-3). Willowdale, Ont: Canadian Forces Personnel Applied Research Unit.

Park, R.E. (1982). *Trends in recruit quality: An analysis of enrollee profiles across recruit zones from July 1980 to June 1982* (Technical Note 5/82). Willowdale, Ont: Canadian Forces Personnel Applied Research Unit.

Pinch, F.C. (1982). Military manpower and social change: assessing the institutional fit. *Armed Forces and Society*, *8*, 575-600.

Sinaiko, H.W., & Scheflen, K.C. (1982). *Correlates of first term attrition: A comparison across TTCP nations.* TTCP Technical Panel UTP-3.

Statistics Canada (1985). *The labour force: July 1985.* Ottawa: Supply and Services Canada.

Tierney, M.C. (1981). *The applicant survey 1979/80: A preliminary analysis of national and regional effects* (Working Paper 81-4). Willowdale, Ont: Canadian Forces Personnel Applied Research Unit.

Tivendell, J., & Gaudet, J.W. (1985). *Socio-demographic trends and related changes in Canadian society affecting the Canadian Forces personnel supply* (Research Report 85-2). Willowdale, Ont: Canadian Forces Personnel Applied Research Unit.

Urban Dimensions Group Inc. (1982). *Occupational and educational plans of Canadian youth as they impact on ROTP subsidized education opportunities* (Research Report 82-9). Willowdale, Ont: Canadian Forces Personnel Applied Research Unit.

Waters, B.K., & Laurence, J.H. (1982). *A comparison of test score trends: Civilian versus military examinees and recruits (1972-1981).* Offprint.

Role, Importance and Availability of GFAF Reservists

Heinz-J. Ebenrett
Federal Armed Forces Office, Bonn, FedRep Germany
presented by
Hans Kuessner
Federal Armed Forces Office

The Mission

The mission assigned to the German Federal Armed Forces (GFAF) by NATO requires the three Services - Army, Air Force, and Navy - to render an essential and decisive contribution to the collective defense of Central Europe and the Baltic Approaches.

With its 345,000 soldiers on active duty, the Army contributes the largest share to the total strength. The Air Force will grow from almost 111,000 active airmen in peace to about twice that size in war while the Navy will increase its present complement of 39,000 sailors 1.75 times. These strength figures can only be achieved by armed forces that are structured as a conscript army with a high percentage of temporary-career and regular personnel and that are supported by strong reserve personnel.

Reserve Personnel

Since the inception of the GFAF, 5.2 million men have performed military service in the "Bundeswehr". 2.4 million of them are still available for military purposes due to age and level of training.
At present, 762,000 reservists are included in plans for assignments which must be filled by mobilization to achieve the full defensive capability of the armed forces. Of these 762,000 reservists firmly planned for mobilization, 40,000 are officers, 196,000 NCOs and 526,000 privates.
In the next two years, it is intended to raise these numbers by 90,000 in order to be able to perform the task of the Wartime Host Nation Support (WHNS) Program. As a result, wartime strength will grow to 1.34 million service personnel, beginning in 1987.

Use of Reserve Personnel

Army and Air Force are currently providing for 60,000 reservists in the "Standby Readiness" component. In a crisis, the Federal Minister of Defense can recall standby readiness personnel to active duty in order to improve the operational readiness of specific combat units of the Army and the Air force. Following a decision by the Federal Government, the men in the Alert Reserve are recalled in the course of mobilization measures. These men will round out units and agencies of all the Services, assume tasks designed to maintain the operational freedom for the armed forces, and support the allied forces. The balance of the

reservists firmly included in the plans will be available to major units as replacement personnel.
The requirement of the armed forces for reserve personnel can be met in terms of numbers, but there are still some deficiencies as far as quality is concerned, since it has so far been impossible to train all reservists during their periods of active military service for their particular wartime assignments. They must be given additional training during reserve duty training periods.

Reserve Duty Training

Training and extension training of reserve personnel is provided by a system of reserve training periods that is geared to the needs of wartime operational readiness:

Individual reserve duty training periods are designed to provide extension training to the individual reservist. Their normal duration is two to four weeks. Reservists often volunteer for them.
Mobilization exercise of a duration up to 12 days are in the first place intended for the training of those elements which in peacetime exist merely as equipment holding units with no personnel or as cadre-strength units.
Mobilization alert exercises are conducted without warning. They last up to three days, and their purpose is to practice mobilization procedures.

The importance which the armed forces attach to the reserve component will continue to grow in the years to come. The reservist concept is currently being updated in connection with studies under way concerning the structure of the German forces. The goal is not only to continue to meet the numerical requirement of the forces for reserve personnel, but also to improve the quality of training through organizational measures and to ensure that the burdens involved are distributed to all reservists as equitably as possible. Spaces for reserve duty training are to rise to 6,600 by 1986. This will make it possible, for the first time to recall more than 200,000 reservists per annum for reserve duty training periods and to narrow the still existing gaps in their training.

In the nineties, the number of spaces for reserve duty training will gradually be increased to 15,000. These measure will enable up to 400,000 reservists per year to be recalled for reserve duty training periods and to be trained for their wartime assignments. Reserve personnel must then expect to be recalled more frequently and perhaps for shorter periods of time. Such a conscripted program would pose an additional burden both on the individuals concerned and on industry and economy.

Model "Reservist Volunteering by Particular Declaration"

There are doubts that these far reaching plans (mentioned above) can be realized; especially with regard to the readiness of reservists to participate repeatedly on extension training periods.

In order to attract qualified reserve personnel for matters of intensified training, a special model has been developed. This model is called "Reservist Volunteering by Particular Declaration". It consists of the following elements:

1. The reservist formally declares to render repeatedly (additional to obligatory reserve duty training) mobilization exercise and training periods of at least 28 days annually (maybe less when realized) for a minimum of 3 years.

2. The status of the "Volunteering Reservist" will be that of a conscript and his payment shall be fixed according to his civilian income.

3. With respect to dates and terms of exercise the reservist is offered to come to an understanding with his mobilization unit.

4. As for incentives the reservist may expect development and promotion in his individual military career.

The model primarily aims at leadership personnel in mobilization units up to the level of a battalion commander. The "Volunteering Reservist" will generally receive military training in the same mobilization unit. The purpose of this is to enhance unit cohesion among reservists as a way of increasing unit effectiveness. Additionally it is expected that the commitment of the "Volunteering Reservist", i.e. their understanding of and corresponding to common defense duty, will have a distinctively positive influence on fellow citizens.

(Supplementary to the model mentioned above there are reflections about the introduction of a status for part-time soldiers. Preliminary considerations aim at civilian operators, mechanics, technicians etc., who can make it possible to serve as part-time soldiers; for example one day per week or several days a month. That type of "reservist" would be earmarked to move and maintain military equipment and material of those elements, which in peacetime exist merely as equipment holding units with no personnel or as cadre-strength units. Nevertheless, at the present time there is no legal basis for part-time soldiers in Germany and therefore this model has not yet exceeded the status of preliminary considerations. If at all, it may be carried on when the model concerning the "Volunteering Reservists" has been realized.)

Inquiry of the degree of Acceptance

The decision whether the "Volunteering Reservist" model will be realized or not is still pending. Inter alia, it depends on the degree of acceptance among the reservists. In order to get valid data for matters of prognosis and planning, the

Psychological Service of the GFAF has been assigned the mission of determinig this acceptance by means of a representative survey.

Currently a representative sample of 2.500 reservists is being conducted. Results will be given not before the end of the year. Nevertheless, some preliminary trends can be drawn from pretest data. The following preliminary data may be of some interest.

Primarily for matters of testing the suitablility of the questioning instrument a sample of 182 reservists in leadership functions, who happened to render a reserve duty training in August 1985 had been asked about their interest in the "Volunteering Reservists" model. The proportion of answers to the central question show a distinctly high degree of acceptance:

Interest in Participating:(N)

	total	NCO's or cand.	Senior NCO's	Officers
"yes"	74	33	16	25
"uncertain"	22			
"no"	85	63	30	15
	(1 missing)			

74 of the 182 reservists (= 41 %) stated an individual interest, the officers in the sample even by majority.

Reason for Rejection

With respect to those subgroups who showed "no" or "uncertain" interest (N = 85 + 22 = 107) it is obvious that the predominant reasons given for rejection referred to occupational claims or hinderances. Claims of family were of minor but significant importance. Although each third of the uninterested reservists manifested doubts in the purpose and sense of military duty, too, it can be said that the main reasons for rejection are more objective hinderances in the occupational and personal sphere and less negative attitudes or reservations towards the military duty:

Reasons for Rejection(multiple answers)
Subgroups: "no interest" (N=85) and "uncertain" (N=22)

	"primarily	"important	(others)
occupational claims	63	18	26
reservations of the boss	28	23	53
occupational drawbacks	30	15	62
claims of family	31	27	49
reservations of the spouse	21	26	60
doubts in purpose and sense of military duty	20	16	71

Keasons for Volunteering

As for the subgroup of those reservists who showed interest in the model it is noteworthy that the actual source of that interest seems to be a positive attitude towards the military in general. The far overwhelming majority of the interested reservists stated a noticeable rate of contacts to the armed forces, positive experiences during terms as well as the utility of military training while on civilian business, The reasons which may attract them to consider a participation on the "Volunteering Reservist" model are given in the following:

Reasons/Conditions for Volunteering
Subgroup: "interested in . . . " N = 74

	"very important"	"important"	(others)
agreements upon dates	53	18	3
good morale in the unit	36	31	7
suitable duties	36	23	15
military leadership development	34	25	15
promotion in rank	20	27	27
actions near home	25	14	35
financial incentives	15	30	29

The condition to come to an understanding with the mobilization unit upon the dates and terms of training periods is of utmost importance. Nearly all of the interested reservists wanted to be able to plan and determinate duty terms in advance. If that is made sure, they seem to be more attracted by idealistic views and incentives (i.e. morale, duty, leadership) and less by objective advantages (i.e. promotion, payment, short distances).

Summary

Although the data may not be considered representative, they arre suggestive of certain tendencies, which can specifically be verified in the main study. With respect to the far-reaching aims of the "Volunteering Reservist" model it is hoped that both substantial trends we mentioned above may be confirmed; i.e. the sufficiently high degree of acceptance as well as the note that in the first line reservists with idealistic views and motives are attracted. Therefore, their commitment and understanding of duty to render repeatedly voluntary training periods may justly be expected to have a lasting positive influence on defense motivation of draftees and civilians, alike, in their social surroundings.

Retirement Readiness as a Function of Transition Assistance and Trade

Major F.P. Wilson
Canadian Forces: Director of Personnel Selection,
Research and Second Careers

Retirement normally represents the end of routine daily employment in order to earn one's living. Mid-career change, on the other hand, involves leaving one job, either voluntarily or involuntarily, with the expectation of undertaking other employment. Although leaving the military after serving 20 or more years is commonly referred to as retirement, the majority of Canadian Forces (CF) leavers go on to second careers (Pinch & Hamel, 1978). Better insight into the problems of military mid-career change can be acquired by considering this phenomenon from the perspective of developmental and vocational psychology theories.

Research completed by Levinson, Darrow, Kline, Levinson and McKee (1978) suggested that there is a human life cycle made up of a series of stages and substages, with everyone more or less passing through this process. Levinson et al. (1978) and others (Lowenthal, Thuristes & Chiriboga, 1975; Neugarten, 1977; Vaillant, 1977) view the 40 to 50 age period as being particularly volatile, frequently marked by personal crises. Career and occupational concerns play a significant role in theories of adult development. Personal problems emanating from the workplace can gain overriding importance in a person's life. Medical research in the US Forces identified the "retirement syndrome" (Berkey & Stoebner, 1968; Druss, 1965; Greenburg, 1965; McNeil & Griffen, 1967; Milowe, 1964) which is characterized by both social and intrapsychic problems. These writers have posited that the social difficulties stem from: the financial inadequacy of the annuity, which must be supplemented by other income from a second career in the civilian labour force; indecision as to where to settle upon leaving the forces; loss of friends with similar jobs and common interests; and, perceived difficulty in finding employment in an environment that harbours misguided stereotypes of what the military retiree has to offer. The intrapsychic problems are in part attributable to the anticipated loss of status and high degree of responsibility for men and equipment, loss of the security which guarantees the individual that he and his family will be cared for if he becomes ill or incapacitated, loss of friends and lifestyle structure, and the spectre of impending ambiguity, uncertainty, complexity, and conflict attached to "starting over" in a civilian career.

According to Druss (1965) and Greenburg (1965), these social and intrapsychic problems during the latter part of the serviceman's career lead to marital discord, excessive drinking, depression, insomnia, a variety of psychosomatic complaints, and reduced work performance. Researchers in the Canadian context have also examined problems associated with mid-career change. Pinch and Hamel (1978) pointed out that many long-term servicemembers lack knowledge concerning the civilian job market, are unable to meet formal educational prerequisites for jobs, and have failed to recognize the necessity for advanced pre-retirement planning. These authors also found that personnel from "hard" military occupations (i.e. trades without readily discernable civilian counterparts) were unemployed longer and were less likely to regain former pay and status levels. Second career concerns affect all members to some degree and can have a deleterious effect on personnel performance during the last several years of service.

Recognizing that providing pre-retirement assistance not only fulfills a moral obligation, but also impacts favourably on operational effectiveness, the Department of National Defence (DND) has instituted the Second Career Assistance Network (SCAN) program. Components of SCAN include instruction in resume preparation and job search techniques, information on financial planning, military/civilian trade accreditation, and, counselling on personal and social adjustment concerns judged to be significant to mid-life career change. The aim of the program is to help the participants become more psychologically and practically prepared to leave the military and begin a second career. Even though SCAN has been functioning since 1978, its effectiveness for readying personnel for retirement has yet to be established. The work reported in this paper is part of research being conducted to assess SCAN's effectiveness in aiding the second career transition process.

To examine the effects of SCAN, the psychological construct vocational or career maturity was considered to be an appropriate conceptual framework. In previous research, career maturity was shown to be a major determinant of educational and occupational success for adolescents (Super, Crites, Hammel, More, Overstreet & Warrath, 1957; Super & Overstreet, 1960; Zelkowitz, 1974). Recognizing the difficulty inherent in measuring vocational maturity beyond adolescence, Super (1977) presented a revised theoretical model better suited to understanding career development in adults. Later modifications (Super & Knasel, 1979; Super & Kidd, 1979; Super, 1983) operationalized the model's dimensions and recommended that, in the case of adults, the term "career adaptability" be used instead of vocational maturity.

Thus, according to Super's adult paradigm, career adaptability is made up of five dimensions: Planfulness, Exploration, Information, Vocational Decision-Making, and Reality Orientation. This paper will concentrate on the Exploration or Exploratory Behaviour (EB) dimension of the model. EB is a broad concept concerned with attitudes toward vocational resources, ability to distinguish their worth from a personal perspective, and willingness to utilize them in searching for a second career. Due to the unique context of this research, the more descriptive term military disengagement readiness (MDR) is used synonomously with career adaptability throughout the paper.

Inasmuch as career adaptability or military disengagement readiness is a multi-dimensional psychological concept (Jordaan & Heydé, 1979; Super & Overstreet, 1960; Super & Knasel, 1979), it provides an appropriate medium through which the effects of the SCAN program can be measured. If SCAN is effective, there should be a positive relationship between participation in the program and MDR. For those with marketable civilian skills (e.g., electrician, plumber), SCAN appears to offer services that are immediately useful (e.g., resume writing, trade certification, job search techniques). However, for those whose trades are only remotely marketable, such as Combat Arms, the services offered by SCAN (e.g., information on academic upgrading, skill retraining courses) offer a more long term solution to their problems. This study treated the SCAN program and military trade category as independent variables and measured their effects upon one of the disengagement readiness dimensions, i.e., the independent variable, EB. More specifically, this research investigated: a) whether there are different initial levels of MDR, as measured by the EB scale, across military occupations; b) the overall effect of SCAN on EB; and, c) whether trades were differentially effected by SCAN.

METHOD

Subjects were 354 male non-commissioned officers divided into two groups: the SCAN treatment group, those who were registered in the program and had attended at least one seminar; and, the No SCAN control group, those who had no contact with the program but intended to register in SCAN before retiring. Within each of these groups, three trade categories were imposed: Cat 1, 'hard' military trades with no civilian counterpart; Cat 2, low to semi-skilled trades with civilian counterpart; Cat 3, highly skilled trades with civilian counterpart. The three trade categories were developed from occupational descriptions contained in the Canadian Classification Dictionary of Occupations (CCDO). Cat 1 contained 18 trades, Cat 2 contained 39, and Cat 3 contained 55.

The Military Disengagement Readiness Inventory (MDRI) was designed to measure disengagement readiness within CF retirees. The MDRI questionnaire contained eight scales reflecting Super's dimensions of career adaptability. The EB scale was developed based on the Super and Knasel (1979) and Jordan (1963) operational definition of vocational exploratory behaviour. It sought to measure the individual's involvement in such second career planning areas as learning about job search techniques, writing vocational tests, determining pension benefits and obtaining information about retraining programs. The scale contained 16 items and was scored using successive categories from 1 to 6. An item analysis indicated a coefficient alpha of .9038 for this scale.

The SCAN group was administered the MDRI at the start of a SCAN seminar, and again two months afterwards. To generate a control group, the MDRI was sent out during the same timeframe to members identified as having five or less years to serve before retirement. Individuals who indicated on the first MDRI that they were planning to register in SCAN at a later date were selected as the No SCAN group and received a second questionnaire at the same time as the SCAN group.

ANALYSIS AND RESULTS

A split plot ANOVA was used with two between subject factors, i.e., trade category and treatment group. The within subject variable, EB, was measured via the two MDRI administrations approximately two months apart. During the intervening period, the SCAN treatment group attended a SCAN seminar and participated in other program activities, whereas the No SCAN group was not exposed to any aspect of the program.

TABLE 1

Mean Exploratory Behaviour Scores of Treatment Group Over Time*

GROUP	TIME 1	TIME 2
SCAN (N=270)	47.87 SD=17.51	57.04 SD=18.45
No SCAN (N=84)	39.85 SD=16.28	42.45 SD=16.83

* Trade group means are not shown as all differences were non-significant.

Mean before and after EB scores for the SCAN and No SCAN groups are presented in Table 1. Overall F value indicated that there is no significant trade by treatment interaction over time. However, the treatment by time interaction was significant, F (347,1) = 14.60, $p < .001$ with the group receiving SCAN showing a greater increase in EB on the second measure of the MDRI. There were also significant main effects for time, F (347,1) = 108.11 $p < .001$, and treatment, F (347,1) = 30.68 $p < .001$, but not for trade, F (347,2) = .4759, $p > .05$. (Reported F values are based on multivariate test results.)

Paired comparisons, using Dunn's test of significance ($p < .05/4 = .0125$) indicated that upon the first testing, there was a significant difference between the SCAN and No SCAN groups, with the SCAN group displaying greater EB, t (359) = 3.57, $p < .001$. The second administration showed even greater difference between these two groups, t (359) = 6.46, $p < .001$. Over time, the SCAN group showed a significant increase in EB, t (269) = 10.63, $p < .001$, whereas the No SCAN group remained unchanged, t (83) = 1.91, $p = .06$.

DISCUSSION AND CONCLUSION

The aim of this research was to determine whether members trained in trades having readily recognizable civilian counterparts are more prepared to make the transition to a second career, whether SCAN has an impact on the readiness of the individual to make the transition, and, whether SCAN differentially affects readiness to seek a second career across trade categories. As demonstrated in this research, SCAN is definitely affecting an individual's predisposition to explore various ways of finding second careers. This corroborates earlier research conducted with adolescents demonstrating that career maturity can be taught (Yates, Johnson, & Johnson, 1979). However, all trades appear to be affected equally by SCAN. That is, those in "hard" military occupations report the same exploratory behaviour as those in more readily marketable occupations. An explanation for this finding may be that the EB items have more obvious applicability to second career concerns than other MDRI scales. Such activities as preparing a career resumé, learning about job search techniques, and writing vocational tests would appear to have equal relevance across trade categories.

However, this finding also may be an indication that the military occupations could be rearranged on a rational or statistical basis to reflect a better delineation of categories.

The results of this research suggest that SCAN is having a positive affect on at least one aspect of second career adaptability. It seems likely that an increase in knowledge and skills directed at easing the transition to the civilian workforce would reduce the stress and anxiety associated with retiring from the military. Thus, this leads to the belief that not only is SCAN fulfilling a moral obligation, but it likewise may be helping to maintain a more effective servicemember over the last several years prior to retirement.

This paper reported on the exploration dimension of the second career adaptability concept (Super, 1983). Although EB is only one of five discrete dimensions encompassed by second career adaptability, the results of this research suggest that the construct can be operationalized and used for gaining greater understanding of military second career transition. However, much stronger evidence to support the validity of the construct would be provided by investigating the post-retirement status of SCAN and non SCAN retirees.

REFERENCES

Berkey, B.R., & Stroebuer, J.B. (1968). The retirement syndrome: A previously unrevealed variant. *Military Medicine*, *133*, 5-7.

Druss, R.G. (1965). Problems associated with retirement from the military service. *Military Medicine*, *130*, 251-255.

Greenburg, H.R. (1965). Depressive equivalents in the pre-retirement years: The old soldier syndrome. *Military Medicine*, *130*, 251-255.

Jordaan, J.P. (1963). Exploratory behaviour. In D.E. Super, R. Statisherky, N. Martin, & J.P. Jordaan. *Career Development: Self-concept theory*. New York: College Entrance Examination Board.

Jordaan, J.P., & Heyde, M.D. (1979). *Vocational maturity during the high school years*. New York: Teachers College Press.

Levinson, D.J., Darrow, C.M., Kline, E.B., Levinson, M.H., & McKee, B. (1978). *The seasons of a man's life*. New York: Alfred Knopf.

Lowenthal, M.F., Thurister, M., & Chiriboga, D. (1975). *Four stages of life: A comparative study of women and men facing transitions*. San Francisco: Jossey-Bass.

McNeil, J.S., & Giffen, M.B. (1967). Military retirement: The retirement syndrome. *American Journal of Psychiatry*, *123*, 848-853.

Milowe, I.D. (1964). A study in role diffusion: The chief and the sergeant face retirement. *Mental Hygiene*, *48*, 101-107.

Neugarten, B. (1964). *Personality in middle and late life*. New York: Atherton Press.

Pinch, F.C. & Hamel, C., (1978). *The transition to civilian life among CF members: Preliminary results, stage II*. (Report 78-3). Toronto: Canadian Forces Personnel Applied Research Unit.

Vaillant, G.E. *Adaption to life*. (1977). Boston: Little, Brown and Co.

Super, D.E. (1977). Vocational maturity in mid-career. *Vocational Guidance Quarterly*, *25*, 294-302.

Super, D.E. (1983). Assessment in career guidance: Towards truly developmental counselling. *The Personnel and Guidance Journal*, May 1983, 555-562.

Super, D.E., Crites, J.O., Hummel, R.C., Moser, H.P., Overstreet, P.L., & Warnath, C.F. (1957). *Vocational development: A framework for research*. New York: Teachers College Press.

Super, D.E., & Overstreet, P.L. (1960). *Vocational maturity of ninth grade boys*. New York: Teachers College Press.

Super, D.E., & Kidd, J.M. (1979). Vocational maturity in adulthood - Toward turning a model into a measure. *Journal of Vocational Behaviour*, *14*, 255-270.

Super, D.E., & Knasel, E.G. (1979). *The development of a model, specifications, and, sample items for blue collar workers. Final report to Canada Employment and Immigration Commission*. National Institute for Careers, Education and Counselling, 1979.

Yates, C., Johnson, N., & Johnson, J. (1979). Effect of the use of the vocational exploration group on career maturity. *Journal of Counselling Psychology*, *26*, 238-270.

Zelkowitz, R.S. (1974). The construction and validation of a measure of vocational maturity for adult males. *Doctoral dissertation, Teachers College, Columbia University*, (University Microfilms No. 75-18, 456).

Development of an Air Force Training Decisions System

Sharon K. Garcia
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

I. Introduction

The management of technical training requires the coordination of efforts among numerous Air Force organizations. Significant roles in determining the nature of technical training are filled, for example, by the Air Force deputy chief of staff for manpower and personnel; by Air Force functional managers; and by agencies at the Air Force Manpower and Personnel Center in charge of job classification, assignments, and management of the on-the-job training programs. The Air Training Command at its headquarters, technical training centers, and other units plays a central role in policy as well as day-to-day training management. The headquarters of all major commands have a major voice in training decisions; and additional inputs are gathered from various ad hoc training study groups, AF laboratories, and research centers.

Because of the scope and complexity of the training and personnel systems, many decisions that impact on training are made, to some extent independently, by different management units responsible for different parts of the training and personnel systems. Conflicting goals are inevitable. As each unit attempts to optimize operations within its own area, the net result is competing objectives and total system suboptimization. Part of the problem is that relevant data for many training decisions are not available. For example, in practice, the amount of resident technical school training is largely determined by pre-defined budgets, and whatever content is not covered in the school is left to on-the-job training, with little data for assessing the impacts on OJT resources, costs, and capacities. In addition, alternative ways in which the overall training system might be restructured in concert with operational changes in personnel utilization are not considered. Hence, a more unified total systems approach to such problems with all relevant data considered was needed. In response to this need, Air Staff and the HQ Air Training Command requested the AF Human Resources Laboratory to conduct research in an attempt to refine current training decision procedures. The research, entitled the Training Decisions System (TDS), has as its objective the development of a computer-assisted decisions system to address the what, when, and where training decisions for any one specialty. What refers to what job tasks, skills, and knowledges should be emphasized in training. When refers to at what point in an airman's career training should be provided (e.g., upon entry, OJT, advanced training). Where, refers to the question of where training should be provided (e.g., resident school, field units, CDC, MAJCOM, other).

II. Development of the Training Decisions System (TDS)

The TDS will involve the development of four basic, user-friendly interactive subsystems. They are the Task Characteristics Subsystem (TCS), Field Utilization Subsystem (FUS), Resource/Cost Subsystem (RCS), and the Integration and Optimization Subsystem (IOS).

The Task Characteristics Subsystem (TCS) will be developed and used for identifying clusters of job tasks that can be appropriately trained as a unit, based on common skill and knowledge requirements and other shared characteristics (e.g., the probability of co-performance). These units will be known as Task Training Modules (TTMs) and will be used as the basic unit of analysis in the Training Decisions System. Once task training modules have been designed, a methodology will be developed for allocating these TTMs to the appropriate training settings or sites which may include initial skill resident training, on-the-job training, or correspondence courses.

The Field Utilization Subsystem (FUS) will be used to describe existing patterns of airman utilization in terms of jobs, training states, and major career paths. In addition, the FUS will attempt to define training/personnel assignment patterns that represent alternative approaches to training, assignment, and use of airmen in a particular specialty, based on management preferences. Both training content and job descriptions will be represented by collections of task training modules.

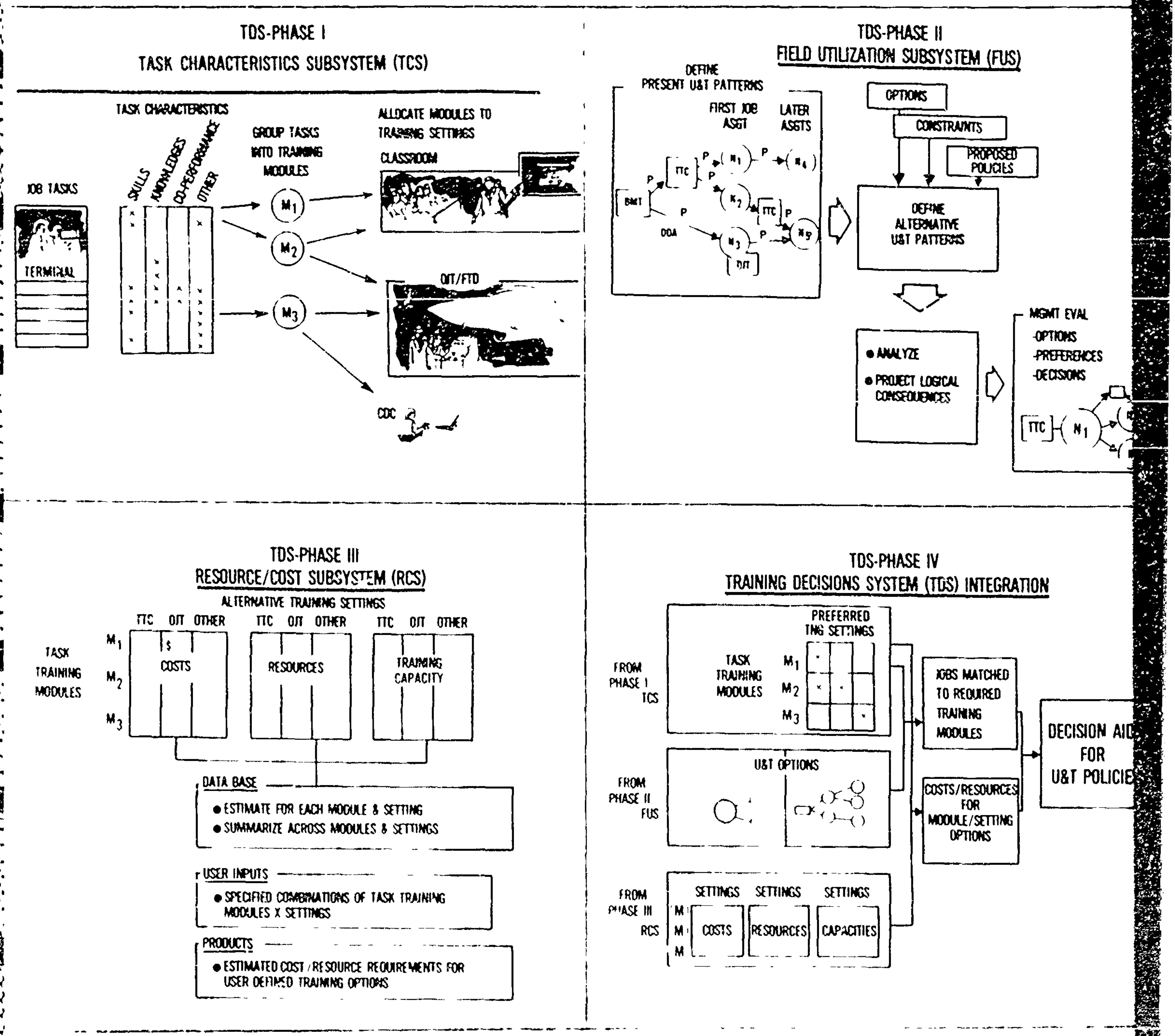
The Resource Cost Subsystem (RCS) will be developed for estimating the costs and resources required for training different clusters of job tasks in alternative training settings; for estimating the training capacities of alternative training settings; and for developing summary estimates of costs, resources, and capacities needed for specified training alternatives.

The Integration and Optimization Subsystem (IOS) will result in the integration of the three previously described subsystems and develop effective decision aids for AF technical training designers. The final product of the IOS will be the Training Decisions System; a user friendly software package, bringing together a complex of elements describing training, personnel, and cost factors, as well as management policy preferences. State-of-the-art decision model technologies will then be applied to these elements to answer "what if" questions for management, and to develop "optimal" training designs, for more cost-effective training decisions. A conceptual diagram of each of the TDS subsystems is provided in Figure 1.

Research and development of the TDS is a four-year contract effort. The prime contractor is McDonnell Douglas Astronautics. The contract began in Sept 83 and will be completed in Sept 87. Initial development of the TDS is being applied to four Air Force career ladders or occupations. They are Avionic Inertial and Radar Navigation Systems, Security/Law Enforcement, Aircraft Environmental Systems, and Electronic Computer and Switching Systems.

III. Conclusions

Once completed, the Training Decisions System will produce a training decisions system that will provide readily available, validated information to the Air Staff and user commands, especially Air Training Command, on costs and consequences of training decision alternatives under different constraints, costs, and personnel utilization patterns. The following benefits are anticipated from the implementation of such a system: (a) enhanced mission readiness through optimizing the match of technical training resources and overall operational demands, (b) increased training efficiency through optimizing the sequence and settings in which training occurs, (c) improved personnel utilization through development of methods for analyzing functional job patterns in relation to optimized training sequences, (d) increased cost effectiveness of training through the formulation of training decisions based on explicit cost and resource consequences, and (e) reduction of excessive operational training commitments through more accurate estimation and analysis of unit capacity to train while meeting ongoing mission demands. In short, TDS will give managers the tool to plan for the best training for the dollar.


TDS-PHASE I
TASK CHARACTERISTICS SUBSYSTEM (TCS)
JOB TASKS
TERMINAL
TASK CHARACTERISTICS
SKILLS
KNOWLEDGES
CO-PERFORMANCE
OTHER
GROUP TASKS INTO TRAINING MODULES
M_1
M_2
M_3
ALLOCATE MODULES TO TRAINING SETTINGS
CLASSROOM
OJT/FTD
CDC
TDS-PHASE II
FIELD UTILIZATION SUBSYSTEM (FUS)
DEFINE PRESENT U&T PATTERNS
FIRST JOB ASGT
LATER ASGTS
BMT
TTC
P
DOA
OJT
M_4
M_5
OPTIONS
CONSTRAINTS
PROPOSED POLICIES
DEFINE ALTERNATIVE U&T PATTERNS
● ANALYZE
● PROJECT LOGICAL CONSEQUENCES
MGMT EVAL
-OPTIONS
-PREFERENCES
-DECISIONS
TDS-PHASE III
RESOURCE/COST SUBSYSTEM (RCS)
ALTERNATIVE TRAINING SETTINGS
TTC
OJT
OTHER
TASK TRAINING MODULES
$
COSTS
RESOURCES
TRAINING CAPACITY
DATA BASE
● ESTIMATE FOR EACH MODULE & SETTING
● SUMMARIZE ACROSS MODULES & SETTINGS
USER INPUTS
● SPECIFIED COMBINATIONS OF TASK TRAINING MODULES X SETTINGS
PRODUCTS
● ESTIMATED COST/RESOURCE REQUIREMENTS FOR USER DEFINED TRAINING OPTIONS
TDS-PHASE IV
TRAINING DECISIONS SYSTEM (TDS) INTEGRATION
FROM PHASE I TCS
TASK TRAINING MODULES
PREFERRED TNG SETTINGS
FROM PHASE II FUS
U&T OPTIONS
FROM PHASE III RCS
SETTINGS
M
COSTS
RESOURCES
CAPACITIES
JOBS MATCHED TO REQUIRED TRAINING MODULES
COSTS/RESOURCES FOR MODULE/SETTING OPTIONS
DECISION AID FOR U&T POLICIE

TDS CONCEPTUAL DIAGRAM

DEFINING TASK TRAINING MODULES:
COPERFORMANCE CLUSTERING

Drs. B. M. Perrin, D. S. Vaughan, R. M. Yadrick,
& J. L. Mitchell
McDonnell Douglas Astronautics Company
Saint Louis, MO & San Antonio, TX

Training decision-making in the Air Force is a process of balancing, either explicitly or implicitly, a number of distinct and often conflicting considerations. Instructional effectiveness, manpower and personnel utilization plans, and financial factors must all be balanced in deciding who gets trained, when during their career, on what skills, and using which modes of instruction. Currently, Utilization and Training Workshops are used as a forum to weigh these considerations in determining training policy for an Air Force Specialty (AFS). Difficult decisions are made even more complex however, because the information available to these groups is often fragmentary, due to the number of distinct skills and knowledges in an AFS, the number of possible instructional modes, and the costs associated with training each skill using each mode.

The Training Decisions System (TDS) will be a means of bringing together information concerning each of these factors--instructional, personnel utilization, and financial--to aid Air Force managers in establishing training policy. The key elements of the TDS are the sets of skills and knowledges for which the relative instructional efficiencies of various training and utilization policies will be determined. These sets of skills and knowledges will be in the form of groups of Occupational Survey (OS) tasks, known as Task Training Modules (TTMs). Ideally, TTMs will be groups of OS tasks that are relatively homogeneous with respect to underlying skills and knowledges and that are relatively distinct from other groups of OS tasks (i.e., other TTMs); consequently, TTMS should capture efficiencies of training that might result from common training materials, content, equipment, and the like.

Additional training efficiencies accrue from training similar tasks together if these tasks are also coperformed (i.e., performed by the same personnel). Thus, TTMs should be composed of tasks which are similar and which are performed by the same personnel, if the information to be provided by the TDS is to be maximally useful. In a separate paper that follows in these proceedings (Yadrick, Vaughan, Perrin, and Mitchell, 1985), we have documented our method of obtaining expert judgments concerning task similarity. In this paper, we present procedures that may be used to hierarchically cluster tasks based on their reported coperformance in the OS. These statistical clusterings will then be compared to experts' groupings of the same tasks. Results are presented for two AFSs--328X4, Avionic Inertial and Radar Navigation Systems; and 811XA, Security, Law Enforcement, and Law Enforcement-Military Working Dog Qualified.

METHOD

Statistical clustering has been used for some time in military occupational analysis to aid job analysts in identifying job-types (i.e., groups of individuals performing similar tasks). The clustering technique that has been applied to this problem and is utilized in the Comprehensive Occupational Data Analysis Programs (CODAP) system is the average linkage clustering procedure (Ward, 1963). This procedure has performed well in empirical studies, as compared to other procedures reported in the statistical literature (Milligan, 1981; Mojena, 1977).

Figure 1 illustrates the relationship between case (or person) clustering, the normal application of CODAP to job-typing, and task clustering, the application to identifying tasks that are performed by the same personnel. (Note: Air Force job-typing normally uses a relative time spent measure for clustering cases rather than the performed/not performed dichotomy depicted in Figure 1; in this case, a number between 0 and 1 representing relative time spent on a task would replace the 1's in the figure.) While job-typing involves grouping persons who perform the same (or similar) sets of tasks, task clustering produces sets of tasks which are frequently coperformed.

Task clustering can be accomplished in CODAP by transposing the raw data file that is input to the system. That is, instead of task performance data for each person as represented on the left side of Figure 1, each record would consist of person performance data for each task (as represented on the right side of Figure 1). This transposed data would then be processed by CODAP, yielding a task coperformance cluster diagram.

Transposing the data input file has the effect of making the number of cases appear to the system as the number of tasks, and the number of tasks becomes the number of cases. This effect can have serious practical implications when the number of cases is large, as CODAP is limited in the number of "tasks" (transposed cases) it can process (the IBM version can handle up to 2000 tasks, and the UNIVAC version can process up to 1700, although a UNIVAC rewrite of CODAP will be able to process 3000 tasks).

FIGURE 1: A COMPARISON OF CASE CLUSTERING AND TASK CLUSTERING

JOB TYPING: CLUSTERING PERSONS WHO PERFORM THE SAME (OR SIMILAR) SETS OF TASKS

		TASKS								
	PERSONS	1	2	3	4	5	6	7	8	9
JOB I	1	1	1	1						
	2	1	1	1						
JOB II	3	1	1	1	1	1	1			
	4	1	1	1	1	1	1			
JOB III	5				1	1	1			
	6				1	1	1			
JOB IV	7							1	1	1
	8							1	1	1

TTM CONSTRUCTION: CLUSTERING TASKS THAT ARE PERFORMED TOGETHER

9-3278

		PERSONS							
	TASKS	1	2	3	4	5	6	7	8
TTM I	1	1	1	1	1				
	2	1	1	1	1				
	3	1	1	1	1				
TTM II	4			1	1	1	1		
	5			1	1	1	1		
	6			1	1	1	1		
TTM III	7							1	1
	8							1	1
	9							1	1

To avoid this limitation, one may compile a task similarity matrix external to CODAP, and then, use this matrix to cluster the task data. The performed/not performed similarity measure used in CODAP is as follows:

$$S_{ij} = ((N_{ij}/N_i) + (N_{ij}/N_j)) / 2$$

where N_{ij} is the number of persons performing both tasks i and j, N_i is the number of persons performing task i, and N_j is the number performing task j. In words, the similarity between pairs of tasks is the average of the two ratios of the number of persons performing both tasks divided by the number performing each task.

Both procedures--the transposition of the raw data and the computation of a similarity matrix--were used to analyze task coperformance in the 328X4 OS data to verify the methods. Because of the number of cases in the 811XX OS sample, in excess of 6000, a task similarity matrix computed outside of CODAP was used to cluster the data.

One measure of the homogeneity of the tasks taken to form a TTM is the between group similarity. Between group similarity is the average of the similarities of the tasks between the groups being merged to form a particular cluster. It is defined as follows:

$$BG_{ij} = \sum\sum S_{ij}/N_{ij}$$

where S_{ij} is the similarity between tasks i and j from the two groups to be merged and N_{ij} is the number of task comparisons between the groups.

Several statistics are available to compare clusterings or groupings of a single sample, based on a pairwise classification of cases for the two solutions. Table 1 illustrates the classification scheme, where each of the four cells specifies a type of agreement or disagreement between the solutions. For example, cell A indicates the number of pairs of cases grouped by both methods, while cell D is the number of pairs grouped separately by both solutions. Cell B and C indicate frequencies of disagreement in which pairs are grouped by one method, but not the other. Three comparison statistics can be defined in terms of these cell frequencies as follows:

Rand (1971): $(A + D) / (A + B + C + D)$
Jaccard (Downton & Brennan, 1980): $A / (A + B + C)$
Fowlkes & Mallows (1983): $A / \sqrt{(A + B) \quad (A + C)}$

Table 1: A pairwise classification scheme used to compare two clustering solutions.

		Solution 1	
		pair in same cluster	pair not in same cluster
Solution 2	pair in same cluster	A	B
	pair not in same cluster	C	D

All three statistics yield a value of 1.00 when the two solutions agree perfectly, and all three have a lower bound of 0. The Rand statistic, despite being the most widely used, is often inflated by using pairs not classified together by either procedure (cell D) as reflective of solution consistency. Both the Jaccard and the Fowlkes & Mallows statistics were devised, in part, to overcome this problem. Sampling distributions are not available for these statistics; consequently, the numbers are not directly interpretable as indicating either agreement or disagreement between methods.

Two complications in comparing expert and statistical groupings of the tasks are worth noting at the outset. First, due to infrequent performance of some tasks within an AFS, the experts dropped some tasks from their groupings. While the number of deleted tasks was generally very small, one group of experts dropped 27 tasks from the 328X4 OS task list. When this complication occurred, these tasks were omitted from the analysis, and so, are not reflected in the comparison statistics.

The second complication resulted from experts placing the same task in two or more TTMs--an action they were directed to take if they believed it was necessary. Again, in absolute terms, this action was taken relatively infrequently, although one task statement was placed in five different TTMs by one group of experts. This complication was handled by counting each occurrence of the task individually; that is, one sorting by the experts may have agreed with the coperformance clustering (and be counted in cell A of Table 1), while a second sorting may have disagreed with the statistical clustering (and be counted in cell B or C).

RESULTS

Unlike case clustering to identify job-types which has more than 30 years of research and practice behind it, very little information exists to aid in identifying TTMs from a task coperformance cluster diagram. Rules-of-thumb which occupational analysts have adopted to narrow the search for important jobs may or may not be relevant to the search for TTMs. Particularly problemmatic is the determination of the number of TTMs. Since the clustering solution is hierarchical, TTMs of any degree of specificity can be identified.

An approach to this problem is to use the same type of heuristics a job analyst would use to interprete a case clustering. One of the authors of this paper, who is familiar with both occupational analysis and the 811XX career field, used this approach. He identified nineteen general task content areas (similar to the higher order job clusters in job analysis) and 67 TTMs within those general areas. The TTMs varied in size from 2 to 34 tasks and averaged just under 10 tasks per TTM. The between group similarities of the TTMs identified from the 811XX OS task coperformance clustering ranged from 26.75 to 91.20, and averaged 58.87.

While these TTMs, by definition, have the desirable characteristic of being composed of tasks which tend to be performed by the same personnel, the degree to which the tasks are similar in terms of skills and knowledges is not known. Some indication of the skill/knowledge homogeneity of the coperformance clusters can be obtained by comparing these results to those obtained from having experts group the tasks.

Two additional issues had to be addressed before the results of the expert and statistical clusterings could be directly compared. First, the number of statistical clusters had to be determined, since the comparison statistics are influenced by the specificity of the results. To promote comparability between the procedures, the number of statistical clusters was set equal to the number of expert task groupings for each comparison. The second issue dealt with how to select the statistical clusters; it was decided to use the task clusters which maximized between groups homogeneity. This purely statistical criterion was chosen because it permitted selection of different sets of TTMs with different degrees of specificity. This approach also eliminated the variability that might have resulted from differences between analysts, had interpretations of the coperformance cluster diagram been used. Thus, the comparisons reflect the degree to which the statistical clusterings, without interpretation, capture expert task groupings.

Table 2 summarizes the results of the comparisons between the coperformance clustering produced by CODAP and two task groupings produced by separate groups of experts. Both the Jaccard and the Fowlkes & Mallows statistics are reported. The comparison between the two expert groupings in the 328X4 AFS produced the highest degree of convergence, substantially higher than that found between the 811XX expert groups. It should be noted, however, that the 328X4 groups consisted solely of technical trainers, while the 811XX groups contained both school and field personnel. The similarity of the backgrounds of the 328X4 experts may partially account for the greater convergence of their results.

The task coperformance clustering in the 811XX AFS matched the expert groupings as well as the experts' results matched each other. Additionally, the coperformance clustering agreed more closely with the experts' classifications in this specialty than in the 328X4 AFS. Again, this result may be due to the influence of the field personnel, whose perspective presumably is more attuned with performance factors.

CONCLUSIONS

The evidence concerning task ccperformance clustering suggests that this procedure is workable and produces task groupings which are homogeneous with respect to content. Once computer software is developed, task coperformance clustering can be accomplished at relatively little additional expense, using existing OS data and CODAP clustering procedures. As this methodology develops, additional items may be added to the OS and computer routines derived to aid in the identification and interpretation of TTMS.

Table 2: Comparison of statistical and expert groupings of tasks in the 328X4 and the 811XX AFSs. The Jaccard and the Fowlkes & Mallows statistics are reported. Expert groupings I and II are results from independent efforts.

328X4 AFS	Expert Groupings - I	Expert Groupings - II	Ccperformance Clusters
Expert Groupings - I	----	.271 .476	.087 .171
Expert Groupings - II		----	.121 .224
Coperformance Clusters			----

811XX AFS	Expert Groupings - I	Expert Groupings - II	Coperformance Clusters
Expert Groupings - I	----	.171 .293	.157 .314
Expert Groupings - II		----	.179 .350
Coperformance Clusters			----

The agreement between the statistical and expert groupings of the tasks, given that it was roughly equivalent to that between independent clusterings by experts, is encouraging. These results are even more promising considering that the task clusters were based solely on a statistical criterion, the between groups similarity. We would not recommend that this technique be directly implemented. Rather, an experienced analyst, working with one (or more) experts in the career field, should be used to identify and refine TTMs from a coperformance cluster diagram. This procedure would almost undoubtedly produce more homogeneous and useful TTMs.

One final word of caution is required. The skill/knowledge homogeneity of the task clusters produced by this methodology is still somewhat suspect, as the overall comparison statistics reported cannot specify the nature of the disagreement between results. Presumably, the experts' groupings might differ only according to subtle variations among tasks that are clustered by one group, but not the other. On the other hand, coperformance clusters, while exhibiting the same absolute number of differences, might include strikingly different types of tasks that are, nonetheless, coperformed. A more stringent validation effort is currently underway to assess this possibility.

REFERENCES

Downton, M., & Brennan, T. (1980, June). Comparison classifications: An evaluation of several coefficients of partition agreement. Paper presented at the meeting of the Classification Society, Boulder, Colorado.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings, (with comments and rejoinder). Journal of the American Statistical Association, 78, 553-584.

Milligan, G. W. (1981). A review of Monte Carlo tests of cluster analysis. Multivariate Behavioral Research, 16, 379-407.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. The Computer Journal, 20(4), 359-363.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66, 846-850.

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58, 236-244.

Yadrick, R. M., Vaughan, D. S., Perrin, B. M., & Mitchell, J. L. (1985, October). Evaluating task training modules: SME clustering and comparisons. Paper prepared for the meeting of the Military Testing Association, San Diego, California.

EVALUATING TASK TRAINING MODULES: SME CLUSTERING AND COMPARISONS

DRS. R. M. YADRICK, D. S. VAUGHAN, B. M. PERRIN
& J. L. MITCHELL

MCDONNELL DOUGLAS ASTRONAUTICS COMPANY
SAINT LOUIS, MO. & SAN ANTONIO, TX.

INTRODUCTION

One of the most straightforward ways to categorize information is simply to have people sort cards into piles. A single instance of the categories to be formed is printed on each card, and the piles that result constitute the categories. This procedure has a long history in psychology for research in cognitive modeling, and seemed suitable for use in forming Task Training Modules (TTMs) in support of the Training Decisions System (TDS) research (Garcia, 1985).

The approach was pilot tested at Scott AFB, with 811XX personnel (law enforcement) job incumbents serving as subjects, or Subject Matter Experts (SMEs). Cards were labeled with individual tasks from the most recent Occupational Survey (OS) on the 811XX AFS. Cards were initially grouped into "starter piles", in which all the tasks shared a common Specialty Training Standard (STS) paragraph reference. Starter piles provided SMEs with initial working units of manageable size, and also with a reasonable conceptual base (the STS paragraphs) from which to start.

SMEs were instructed to simply rearrange the starter piles to form groups of tasks that "should be trained together". Piles could contain many tasks, few tasks, or only a single task. Duplicate task cards could be placed in different piles. These rather flexible instructions and liberal sorting options were incroduced in order to allow maximum expression of SME opinion and avoid forcing upon them any of our own notions about how TTMs should be formed.

There were two "passes" for each within-job sort. That is, SMEs initially rearranged the STS piles into their own piles. They then went through their own piles to make any needed changes. They worked at their own pace with no time constraints.

The results of this pilot test are reported elsewhere (Vaughan, Yadrick, Dunteman, and Clark, 1984) and need not be examined here in detail. Briefly, the resulting TTMs seemed like reasonable task groupings, and the overall card-sorting approach was feasible and deserved field testing. SMEs found it easy to work with the starter piles, performed conscientiously, and were satisfied with their own results.

In the present effort, the card-sorting method was field tested for 2 AFSs. The objectives of this project were to develop an operational system for TTM construction, and to get task clusters that could be compared to computer-generated clusters (Perrin, Vaughan, Yadrick, and Mitchell, 1985).

METHOD AND PROCEDURE

The card sorting process was modified and expanded for field testing purposes. Some of the modifications reflected policy decisions on our part (e.g., we decided not to have separate card sorts for each job in each specialty, but rather to have whole-AFS sorts), but most were minor changes designed to streamline the process.

The Security Police and Law Enforcement (811XX) card sorting was conducted at Lackland AFB, Texas, in May, 1985. The fourteen SMEs present represented a relatively even mix of the three 811XX shredouts, namely security police, law enforcement personnel, and military working dog (MWD) handlers. Technical trainers from ATC were represented in fairly equal proportion with operational personnel.

The SMEs were divided into four separate groups. These groups worked independently of the other groups, except during the final stage of the process. This stage will be described later. Again the groups were formed to provide the most even mix of SME backgrounds, shreds, etc., as possible. SMEs then received instructions and began sorting.

Two groups received starter piles in which all tasks shared a common STS reference, as in the pilot study. The other two groups received starter piles composed of tasks which had clustered together in the coperformance clustering process (Perrin, et al., 1985). The resulting TTMs could then be compared and evaluated. In addition, this would help mitigate any effect of starter pile content upon the final TTMs.

SMEs made three passes, all at their own pace. On the first pass, they oriented themselves to the exercise, examining piles for homogeneity of coperformance, skills and knowledges, combining and subdividing the piles as necessary. On the second pass, the newly created piles were recombined to refine the coperformance, knowledge and skill groupings. SMEs made further checks and refinements ("fine tuning") on the final pass. The results (TTMs) of each pass were recorded.

The four groups finished at very different times, as might have been expected. Two groups finished about halfway through the second day. Various groups dynamics were clearly observable, such as the domination of one group and of the sorting results by a single forceful member of the group.

Different groups worked as though they had quite different interpretations of the instructions. For example, one group arranged piles so as to train what they came to call a "super cop". They carefully arranged sequences of TTMs so that, by receiving training on each full TTM in their prescribed sequence, the result would be an airman who could perform literally all the tasks and jobs in all the shreds of the specialty. Most groups, however, adopted very different and more realistic strategies.

The final phase of the process was conducted to reconcile the different sortings between groups. In each reconciliation, groups which had started with the same piles (e.g., the groups with STS starter piles) were teamed together.

Both reconciliation groups were each to provide a single set of "final" TTMs. Unfortunately, only one reconciliation group was actually able to finish. The other reconciliation group, made up of the two original groups which did not finish their own sortings in the first two days, were forced to rush through the reconciliation step.

The same essential process was carried out at Keesler AFB, with Avionic Inertial and Radar Navigation Systems (328X4) SMEs in August, 1985. There were, however, important differences. All of the SMEs available were instructors at the technical training school, and no operational people were present. Also, there were only enough SMEs to form two groups instead of four. As a result, the desired replication could not be done. Operational workers reviewed and refined the TTMs obtained at Keesler, although these refinements have not yet been examined in detail. The two 328X4 groups received essentially the same instructions a had the 811XX groups, although we stressed somewhat more the idea that TTMs should reflect common skills and knowledges required to do particular jobs. This change was made mainly to avoid any misinterpretation of instructions.

RESULTS

In the 811XX AFS, the reconciliation effort resulted in 69 TTMs from the groups that had coperformance cluster starter piles. Of these, only four TTMs contained a single task, and 52 contained five or more tasks. The largest contained 42 tasks. These were sorted from 19 starter piles containing a total of 666 tasks.

The other reconciliation effort (groups with the STS starter piles) resulted in 65 TTMs formed from 39 started piles. Only one contained a single task. The largest contained 31 tasks.

Although these results appear to be in reasonably good agreement. the similarities are only superficial. As reported in another paper in this session (Perrin, et al. 1985), both the Jaccard index (Downton & Brennan, 1980) and that from Fowlkes and Mallows (1983) were relatively low (0.171 and 0.293, respectively) showing relatively poor agreement between the two groups in assigning tasks to TTMs. Indeed, both indices were higher (0.179 and 0.350, respectively) and showed greater agreement between the STS starter pile reconciliation and the pure coperformance clustering than between the two card sorting reconciliation groups.

In the 328X4 AFS, the group which started with STS piles formed 140 TTMs from 31 original piles containing 778 tasks. For the most part, these were small. Eighty-six contained fewer than 5 tasks, and many of these contained only one task. However, two were especially large, containing 78 and 48 tasks, respectively.

The group that started with coperformance piles formed 33 TTMs from 58 original piles. Eleven of these contained over thirty tasks, and two contained over eighty tasks.

Although there appears to be considerable disagreement between the groups, it is apparently superficial. Table 1 reports Jaccard and Fowlkes-Mallows indices for several comparisons of these card sorts with each other, with the later reconciliation sort done between the two groups, and with coperformance clustering. In general, the two groups actually agreed relatively well. The obvious conclusion is that the 140 TTMs formed by the one group were generally subgroupings of the 33 TTMs formed by the other. Indeed, the reconciliation data indicate fairly good agreement with both groups (see Table 1). Also, reconciliation

Table 1: Comparison of SME card sorts and task coperformance statistical clustering in the 328X4 AFS*

	Reconciliation Card-Sort	Card-Sort From Coperformance Starter Piles	Card-Sort From STS Starter Piles	Task Coperformance Clusters
Reconciliation Card-Sort	----	.326 .520	.637 .787	.127 .244
Card-Sort From Coperformance Starter Piles		----	.271 .476	.087 .171
Card-Sort From STS Starter Piles			----	.121 .224
Task Coperformance Clusters				----

*For each table entry, the Jaccard statistic is on top, and the Fowlkes-Mallows statistic is on the bottom.

resulted in 75 TTMs, roughly halfway between the numbers of TTMs produced by the groups separately. It would seem that they found compromise easy to make. Their TTMs, however, do not agree very well with the computer-generated coperformance task clusters. As Table 1 shows, both the Jaccard and Fowlkes-Mallows statistics were less than .25 for all such comparisons.

DISCUSSION

The results reported here are tentative. Final TTMs have not yet been produced for either AFS studied. The 328X4 TTMs must be replicated with additional field data. Nonetheless, even at this point it appears that neither the clusters derived from task coperformance clustering nor those constructed by a single group of SMEs would yield stable TTMs. Instead, some combination of methods will be required that allows tentative TTMs to be crossvalidated and refined by representative groups of field experts. Once a method is developed that produces final, stable TTMs for each specialty, these TTMs can be used as criteria against which more economical clustering methods can be validated.

REFERENCES

Downton, M., & Brennan, T. (1980, June). Comparing classifications: An evaluation of several coefficients of partition agreement. Paper presented at the meeting of the Classification Society, Boulder, Colorado.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings, (with comments and rejoinder). Journal of the American Statistical Association, 78, 553-584.

Garcia, S. (1985, October). Overview of the Training Decisions System. Paper prepared for the meeting of the Military Testing Association, San Diego, California.

Perrin, B. M., Vaughan, D. S., Yadrick, R. M., & Mitchell, J. L. (1985, October). Defining task training modules: Coperformance Clustering. Paper prepared for the meeting of the Military Testing Association, San Diego, California.

Vaughan, D. S., Yadrick, R. M., Dunteman, G. H., & Clark, B. L. (1984). Feasibility of task training module construction methods and preliminary task characteristics subsystem project design. Saint Louis, MO: McDonnell Douglas Corporation.

NONHIERARCHICAL CLUSTERINU OF AIR FORCE JOBS AND TASKS

Jimmy L. Mitchell, PhD
McDonnell Douglas Astronautics Co.
San Antonio, Texas

William J. Phalen
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

INTRODUCTION

One of the problems with the typical occupational analysis project is the limitation of coverage of job groups in the normal Comprehensive Occupational Data Analysis Programs (CODAP) hierarchical clustering solution. In an operational study, the groups of minimally acceptable size existing at any stage of the clustering, as reflected on the diagram of the occupation, will not include all the cases, except at a very low stage, where between-group overlap values are also very low (see Figure 1). This is a function of both the degree of homogeneity of jobs within the specialty and the order in which the cases enter the initial groups. It is not unusual to have 5 to 20 percent of the cases in a study excluded from the final job groups identified in an occupational survey report (OSR).

While the identification of the job groups in an occupation or Air Force Specialty (AFS) has considerable utility in the personnel classification and training system, the final job descriptions for such groups have some inherent limitations (Carpenter, 1974; Pass & Robertson, 1978). The hierarchical clustering procedure normally used is an iterative process whereby, once a case is clustered in a group, it is no longer considered in terms of its similarity of tasks performed or relative time spent performing against later-formed groups. Some researchers have proposed that other clustering methodologies or other similarity measures be used (Pass, 1980). No one approach, however, has emerged as an optimum clustering alternative.

In the present study, a nonhierarchical clustering methodology was used to refine the job types identified in the normal CODAP process. The usual hierarchical clustering procedure was used to identify the "seed" groups and an iterative, nonhierarchical clustering method was added to refine these groups. The expectation was that a combination of the two methods would provide superior results in terms of group membership, percent of cases accounted for, and realism of the resulting job descriptions.

METHODOLOGY

A runstream of existing CODAP programs has been put together to give the CODAP system a nonhierarchical capability (Datko, 1985; Phalen, Weissmuller, & Staley, 1985). This new procedure was developed to facilitate analysis of various types of officer job rating scales (relative time spent versus part of the job, complexity, etc.), and has already proven very useful in that context.

If a set of CODAP-generated job descriptions is input to this nonhierarchical clustering program as "seed" profiles, an unlimited number of cases, as represented by their individual job descriptions, can be classified into homogeneous groups, according to which one of the "seed" profiles each case resembles most closely. Multiple iterations can be run to ensure that each case has the opportunity to switch to a more compatible group, or to switch to a group whose homogeneity is increased the most by classifying the case in it, until the number of reclassifications has reached a minimum. The reclassification of cases, previously grouped by a hierarchical clustering, is a way to improve within-group homogeneity and between-group differences (Phalen et al., 1985). Also possible in the system is multiple group membership by a case or group of cases.

While the membership of the nonhierarchically clustered job types will overlap to a very substantial degree with those produced by the normal CODAP clustering, this new procedure permits the "isolates" or unclustered cases to be included in the covered sample. In addition, the resulting job descriptions should be considerably more homogeneous in terms of both the "core" tasks of the group and the relative time spent on those tasks. Thus, the relative worth of the new procedure could be assessed by examining the increase in the homogeneity of the groups. Such a procedure, then, should result in a clearer definition of the variety of jobs within a specialty, and a more realistic determination of the training requirements for each such job.

SAMPLE

CODAP case data were readily available for a sample of first-enlistment Security Policemen (AFSC 811X0) who were being studied in another line of research (Perrin, Vaughan, Yadrick, & Mitchell, 1985). The sample included 3,302 first-enlistment individuals from the most recent Security Police occupational analysis study (Alton, 1984). A separate first-enlistment diagram was analyzed and 82 811XX first-term seed job groups were identified which had reasonable internal homogeneity (overlap between combining groups ≈ 35 and overlap within combined groups ≈ 50). Group size ranged from 3 to 210; the average group size was 27. Some very small groups were included which were less than the starter group size of 10; these smaller groups were identified by closely examining several small, undefined heterogeneous groups with mixed job titles and low within-group overlap values (such as Customs or Kennel Support). The reason for including these very small groups was to see if they would gain sufficient additional members in the nonhierarchical process to become legitimate job groups.

The 82 groups accounted for only 67 percent of the cases in the sample, which illustrates the coverage problem discussed earlier. Each group was named and related groups were given overlapping component names to facilitate analysis of their relationships.

RESULTS

The initial nonhierarchical grouping yielded a marked reduction in the number of cases not classified into the 82 seed job types. Only five cases remained unclassified for a 99.8 percent coverage (see Figure 2). In succeeding iterations, the number of unclassified cases increased slightly, but even at the sixth iteration, the proportion of cases covered was greater than 99 percent. Thus, it would appear that almost all of the unaccounted-for cases in the original clustering were reasonably similar to the major job types identified in the initial analysis. The question now became one of assessing the impact of adding these previously unaccounted-for cases to the groups.

One way to address this issue was to examine the within-group overlap values, averaged across all 82 groups. For the initial iteration of the nonhierarchical clustering process, the average within-group overlap was 48.23. The within-group standard deviation averaged across all groups was 10.68. For the second run, the mean within-group overlap value was 48.65 (vice 48.23) with an average S.D. of 8.81 (vice 10.68). Thus, this second iteration had little if any impact on the average within-group overlap but resulted in a substantial decrease of the within-group variance. This change in values indicated that the new job group descriptions were considerably more homogeneous. This is, however, a summary statistic and, in order to demonstrate a meaningful impact on the job-typing process, we must examine how the addition of the unclassified cases changed individual job groups. Some representative results were selected to illustrate several observed trends.

Three related Law Enforcement (LE) jobs identified in the original hierarchical clustering study included: LE Desk Sergeants (Grp 594), LE Patrolmen (Grp 785), and a group which performs both as Desk Sergeants and Patrolmen (Grp 921). In the original study, group size for these groups was 15, 12, and 90, respectively (see Figure 3). In the initial iteration of the nonhierarchical clustering process, both the Desk Sergeant and Patrolmen groups more than doubled in size, whereas the mixed group dropped in membership. It is reasonable to assume that the 22 members lost from the combined LE Desk Sgt/Patrol group were those performing more Desk Sergeant functions, since 22 new members appeared in the Desk Sergeant group (Grp 594). Note that once these 22 new cases were added, the Desk Sergeant group membership essentially stabilized--the number of cases, the mean within-group overlap, and the standard deviation remained about the same for iterations 1 through 6. Note also that the standard deviation for this larger group (N = 37) dropped considerably (from 9.2 to 7.3), which demonstrates the development of a more homogeneous group.

The composite LE Desk Sgt/Patrolmen group (Grp 921) attritted more of its membership with each iteration of the process. Its mean within-group overlap dropped slightly at each stage and its SD increased, indicating that the remaining cases were more heterogeneous than in the original seed group. Presumably, if additional iterations had been run, this group would essentially disappear.

The LE Patrolmen (Grp 785), on the other hand, first grew in membership and then shrank, while its mean within-group overlap steadily increased and its SD dropped. The group gained in membership from other groups not included in this comparison (there is a composite LE Patrolmen/Entry Controller group, and other possible contributors). The point is that even at the sixth iteration, this group had not completely stabilized, and additional runs might be needed to maximize the within- group overlap and minimize the within-group variance.

A set of three related Alert Area Security groups illustrates some additional trends (see Figure 4). All three groups involve tasks being performed by Security Policemen in controlling access to and guarding alert aircraft. Grp 283 appears to be the most meaningful of the three groups, in that its membership continued to expand through all six iterations, whereas the other two groups fluctuated. We might need to run additional iterations to see how large this group would become before the mean overlap is maximized and the S.D. is minimized. For the other two groups, we need to examine their job descriptions to determine how they differ from Grp 283 (that is, what makes them distinct groups). Only then can a judgment be made whether these two small, low-overlap groups should be retained.

DISCUSSION

Preliminary results strongly suggest that the nonhierarchical clustering process has considerable potential in refining the existing occupational analysis process. The drastic drop in the number of unclustered or unaccounted-for cases represents the greatest benefit, since the group job descriptions resulting from increased sample size should be much more stable. We need, however, to extend our analysis to include study of such variables as the core tasks for each group and the amount of core job time accounted for by such core tasks. With lower within-group variance and somewhat increased within-group overlap, the set of core tasks should account for a significantly greater proportion of total work time. Such a finding would have important implications for the identification of training requirements within a specialty, both for resident programs and on-the-job training programs. In addition, the more discretely defined job groups resulting from this type of analysis should have considerably more construct and content validity in the eyes of other occupational analysis users, such as functional managers in the manpower and personnel communities.

Clearly, further work needs to be done to define the proper techniques for using this type of system. Data displays should be developed which will permit an analyst to track what is happening to each group across iterations. Also needed is a method of setting aside stabile groups at the end of each iteration and then continuing to run the program on the remaining groups. The analysis process will have to be sufficiently flexible to permit multiple runs of the later iterations in order to optimize both the mean group overlap and standard deviation values. Such developments are planned over the next few months.

The system could also be used to cluster a transposed task-by-person matrix (Perrin et al., op cit) to refine Task Training Modules (TTMs) from coperformance data. Such an application would have considerable advantages in terms of reduced cost and multiple TTM membership for technical tasks that apply to more than one module.

ACKNOWLEDGEMENTS

Mr William J. Phalen, AFHRL/MOMM, developed the concept and wrote the specifications for the nonhierarchical clustering algorithm used in this research. Sgt Kevin Kumba, AFHRL Technical Services Division, wrote the computer programs involved, and Mr Lou Datko, AFHRL/MODJ (now of the USAF Manpower and Personnel Center, Randolph AFB, TX), contributed substantially to their refinement during initial testing.

REFERENCES

Alton, R.L. (1984, November). Occupational survey report, security police career ladders (AFSCs 811X0, 811X2, and 811X2A). Randolph Air Force Base, TX: USAF Occuational Measurement Center, Occupational Analysis Program.

Carpenter, J.B. (1974, March). Sensitivity of group job descriptions to possible inaccuracies in individual job descriptions (AFHRL-TR-74-6, AD-778 839). Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory.

Datko, L.M. (in preparation). Rating scales for officer occupational analysis. Brooks AFB, TX: Air Force Human Resources Laboratory. Draft Technical Report.

Pass, J. (1980, May). An empirical comparison of proximity measures for CODAP cluster analysis. Proceedings of the Third International Occupational Analysts Workshop. Randolph AFB, TX. USAF Occupational Measurement Center.

Pass, J.J., & Robertson, D.W. (1978). Sample size and stability of task analysis inventory response scales. Proceedings of the 20th Annual Conference of the Military Testing Association. Oklahoma City. U.S. Coast Guard Institute.

Perrin, B.M., Vaughan, D.S., Yadrick, R.M., & Mitchell, J.L. (1985, October). Defining task training modules: Coperformance clustering. Proceedings of the 27th Annual Conference of the Military Testing Association. San Diego, CA: Naval Personnel Research & Development Center.

Phalen, W.J., Weissmuller, J.J., & Staley, M.R. (1985, May). Advanced CODAP: New analysis capabilities. Proceedings of the Fifth International Occupational Analysts Workshop. Randolph AFB, TX. USAF Occupational Measurement Center.

```
1213 0055        *           2241 0085        *       *    1726 0082
0001-0055        *           0255-0339        *       *    0465-0546
39.1 52.2        *           46.3 51.4        *       *    43.1 48.7
    *            *                *           *       *         *
    *            *                *           *       *         *
1191 0058    2167 0268            *           *       *    1679 0089
0001-0058    0072-0339 * * * * * *            *       *    0465-0553
39.0 42.7    45.8 49.2                        *       *    42.8 47.9
    *            *                            *       *         *
    *            *                            *       *         *
1050 0060    1997 0310                        *       *    1510 0097
0001-0060    0072-0381 * * * * * * * * * * * *        *    0460-0556
37.7 46.6    44.8 48.2                                *    41.5 47.0
    *            *                                    *         *
    *            *                                    *         *
0796 0071    1758 0318                                *    1294 0109
0001-0071    0072-0389                                *    0460-0560
34.7 44.0    43.3 48.0                                *    39.8 45.6
    *            *                                    *         *
    *            *                                    *         *
    *        1137 0388                                *    1146 0124
    *        0072-0583 * * * * * * * * * * * * * * * *     0460-0583
    *        38.4 45.3                                     38.5 44.2
    *            *                                              *
    *            *                                              *
    *        1028 0512                                          *
    *        0072-0583 * * * * * * * * * * * * * * * * * * * * *
    *        35.6 42.3
    *            *
    *            *
0079 0583        *
0001-0583* * * * *
34.5 40.6
    *
    *
    *
```

Figure 1. Clustering of Security Police Time-Spent Ratings - 811XX

Iteration	No. of Cases Classified	Number Unclassified	Percent of Coverage
0 (input)	2238	1065	67.7%
1	3298	5	99.8%
2	3292	11	99.6%
3	3285	18	99.4%
4	3281	22	99.3%
5	3279	24	99.3%
6	3278	25	99.2%

Figure 2. Changes in the Number of Cases Classified

Group	Variable	Seed	Iteration 1	2	3	4	5	6
Grp 594 LE Desk Sergeant								
	Group Size	15	37	34	33	32	32	33
	No. Lost	-	-	3	2	1	0	0
	No. Gained	-	22	0	1	0	0	1
	Mean	-	44.6	42.5	42.7	42.4	42.5	42.6
	S.D.	-	9.2	7.4	7.6	7.4	7.3	7.3
Grp 921 LE Dsk Sgt/Patrol								
	Group Size	90	68	38	20	14	10	5
	No. Lost	-	22	30	18	7	4	5
	No. Gained	-	0	0	0	1	0	0
	Mean	-	57.4	56.2	56.9	55.3	54.4	49.2
	S.D.	-	7.2	7.4	8.2	9.9	10.5	10.9
Grp 785 LE Patrolmen								
	Group Size	12	38	55	90	84	67	51
	No. Lost	-	-	6	12	29	25	20
	No. Gained	-	26	23	47	23	8	4
	Mean	-	51.0	52.1	55.5	57.2	57.6	57.9
	S.D.	-	10.9	7.8	7.4	7.2	6.9	5.9

Figure 3. Changes in Several Law Enforcement Groups

Group	Variable	Seed	Iteration 1	2	3	4	5	6
Grp 283 Alert Area Sentry/Entry Controll r								
	Group Size	54	73	113	170	231	277	279
	No. Lost	-	-	4	5	10	17	34
	No. Gained	-	19	44	62	71	63	36
	Mean	-	44.2	46.2	47.9	48.3	49.1	49.8
	S.D.	-	10.7	9.9	9.7	9.5	9.6	9.4
Grp 361 (Same Title)								
	Group Size	14	45	46	34	33	31	30
	No. Lost	-	-	9	14	4	3	1
	No. Gained	-	31	10	2	3	1	0
	Mean	-	36.4	33.7	31.0	32.4	32.1	32.1
	S.D.	-	14.9	11.6	10.4	8.9	8.4	8.1
Grp 470 (Same Title)								
	Group Size	11	18	14	14	14	13	11
	No. Lost	-	-	4	0	0	1	2
	No. Gained	-	7	0	0	0	0	0
	Mean	-	41.4	39.3	40.7	40.7	40.1	38.6
	S.D.	-	17.7	11.7	10.3	10.3	10.5	10.3

Figure 4. Changes in Related Alert Area Security Jobs

On the Study of Differential Item Performance without IRT

Paul W. Holland
Educational Testing Service
Princeton, New Jersey 08541

1. INTRODUCTION

The problem of identifying items for which the performance of certain subpopulations -- often women and minorities -- is unusual and out of line with their performance on other items or test results has a substantial history. The book by Berk (1982) summarizes the state of the art as of 1980 and the work of Lord (1980), Scheuneman (1979), Shepard, et.al. (1981), among others are relevant. From a statistical point of view, modern methods for the analysis of multi-way contingency tables seem particularly appropriate to this problem and some suggestions for their use have been made, (Marascuillo and Slaughter, 1981). In this spirit, the present paper proposes the well-known method of Mantel and Haenszel (1959) for the analysis of 2×2×K contingency tables as an easily implemented, powerful technique for the measurement of the degree to which two subpopulations of examinees perform differently on a given test item. Modern references to the Mantel-Haenszel procedure include Breslow (1981), Hauck (1979), and Breslow and Liang (1982). The basis for the use of the Mantel-Haenszel (herein MH) procedure in the study of differential item performance is the fundamental notion of the need to compare *comparable* people when examining the relative performance of two groups of examinees on an item. This is the problem of matching and is discussed in section 2. Section 3 gives the relevant facts about the MH procedure while section 4 discusses various aspects of the MH procedure and related methods in the context of measuring differential item performance.

2. MATCHING VERSUS CONTROLLING FOR ABILITY

The need to "control for ability" is well established in the differential item performance literature. It is the fundamental basis for the proposed use of item response theory methods to study "item bias." Other methods, such as those of Scheuneman (1979), use test performance as a proxy for ability. The "delta-plot method," Angoff (1982), controls for ability indirectly by concentrating attention on the covariance between the item difficulty indices for the two groups rather than on their respective mean values.

In my opinion, the "need to control for ability" is an inadequate way to express a more fundamental idea. When we compare two subpopulations on *any* criterion, it is always important to be sure that only *comparable* members of the two groups are being compared. What constitutes *comparability* will depend on the problem at hand. In the study of differential item performance we are interested in learning something about a test item and how members of one subgroup (the "focal group") might react differently to it than do the members of another subgroup (the "reference group"). If our criterion is performance (i.e., right or wrong on the test item) then it is improper to compare the performance of reference and focal group members who differ in significant and measurable ways that are related to their performance on the item. Differential item performance means differences in performance on an item between focal and reference group members that is attributable to characteristics of the item and not to differences in characteristics of the groups of examinees.

When we confound both examinee characteristics and item characteristics and simply look at differences in the performance on an item of reference and focal group members we are measuring what is called <u>impact</u> rather than differential item performance. For example, comparing the proportion of reference and focal group members who give correct answers to a given item is a measure of the item's impact on the focal group relative to the reference group In measuring differential item performance members of the reference and focal groups are first divided into sets of examinees who are <u>matched</u> on relevant criteria before their performance on the item is compared. Examples of relevant matching criteria are: scores on related tests, schooling measures, and other group membership. In many practical settings, matching will be done on related test scores since these are both available and accurately measured.

The 2×2×K Table: For a given item, say item j, the data from the i^{th} matched group of reference and focal group members can be arranged as a 2×2 table:

	Right on item j	Wrong on item j	
Reference	A_i	B_i	n_{Ri}
Focal	C_i	D_i	n_{Fi}
Total	R_i	W_i	n_{+i}

(1)

For i=1,..., K = number of matched groups. In (1) A_i denotes the number of reference group members in the i^{th} matched group who answered item j correctly. B_i, C_i and D_i have corresponding interpretations. n_{Ri} and n_{Fi} denote the number of reference and focal group members, respectively, in the i^{th} matched group, while n_{+i} denotes the total number in the i^{th} matched group of examinees. R_i and W_i denote the number in the ith matched group who get the item right and wrong, respectively. Considered together these K 2×2 tables form one big 2×2×K table. There is one such 2x2xK table for each item being considered. It is worth emphasizing that once the criteria for matching have been selected, the 2×2×K table of data can be formed from samples of data from the reference and focal group members. It should also be emphasized that the choice of matching variables is important and will depend on the availablity, amount, and accuracy of data as well as on its relevance to item performance.

3. THE MANTEL-HAENSZEL PROCEDURE

In the i^{th} matched group, the odds that a reference group member gets item j correct is A_1/B_1, while the corresponding odds for a focal group member is C_i/D_i. The MH procedure measures the advantage (or disadvantage) on item j that reference group members have relative to their matched focal group colleagues by the ratio of these two odds. This gives us the odds-ratio estimate

$$\hat{\alpha}_i = \frac{A_i}{B_i} / \frac{C_i}{D_i} = \frac{A_i\ D_i}{B_i\ C_i} \ . \tag{2}$$

The $\hat{\alpha}_i$ estimate a population cross-product-(or odds-) ratio, α_i, for the i^{th} matched group.

The Mantel-Haenszel common-odds-ratio estimate is a weighted average of the $\hat{\alpha}_1$ that uses the following weighted formula:

$$\hat{\alpha}_{MH} = \frac{\Sigma\ \omega_i\ \hat{\alpha}_i}{\Sigma\ \omega_i} \ , \tag{3}$$

where

$$\omega_i = \frac{B_i\ C_i}{n_{+i}} \ . \tag{4}$$

Substituting (4) into (3) yields the usual formula for $\hat{\alpha}_{MH}$:

$$\hat{\alpha}_{MH} = \frac{\Sigma\ A_i\ D_i/n_{+i}}{\Sigma\ B_i\ C_i/n_{+i}} \ . \tag{5}$$

The Mantel-Haenszel estimate, $\hat{\alpha}_{MH}$, is the average factor by which the likelihood that a reference group member gets item j correct exceeds the corresponding likelihood for comparable focal group members. (Likelihood is measured by the odds of getting item j correct). For example, if $\hat{\alpha}_{MH} = 1$ then reference and focal group members are, averaging across all the matched groups, equally likely to be correct on the item. When $\hat{\alpha}_{MH} > 1$ then the reference group has the advantage whereas when $\hat{\alpha}_{MH} < 1$ the focal group has the advantage.

Associated with the estimate $\hat{\alpha}_{MH}$ is a one-degree-of-freedom chi-square test of the hypothesis that all of the population cross-product ratios in all of the 2×2 layers of the 2×2×K table are unity (i.e., $\alpha_i = 1$ all i). This test is given by the formula:

$$\chi^2_{MH} = \frac{\left(\left|\sum_i A_i - \sum_i \mu_i\right| - \frac{1}{2}\right)^2}{\sum_i \sigma^2_1} \tag{6}$$

where

$$\mu_i = E(A_i | \alpha_i = 1) = \frac{n_{Ri}\ R_i}{n_{+i}} \tag{7}$$

and

$$\sigma_i^2 = \mathrm{Var}(A_i | \alpha_i = 1) = \frac{n_{Ri}\ n_{Fi}\ R_i\ W_i}{(n_{+i})^2(n_{+i}-1)} . \tag{8}$$

The χ^2_{MH} from (6) will be large if $\hat{\alpha}_{MH}$ differs from 1.0 significantly in either direction. Thus, this test will detect differential item performance that favors either the reference or the focal group.

4. DISCUSSION

The MH procedure is closely related to log-linear model procedures for estimating a constant two-way interaction across a series of 2×2 tables (see Bishop, Fienberg, and Holland, 1975). In practical terms, $\hat{\alpha}_{MH}$ is usually nearly identical to estimates of the common-odds-ratio that involve complicated iterative procedures. While the formula foi $\hat{\alpha}_{MH}$ is a simple weighted average of the sample odds-ratio α_i, it has been shown (Breslow, 1981) that, over the range of values relevant to this application of the MH procedure, $\hat{\alpha}_{MH}$ is nearly optimal as an estimator. In other words, no other estimate of the common-odds-ratio can have a substantially smaller variance. The chi-square test based on χ^2_{MH} is of high power because it is concentrated into a single degree of freedom rather than dissipated across several degrees of freedom.

If there is more than one pair of groups that could serve as the reference and focal group in an analysis then values for $\hat{\alpha}_{MH}$ and χ^2_{MH} can be computed for all such pairings.

The parameter $\Delta = -2.35 \ln(\alpha)$ is (approximately) in the scale of differences in delta-units of difficulty where delta-units are those used by ETS in their normal item analysis procedures. This transformation can be used to put $\hat{\alpha}_{MH}$ values into units which are familiar to those who use the delta-scale in test construction and analysis:

$$\hat{\Delta}_{MH} = -2.35 \ln(\hat{\alpha}_{MH}). \tag{9}$$

Thus, $\hat{\Delta}_{MH} = -1.0$ means that the focal group found the item one delta-unit harder than did comparable members of the reference group. The parameter $\hat{\Delta}_{MH}$ is similar to an average shift to the right of $-\Delta_{MH}$ in the ICC of the focal group relative to the ICC of the reference group in an IRT model (as estimated by LOGIST).

The MH procedure can easily be expanded to include an analysis of distractor choice for multiple choice tests. For five-choice responses the 2x2 table in (1) is replaced by the followin 2x6 table

Response on item j

	A	B	C*	D	E	Omit	Total
Reference							n_{Ri}
Focal							n_{Fi}
Total							n_{+i}

C* is correct answer, for example. Then the MH procedure is applied to the five 2×2 tables formed by juxtaposing the column for the correct answer with a column for one of the five ways of producing incorrect answers. E.g.,

	C*	A		C*	B		C*	Omit
Reference								
Focal						....		

This yields five MH cross-product estimates and five chi-square tests for each item. In some cases these may be used to see if a significant value of $\hat{\alpha}_{MH}$ is due to a single type of incorrect answer.

There are a number of important research issues that need to be addressed in the use of the MH procedure in the study of differential item performance. What aspects of the criteria for matching examinees seriously affects $\hat{\alpha}_{MH}$ in practical settings -- the reliability of the criteria, the fineness of the matching, the use of other examinee attributes, etc.? How stable are the values of $\hat{\alpha}_{MH}$ across different examinee populations? What are the relationships between the values of $\hat{\alpha}_{MH}$ and other statistical indices used to construct tests -- i.e., difficulty and discrimination? How should values of $\hat{\alpha}_{MH}$ for several pairs of reference and focal groups be combined for the same test item?

The MH procedure promised to be a relatively inexpensive and yet statistically powerful technique for identifying test questions that are potentially "biased" or unfair in some way to identified groups of examinees. ETS is currently embarked on a variety of research projects to see how to best use this tool for such purposes.

References

Angoff, W. (1982) "Use of difficulty and discrimination indices for detecting item bias" in Berk, R. (ed.) *Handbook of Methods for Detecting Test Bias*. Baltimore and London: Johns Hopkins University Press.

Berk, R. (Ed.) (1982) *Handbook of Methods for Detecting Test Bias*. Baltimoie and London: Johns Hopkins University Press.

Breslow, N (1981) Odds ratio estimates when the data are sparse. *Biometrika*, *68*, 73-84.

Breslow, N. and Liang, K. (1982) The variance of the Mantel-Haenszel estimator. *Biometrics*, *38*, 943-952.

Hauck, W. (1979) The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. *Biometrics*, *35*, 817-819.

Lord, F. (1980) *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.

Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Marascuillo, L. and Slaughter, R. (1981) Statistical procedures for identifying possible sources of item bias based on chi-square statistics. *Journal of Educational Measurement*, *18*, 229-248.

Scheuneman, J. (1979) A new method for assessing bias in test items. *Journal of Educational Measurement*, *16*, 143-152.

Shepard, L., Camilli, G., and Averill, M. (1981) Comparisons of six procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, *6*, 317-375.

REVIEWING AN ITEM POOL FOR ITS SENSITIVITY TO THE CONCERNS OF WOMEN AND MINORITIES: PROCESS AND OUTCOMES

Susan Wilson Kershaw
Howard Wainer
Educational Testing Service
Princeton, New Jersey

Why a Sensitivity Review

An item pool for a computerized adaptive version of the ASVAB has been developed and must be subjected to numerous editing and review processes before being made operational. Among these processes is a sensitivity review of the entire item pool for material that may be potentially offensive to minority groups and/or women. During a sensitivity review, all test items are screened by trained sensitivity reviewers to ensure that the test is as free as possible from perceived bias and offensiveness. In order for each examinee to perform at his or her optimum level of ability, it is necessary to eliminate any material that may convey negative or distracting messages to a particular subgroup of the test-taking population. This type of review is particularly important in the development of the CAT-ASVAB since the test-taking population will be drawn broadly from many cultural and socio-economic backgrounds and because it is difficult to forsee what particular combination of items will be presented to a given examinee.

The Sensitivity Review Process

ETS Objectives

The overall objective of the sensitivity review process at ETS is to eliminate any material from tests that may be potentially offensive or inappropriate for identifiable subgroups of the test-taking population, or that reinforces negative attitudes toward these subgroups. The ETS test evaluation criteria consist of a general set of criteria that can be applied to all people, and specific criteria that are particularly relevant for five subgroups--Asian Americans, Black Americans, Hispanic Americans, Native Americans, and women.

Test Evaluation Critieria

The following test evaluation criteria have been excerpted from ETS Test Sensitivity Review Process (Hunter & Slaughter, 1980). For a more detailed description of the sensitivity review process, refer to their document.

A. Definitions

Group reference items reflect the multicultural nature of our society. There are two basic classes of subgroup reference items: representational items and substantive items.

1. Representational items are those in which references are made to minorities and women, but where the subject matter content of the test is intended to measure factors unrelated to such groups. Reading passages, charts and graphs, pictures, cartoons, and writing ability times are most easily adapted to this purpose.

2. Substantive items are those designed to directly measure knowledge about a population group. Such items might ask about the role of the Black church in Black life, the migration patterns of Chicanos in North America, or the factors that have led to the increasing numbers of women enrolling in graduate-level programs.

B. Evaluation Requirements
All test items will be reviewed and identifiable group reference items will be evaluated from the following perspectives:

1. Cognitive/Affective
These two dimensions should be considered when reviewing all group reference items. The cognitive dimension deals with the factual basis of item content and the affective dimension reflects the positive or negative feelings the item may evoke on the part of group members.

2. Controversial Material
Highly controversial issues, such as legalized abortion or hypotheses about genetic inferiority, must not be included in any test item unless they are both relevant and essential to effective measurement.

3. Examinee Perspective
All group reference items should be reviewed from the perspective of test takers who do not have access to an answer key. When an examinee must know the correct key to prevent an item from reinforcing negative attitudes, the item should be rejected. This situation most often arises when an item writer attempts to mislead test takers who hold stereotypical beliefs by using one or more distractors that are obvious stereotypes (obvious to the writer). In these cases, examinees who select a sterotype as the correct response are not routinely informed that their response was incorrect. Such practice may reinforce their belief in the legitimacy of the stereotype. This is particularly likely for examinees who subsequently receive a high score.

4. Stereotyping
All ETS tests will be reviewed to ensure that they do not contain language or symbols that reinforce sterotypes judged to be generally offensive. While it is clear that no ETS test intentionally contains blatant references to such offensive stereotypes, the potential exists for including material in a test that could be perceived as manifestations of such sterotypes by group members who take the test. Offensive stereotypes generally imply an inferiority or deficiency

of one of more groups in physical characteristics (for example, height, weight, attractiveness, strength) and psychological characteristics (for example, intelligence, ethics, emotions, behavioral patterns) generally regarded as desirable by the majority culture.

5. Caution Words and Phrases
Through experience, those currently reviewing tests for sensitivity have learned that certain key words and phrases are more likely to accompany sensitive material. While for the most part use of these words and phrases in tests are proper and legitimate, they will receive extra attention by sensitivity reviewers when any two or more words are found in an item because they indicate an increased potential for including offensive material. Examples of such words are lower-class, discrimination, race, and housewife.

6. Special Review Criteria for Women's Concerns
During the past decade, a great deal of progress has been made in identifying numerous manifestations of sexism in our society. This progress has included several efforts to identify and eliminate sexism in written language. Two notable efforts in this direction were the ETS Guidelines for Sex Fairness in Tests and Testing Programs and the McGraw-Hill Guidelines for Equal Treatment of the Sexes. Much of the material contained in these documents has been recast to serve as evaluation criteria. Sensitivity reviewers will ensure that all ETS tests are in compliance with these criteria.

7. Underlying Assumptions
An underlying assumption is a subtle secondary premise in test material that reflects an individual's ethnocentric beliefs.

8. Context Considerations
Reviews for sensitive material frequently require judgments relative to the context in which it is presented. In some cases, it may be necessary to measure one's knowledge of a domain by using material that some groups may feel is sensitive. There are four areas in which this occurs with some frequency: historical domain, literary domain, legal domain, and psychological domain.

Formal Structure and Procedures

The formal test sensitivity review is conducted by trained sensitivity reviewers who are often members of the Test Development staff at ETS. The review procedure involves an evaluation of the test by the sensitivity reviewer in accordance with the standard test evaluation criteria. Any comments and recommendations by the sensitivity reviewer are documented on the Test Sensitivity Review Form and returned to the test assembler. The test assembler discusses recommended changes with the sensitivity reviewer. If agreement is reached concerning changes, both sign and date the review form. If agreement cannot be reached, the matter is referred to the area Test Development Director and, if necessary, to an arbitration committee for resolution.

The sensitivity review of the CAT-ASVAB item pool followed the standard ETS procedures except for the final steps in which actual modifications are made to unacceptable test items. Rather, recommendations for changes by a minority panel were submitted for Government review in a Confidential Appendix to the ETS report. The Sensitivity Review of the CAT-ASVAB Item Bank (Wilson & Wainer, 1985.)

The CAT-ASVAB Sensitivity Review Board

The Sensitivity Review Board for this project was composed of twelve nationally renowned educators and test developers representing minority group concern. Asian American, Black American, Hispanic Americans, and both gender groups were represented on the panel.

The following ETS criteria were considered in selecting sensitivity reviewers:

A. Ability to perceive offensive material
B. Ability to review tests from multiple perspectives, not simply from the viewpoint of one group, or social/political perspective.
C. Coverage among the reviewers of key subject areas such as humanities and social sciences.

The Meeting

The CAT-ASVAB Sensitivity Review was conducted over a two-day period. The meeting began with a brief orientation session to familiarize panel members with the purpose, plan, and procedures for the sensitivity review. Each panel member had also received a booklet containing copies of relevant sections from ETS Test Sensitivity Review Process (Hunter & Slaughter, 1980) to review prior to the meeting. The ETS test evaluation criteria, guidelines for recognition of unacceptable sterotypes, caution words and phrases, and special review criteria for women's concerns were detailed in this booklet and were to be referred to by panel members as guidelines for item review and reporting of results.

Following the task orientation session, panel members were assigned to review CAT subtests in review teams. Each review team consisted of one male and one female of different ethnic affiliations, to ensure that the sensitivity review process was balanced across sex and ethnicity as much as possible. The CAT-ASVAB item bank consists of 2118 test items, representing nine different content areas: Electronics Information, Mechanical Comprehension, Shop Information, Automotive Information, Mathematics Knowledge, Arithmetic Reasoning, Paragraph Comprehension, Word Knowledge, and General Science. The 2118 times were divided up approximately equally among the reviewers, with each panel member reviewing between 334-363 items. Panel members were assigned to subtests in pairs so that each item would be reviewed by two people.

Description of Sensitivity Review Outcome

The following table was used to summarize the reviewers`conclusions:

	ACCEPTABLE		UNACCEPTABLE	
Subtest	As Is	With Revision	Potentially Offensive	Item Construction Flaw
Automotive Information				
Mathematics Knowledge				
Shop Information				
Mechanical Comprehension				
Electronics Information				
Arithmetic Reasoning				
General Science				
Word Knowledge				
Paragraph Comprehensive				
Totals				

Among the "acceptable" items some were judged to be acceptable only after modification. "Unacceptable" items were found to be so for two quite different reasons. Sometimes this was because of their potential offensiveness to some sub-groups of examinees. A second reason for finding an item unacceptable was because of a formal flaw in the item construction. This usually yielded the situation in which the number of correct responses to an item was unequal to one. Although some of the screened items had such flaws we left their description to another account; that aspect of the item pool is outside of the purview of this project. Therefore, "unacceptable" items in the CAT-ASVAB item bank, refer to items which were found to be unacceptable from a sensitivity review perspective.

Due to the confidential nature of this project, the results of the sensitivity review as reported by panel members must be treated in a secure manner. Therefore, specific details on the test items and the sensitivity reviewers' comments and recommendations cannot be presented in this paper, but can be found in the Confidential Appendix to the Wilson & Wai' er (1985) report.

However, the panel members did make some general comments and suggestions concerning the further improvement of the sensitivity review process itself.

In line with these and other suggestions we recommend in subsequent reviews the following additions/changes:

1) extend the review process sufficiently to allow a more extensive training period. This is important initially to fully inform the panel about the special aspects of CATs.
2) a slight modification of the review form to more closely correspond with the summary shown in the Table above.
3) have available to the review panel an information sheet on each item that contains:
   a) the item key,
   b) the content specifications that each item fulfills,
   c) the statistical results of the item's pretesting,
   d) comments of previous reviews (if any).

The availability of item history in a form as described in 3 above is *pro forma* in traditional test development. Although our recommendations deal with improving the ease and quality of the item review process, it has been the experience of test developers that such a procedure is quite useful throughout the item development process.

References

Hunter, R.V. & Slaugher C.D. (1980, July).
EIS test sensitivity review process.
Princeton, NJ: Educational Testing Service.

Wilson, S. & Wainer, H. (1985, September).
The sensivitivity review of the CAT-ASVAB Item Bank. (Contract No. N66001-84-D-0083, D.O. 7J03). Submitted to Navy Personnel Research and Development Center.
Princeton, NJ: Educational Testing Service.

MEASUREMENT PRECISION AND "RELIABILITY": SOME CONSIDERATIONS OF METRICS AND STOPPING RULES IN CAT

David Thissen
University of Kansas

Reliability has been defined as "the degree to which a test is free from error." While that definition is certainly attractive, it is not particularly mathematically tractable. In a more practical vein, classical test theory defined reliability as

$$rho^2{}_{YT} = rho_{XX'} = \frac{\text{"True score variance"}}{\text{"Observed score variance"}} \quad ; \qquad (1)$$

Since the error of measurement in classical theory is independent of true score, and has a constant variance, the form in equation (1) provides a standardized description of the relative size of the (lack of) error of the test score. It is also the "percentage of variance" in the observed scores which is due to true-score variation, and a predicted value of the correlation between observed scores on parallel forms. Here we will consider the extension of that concept to an IRT-scored Computerized Adaptive Test (CAT). This is not as hopeless as it may appear; CAT is sufficiently flexible that it can be made to produce results which meet the assumptions of classical test theory (which is something classical test scores never do).

The crucially wrong assumption of classical test theory, with respect to test scores, is that the error variance of all test scores is the same. The restriction of range of raw test scores alone makes that impossible, except in degenerate cases. However, in a CAT system, it is possible that all of the "test scores" could have the same error variance. Here we use the notation est(θ) as a generic term for an estimate of θ; est(θ) is some estimate of the location (like a mode or mean) of the likelihood over θ for a response vector, in which the likelihood is computed as the product of the trace lines for the responses. We use the notation se^2 for the variance of that likelihood.

Equal se^2 CAT

It would be difficult to construct a test giving constant se^2s: the item pool would have to be so large and "broad" that, regardless of the ability of the examinee, the system could continue to administer increasingly easy or difficult items until the person gave sufficient numbers of both correct and incorrect responses to have a likelihood over θ with a constant variance. The CAT "stopping rule" would be a fixed value for the error variance for θ;

calibration of an item pool with the necessary properties is possible, but beyond the scope of this paper.

Under such a CAT system, strong parallels exist between the classical theory and the IRT CAT system, as summarized in Table 1.

TABLE 1

Classical Test Theory	IRT CAT (fixed se^2 stopping)	
"True score"=T; var(T)	θ; var(θ)=1	
"Error"=E; var(E)	error; var(error)=$se^2(\theta)$	

"Observed score"=X;
X = T + E,
(independent),
var(X)=var(T)+var(E).
	est(θ);
	est(θ) = θ + error,
	(independent),
	var(est(θ))=$1+se^2(\theta)$.	

rho^2_{XT} = var(T)/var(X) from (1).	$\text{rho}^2 = 1/[1+se^2(\theta)]$ from (1).	(2)

Using slightly different notation. Samejima (1977) notes that rho^2 in equation (2) represents the reliability, or expected correlation between parallel test scores, for an IRT-scored test. She notes that this form is impossible or deceptive if $se^2(\theta)$ varies as a function of θ; However, in a CAT system, it is possible that $se^2(\theta)$ is a constant and equation (2) is precisely correct.

Further parallels may be drawn between the classical theory and IRT CAT systems. In classical test theory, the best prediction of the true score is the so-called Kelley (1947) "regressed" estimate, which is

$$\text{rest}(T) = \text{rho}^2_{XT}\ X \qquad (3)$$

if X, T, and E all have mean zero and var(T)=1. If one "bends" IRT slightly, by claiming that the likelihood over θ is exactly (instead of approximately) Gaussian with mean est(θ) and variance $se^2(\theta)$, then

$$\text{rest}(\theta) = \text{rho}^2\ \text{est}(\theta) \qquad (4)$$

is exactly equal to either the Bayes modal estimate of θ or the expected a posteriori estimate of θ computed with a population distribution which is N(0,1). The mode and the mean become the same when "everything is Gaussian"; that is why we are using the generic notation est(θ) and rest(θ) for the un-regressed and regressed estimates of θ respectively.

Note that in the context of an IRT CAT system using a fixed (equal) $se^2(\theta)$ stopping rule, the concept of reliability is not "dead". It is, as a matter of fact, enhanced: it is "righter" than it ever was under the classical theory. It is "righter" because it is based on the idea of equal error variance, which was never true for classical test scores, but which can be *made* true in *CAT*. The value of rho is both a prediction of the (hypothetical, "washed-brain") test-retest correlation between estimates of θ, and it is the "regression" or "shrinkage" constant in the Bayes estimates of θ.

Lord and Novick (1968) distinguish among three different "errors" in their discussion of the classical theory of reliability and those distinctions are useful here as well. The three error variances discussed by Lord and Novick, and their *IRT* counterparts in equal-$\underline{se}^2$ *CAT* systems are tabulated in Table 2.

TABLE 2

	Classical Test Theory	*IRT*
Variance of Measurement		
	$\underline{var}(E) = \underline{var}(X)[1-\text{rho}_{XX'}]$	$\underline{se}^2(\theta) = \underline{var}[\underline{est}(\theta)][1-\text{rho}^2]$
Variance of Estimation		
	$\underline{var}(T)[1-\text{rho}_{XX'}]$	$1/[1+1/\underline{se}^2(\theta)] = \underline{var}[\theta][1-\text{rho}^2]$
Variance of Prediction		
	$\underline{var}(X)[1-\text{rho}^2_{XX'}]$	$1/[1+1/\underline{se}^2(\theta)] + \underline{se}^2(\theta) = \underline{var}[\underline{est}(\theta)][1-(\text{rho}^2)^2]$

The discussion in Lord and Novick (1968, p. 67-8) is cast in terms of standard errors, instead of the variances given above; the standard errors are the square roots of these variances. The standard error of estimation is the square root of $\underline{se}^2$; the standard deviation of the Bayesian posterior over θ is the square root of the variance of estimation, and the standard error of prediction of a subsequent estimate of θ from the regressed estimate is the square root of the variance of prediction. Those three values are not generally the same, which is one of the things wrong with defining reliability as "the degree to which a test is free from error": which error?. But they are all functions of rho^2.

It is also possible and reasonable in this context to consider the reliability of composite scores C obtained as linear combinations of estimates of two or more θs. The simplest example is a composite of two tests. If

$$\underline{est}(C) = \underline{est}(\theta_1) + \underline{est}(\theta_2)$$

$$= (\theta_1 + error_1) + (\theta_2 + error_2),$$

and the θs are correlated r with each other and the errors are uncorrelated with everything, then

$$\underline{var}(\underline{est}(C)) = \underline{var}(\theta_1) + \underline{var}(\theta_2)$$

$$+ \underline{var}(E_1) + \underline{var}(E_2) + 2r\ \underline{sqrt}[\underline{var}(\theta_1)\underline{var}(\theta_2)]$$

and

$$\underline{var}\ (C) = \underline{var}(\theta_1) + \underline{var}(\theta_2) + 2r\ \underline{sqrt}[\underline{var}(\theta_1)\underline{var}(\theta_2)],$$

so

$$rho_{CC'} = \underline{var}(C)/\underline{var}(\underline{est}(C))$$

as per equation (1). Generalization to composites of more than two scales is obvious.

Other metrics

If the test scores are to be reported in some metric other than the θ-metric, such as "expected raw score" (E[score]), and composites are to be computed as linear combinations of scores in that metric, then it is obligatory that reliability be reported in the transformed metric. The simplicity described above can still be obtained in a CAT system if the stopping rule is based on equal $\underline{se}^2$s in the E[score] metric. This is somewhat deceptive psychologically: equal $\underline{se}^2_{E[score]}$ represents very unequal $\underline{se}^2$s in the θ metric, as E[score]s at the extremes have very small variances when the associated estimates on the ability dimension still have very large variances. But rho^2 would have its correct meaning in the metric in which the test is being described.

Unequal $\underline{se}^2$ stopping rules

CAT systems may be implemented with stopping rules other and those which give constant $\underline{se}^2$s. Such stopping rules clearly produce unequal $\underline{se}^2$s for different values of $\underline{est}(\theta)$. What happens to the concept, and the computation, or rho^2 under such alternative stopping rules?

There are several problems. The most obvious is that rho^2 cannot be computed as in equation (2), since that requires a constant value for $\underline{se}^2(\theta)$, which does not exist. Samejima (1977) suggests that for some purposes equation (2) may be replaced by

$$\text{rho}^2 = 1/(1+\text{average}(\underline{se}^2(\theta))) \qquad (5)$$

in which the error variances are averaged over the distribution of θ for the group being tested. Several further problems immediately arise. One is that the estimate of "reliability" obtained with equation (5) depends on the distribution of θ used in the "average"; if it is the theoretical population distribution, it depends on the theory, and if it is an empirical distribution, it depends on the sample.

A second problem is that it is not clear what purpose rho^2 computed using equation (5) could serve. It remains true that it is an estimate of the correlation that would be obtained between parallel tests. However, it is not an indicator of the size of the error variance for any particular $\underline{est}$(θ). Here, all $\underline{est}$(θ)s have different $\underline{se}^2$s and rho^2 computed from (5) reflects only their average. An average of a set of numbers which are known to vary systematically is not particularly useful. It would be much more useful, if the goal was to characterize the size of the errors of measurement, to abandon the concept of reliability altogether and report the sizes of the $\underline{se}^2$s as a function of θ, in either graphical or tabular form.

We noted above that, in the equal-$\underline{se}^2$ situation, the value of either $\underline{se}^2$ or rho^2 is indicative of the amount of "shrinkage" induced by the population distribution in regressed estimates of θ. Unfortunately, rho^2 from (5) is not informative about shrinkage. Indeed, there is rather serious theoretical problem underlying this loss-of-usefulness of the concept of reliability. When the $\underline{se}^2$s of $\underline{est}$(θ)s are unequal, each is regressed a different amount (proportional to its own variance) in Bayesian estimation schemes. That means that the population distribution has differential effects on different individuals, which may be unfair.

Consideration of the reliability of composite scores is complicated extraordinarily by unequal $\underline{se}^2$s at the individual score level. Even in the simplest case in which a composite is the sum of two component scores, if those two component scores may have different $\underline{se}^2$s associated with different values, then a single value of the composite could have a wide variety of $\underline{se}^2$s. This is true because a single value of the composite could be produced by different combinations of θ_1 and θ_2, in which each θ could have different $\underline{se}^2$s which are combined to give the $\underline{se}^2$ for that particular way of obtaining that score on the composite.

While it is true that the distribution of $\underline{se}^2$s possible for each value of the composite could be averaged to give a single average($\underline{se}^2$) for that value of the composite score, that could be a fairly deceptive value when applied to any particular instance of that composite score. And each value of the composite score would still have a different $\underline{se}^2$; those would have to be averaged again to produce a single "reliability coefficient" for the composite score as a whole. It is possible to do this numerically, or actually compute empirical test-retest reliability coefficients; but what is the use of such multiply-averaged numbers when we know that the average value applies to no particular test score?

Conclusion

There is a great deal to be said for the concept of reliability: it predicts test-retest correlation, describes the error variance of test scores, and specifies the amount of "shrinkage" inherent in Bayes estimators. There is even more to be said for the construction of tests which unambiguously have some specific reliability. It appears that such tests must be administered by CAT systems with a constan+ $\underline{se}^2$ stopping rule.

References

Kelley, T.L. (1947). Fundamentals of statistics. Cambridge: Harvard University Press.

Lord, F.M. & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

Samejima, F. (1977). A use of the information function in tailored testing. Applied Psychological Measurement, 1, 233-247.

Multi-dimensional Performance Measurement

Lawrence M. Hanser and Jane M. Arabian
U.S. Army Research Institute[1]

Lauress Wise
American Institutes for Research

Introduction and Background

This paper is based on data collected for the large Army personnel research project titled "Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Project A" (Eaton, Hanser, & Shields, 1985). This project was conceptualized and planned during the 1980 to 1981 time period, and a contract was signed with the prime contractor, Human Resources Research Organization (HumRRO), in 1982. It is being conducted jointly by scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI), HumRRO, the American Institutes for Research (AIR), and Personnel Decisions Research Institute (PDRI).

Early in the planning for Project A, it was recognized that a large proportion of the research would have to be devoted to criterion development. Plans called for the development of several different measures of performance: (a) tests of hands-on performance, (b) paper and pencil tests of job knowledge, and (c) ratings of typical performance. Each of these broad categories of criteria were further subdivided. Hands-on tests included tasks which were specific to each Military Occupational Specialty (MOS) as well as tasks common to all MOS. Two kinds of paper and pencil tests were constructed: (a) to emphasize the content of formal school training, and (b) to emphasize MOS-specific task performance. Rating forms were constructed both for MOS-specific task performance as well as for non MOS-specific Army-wide performance that we have labelled broadly as "soldiering."

The initial impetus for developing such a comprehensive set of criterion measures was largely a function of our underlying theory of performance measurement. This underlying theory states rather simply that performance in a job is multi-dimensional, and that it is not possible to capture that multi-dimensionality using only one measurement method. A method of measurement may be intrinsic to some tasks. For example, having the requisite knowledge of how to take a person's blood pressure may not be the same as actually being able to perform the task accurately, yet both are important. An individual may score high on a paper and pencil test on this task, but might not score as high on a hands-on test of this task. In order to be successful in performing this task on the job it requires: (a) the knowledge of how to do the task, (b) the physical skills to perform the task, and (c) the motivation to do it. Or to put it in another well known way: performance = f(ability x motivation).

[1] The views expressed in this paper are those of the author and do not necessarily reflect the view of the U.S. Army Research Institute or the Department of the Army

Because of the complexity of the criterion space being measured in this project it is extremely important that it be fully understood prior to choosing a final set of predictors and recommending changes to the Army's selection and classification procedures. Several recent papers by project scientists have begun to address the issues associated with criterion development (c.f., Borman, White, Gast, & Pulakos, 1985; Campbell & Harris, 1985; Rumsey, Osborn, & Ford, 1985). Borman et al. constructed and tested a path model of supervisory and peer ratings to examine how each are related to other measures of performance. They found that both job knowledge and hands-on task proficiency are related to ratings, with the dominant path between ratings and hands-on proficiency. They conclude, however, that "... for the most part different methods of measuring job performance yield quite different results." Campbell and Harris describe the results of attempting to interpret criteria using a group of "concerned psychologists." They also present a "working model of job performance for the domain of skilled jobs." In examining the correlation matrices of hands-on and job knowledge tests and rating scales, they state "... the methods correlate more highly within themselves than they do across measures." Rumsey et al. examine the relationships between job knowledge tests and hands-on tests of job proficiency. In each of these papers, a central theme is the multi-dimensionality of performance and the importance of using different measurement methods to capture performance adequately.

The intent of this paper is to further explore the criterion space measured in Project A. Previous research has focused on aggregate measures of performance such as total scores on hands-on or paper and pencil tests or average ratings across several dimensions. In this paper we focus on task level measures in order to begin to understand better the relationships between kinds of tasks and methods of measuring performance on them. Through this we hope to gain a better understanding of the method variance associated with measures of task performance.

Method

Subjects

Date reported in this paper were collected in 1984 as part of field tests of the criterion measures developed by Project A scientists. Participants included first tour soldiers in two Army MOS: (a) 178 Infantrymen (MOS 11B) and (b) 167 Medical Specialists (MOS 91A). A complete description of the data collection methods can be found in Campbell and Harris (1985).

Variables

Percent correct steps per task and average supervisory rating per task provided the major variables used in these analyses. Percent correct scores were obtained on both hands-on and written tests. For each MOS reported here, approximately 15 tasks were scored using all three measurement methods: (a) hands-on performance, (b) multiple choice paper and pencil test, and (c) average supervisory rating of task performance. Approximately 15 additional tasks per MOS were tested in the paper and pencil test, and these were also included in the analyses. In addition, total score on a paper and pencil test focusing on training course content, average supervisory rating on overall performance, and Armed Services Vocational Aptitude Battery (ASVAB) subtest standard scores were included. This resulted in a total of approximately 71 variables per MOS to be included in these analyses. Although these are a relatively small number of subjects given the number of variables, the limits of analysis are a function of the number of factors extracted. These sample sizes will support the extraction of a maximum of five to seven factors per MOS.

Analyses

Though some "feel anxious in the presence of too many partial or semi-partial correlations" (Campbell & Harris, 1985), we decided to explore these data using factor analysis. Our specific plans were as follows: (a) extract a set of oblique factors for each MOS, (b) examine the inter-factor correlation matrices, and (c) examine the patterns of loadings within and across MOS. We used a principal axis solution with an iterative solution for the communalities and a Promax rotation. We decided on the number of factors to extract based on an inspection of the scree and interpretability of various solutions. In order to conserve space, descriptive statistics and reliabilities are not reported here. They are, however, available elsewhere (Borman et al., 1985; Campbell & Harris, 1985; Rumsey et al., 1985).

Results and Discussion

The data on the Medical Specialists yielded a five factor solution. Table 1 shows the oblique solution. Variables reported in the table are limited to the three highest loading on any factor, any variable with an absolute loading of greater than .30 on a cross-method factor, and any variable with loadings greater than .30 on two or more factors.

Table 1. Rotated Factor Pattern (STD REG COEFS)

I	II	III	IV	V	
80	.	.	.	.	Rating:Splint Suspected Fracture <Supv>
77	.	.	.	.	Rating:Put on Field/Pres Dressing <Supv>
75	.	.	.	.	Rating:Perform CPR <Supv>
58	.	.	-35	.	Rating:Measure/Record Respir. <Supv>
53	.	30	.	.	Rating:Measure/Record Pulse <Supv>
.	57	.	.	.	P&P:D9-Replace Filters in M17 Mask
.	51	.	.	.	P&P:I4-Measure/Record Respirations
.	47	.	.	.	P&P:I9-Estab/Maintain a sterile fld
.	43	.	.	.	HO: A4-Put on Field/Pres Dressing
.	34	.	.	.	HO: A9-Init a Field Med Card
.	.	68	.	.	ASVAB SUBTEST SCR-Arithmetic Reasoning
.	.	57	.	.	ASVAB SUBTEST SCR-Math Knowledge
.	.	52	.	.	ASVAB SUBTEST SCR-Coding Speed
.	.	49	.	.	P&P: I6-Assemble Needle & Syringe
.	.	49	.	.	P&P: K2-Draft/Fire TPR Charts
.	.	42	.	.	P&P: A6-Open Airway
.	.	40	.	.	P&P: I7-Change a Sterile Dressing
.	.	41	32	.	School: All Items
.	.	.	76	.	ASVAB SUBTEST SCR-Auto/Shop
.	.	.	71	.	ASVAB SUBTEST SCR-Electronics Information
.	.	.	59	.	ASVAB SUBTEST SCR-Mechanical Comprehension
.	.	.	37	.	P&P: G3-Vehicle Recognition
.	.	.	.	68	HO: I3-Measure/Record Pulse
.	.	.	.	51	HO: I9-Est/Maintain Sterile Field

.	.	.	.	47	HO: I4-Measure/Record Respir.
.	.	.	33	35	HO: AB-Splint Suspected Fracture

As expected, there are strong method factors, with little overlap of variables across method factors. Note, however, that two ratings overlap with the ASVAB factors, and one of the hands-on tasks overlaps with an ASVAB factor. Two hands-on tasks have loadings greater than .30 on Factor II, the paper and pencil job knowledge test factor. Several of the job knowledge test tasks load on the two ASVAB factors. Also, ASVAB splits into two factors, a math/speed factor and a technical factor. Table 2 provides the factor correlations.

Table 2. Inter-Factor correlations

	I	II	III	IV	V
I	100	1	7	-11	17
II	1	100	15	27	-2
III	7	15	100	-6	19
IV	-11	27	-6	100	-8
V	17	-2	19	-8	100

Not surprisingly, the paper and pencil job knowledge test factor, Factor II, and an ASVAB factor, Factor IV, have the highest correlation. Note, however, that none of the ASVAB subtests have loadings of .30 or higher on Factor II, and that it is the ASVAB technical factor which correlates highest with the job knowledge paper and pencil test factor. The ASVAB Verbal subtest did not meet the criteria for inclusion in this table. These results would seem to indicate that correlations between ASVAB and paper and pencil job knowledge measures are not simply the result of shared method variance.

The next highest inter-factor correlations are between the hands-on factor, Factor V, and the ASVAB math/speed and rating factors, Factors III and I respectively. While the hands-on factor is a relatively pure method factor, its correlations with the other factors strengthen the conclusions of Borman et al. Each method appears to measure a different but related piece of job performance.

Table 3 contains the oblique promax factor pattern for Infantrymen. Seven factors were extracted. The choice of variables to report was based on the same rules as for the previous table of loadings.

Table 3. Rotated Factor Pattern (STD REG COEFS)

I	II	III	IV	V	VI	VII	
64	.	.	.	.	.	.	P&P: E5-Oper as Station in Radio Net
64	.	.	.	.	.	.	School: All Items
61	.	.	.	.	.	.	P&P: B4-Perform OP Maint. on M16A1
59	.	.	.	.	.	.	P&P: H1-Perform Tracked Vehicle Maint
56	.	.	-39	.	.	.	P&P: E1-Collect/Report Info
.	66	.	.	.	.	.	Rating: Install/Fire/Recover M18A1 <Supv>
.	65	.	.	.	.	.	Rating: Load/Clear M60 <Supv>
.	59	.	.	.	.	.	Rating: Prepare Range Card for M60 <Supv>
.	54	.	37	.	.	.	Rating: Mean non MOS-Specific<Supv>
.	50	.	33	.	.	.	Rating: Navigate on Ground <Supv>
.	38	.	39	.	.	.	Rating: Set Headspace/Timing on .50 <Supv

.	31	.	.	.	.	.	P&P: G8-Estimate Range
.	.	76	.	.	.	.	ASVAB SUBTEST SCR-Auto/Shop
.	.	74	.	.	.	.	ASVAB SUBTEST SCR-Mechanical Comprehension
.	.	73	.	.	.	.	ASVAB SUBTEST SCR-General Science
.	.	73	.	.	.	.	ASVAB SUBTEST SCR-Verbal
.	.	.	79	.	.	.	Rating: Op as Station in Radio Net <Supv>
.	.	.	76	.	.	.	Rating: Op Radio Set AN/PRC-77 <Supv>
.	.	.	44	.	.	.	HO: E5-Op as Station in Radio Net
.	.	.	31	.	.	.	HO: BC-Engage Targets w LAW
.	.	.	.	68	.	.	HO: C6-Call/Adjust Indirect Fire
.	.	.	.	67	.	.	HO: G8-Estimate Range
.	.	.	.	55	.	.	HO: B4-Perform Op Maint on M16A1
.	39	.	.	37	.	.	Rating: Call/Adjust Indirect Fire <Supv>
30	.	.	.	32	.	.	P&P: B9-Engage w Hand Grenades
32	.	.	.	.	.	.	HO: BB-Prepare Range Card for M60
.	.	.	.	.	58	.	HO: J1-Movement in Urban Terrain
.	.	.	.	.	56	.	HO: BA-Prepare Dragon for Firing
.	.	.	.	36	50	.	HO: B9-Engage Targets w Grenades
.	.	.	.	.	47	.	HO: I1-Install/Fire/Recover M18A1
.	.	.	.	.	35	.	P&P: BA-Prepare Dragon for Firing
.	.	.	.	.	.	71	ASVAB SUBTEST SCR-Numerical Operations
.	.	.	.	.	.	59	ASVAB SUBTEST SCR-Coding Speed
.	.	40	.	.	.	54	ASVAB SUBTEST SCR-Math Knowledge
.	.	41	.	.	.	53	ASVAB SUBTEST SCR-Arithmetic Reasoning

While similar method factors emerge, the factor space for infantrymen is slightly more complex. The ASVAB Factors III and VII are quite clean, though Factor III and the paper and pencil job knowledge test Factor I are relatively oblique (Table 4.). These factors are substantially more correlated than are the two ASVAB factors with each other. Note also that the ASVAB math/speed Factor VII has a lower correlation with the paper and pencil job knowledge test Factor I, than the more technical ASVAB Factor III. If there is a simple "written test" factor, it failed to emerge in either of these solutions.

Perhaps most interesting are Factors IV and V. Each of these factors has a mixture of variable loadings representing different measurement methods. On Factor IV the supervisory rating and hands-on test for operating as a radio station in a net both load substantially. On Factor V the supervisory rating and hands-on test for call/adjust indirect fire both load substantially, and the paper and pencil and hands-on tests for engage targets with grenades also both load substantially.

Table 4 gives the correlations among the factors for Infantrymen. This solution is considerably more oblique than the solution for Medical Specialists.

Table 4. Inter-Factor correlations

	I	II	III	IV	V	VI	VII
I	100	30	53	36	18	25	24
II	30	100	13	40	18	19	-3
III	53	13	100	6	13	-1	29
IV	36	40	6	100	21	34	-5
V	18	18	13	21	100	-2	1

VI	25	19	-1	34	-2	100	0
VII	24	-3	29	-5	1	0	100

The highest correlation is between Factor I, the paper and pencil job knowledge test factor, and Factor III, the ASVAB technical factor. This result is similar to that noted previously. The two primarily supervisory rating factors, II and IV, are quite highly correlated with the paper and pencil test of job knowledge factor. In fact, Factor IV correlates almost as highly with Factor I (r=.36) as it does with the other rating factor, Factor II (r=.40). The two hands-on test factors, V and VI, are uncorrelated with each other. Factor VI has respectable correlations with both the paper and pencil job knowledge test factor, Factor I, and the rating factor, Factor IV.

Conclusions

Our tendency as psychologists is to abhor method variance as something to be avoided. This should not necessarily be the case in the realm of job performance measurement. Performance of a task requires first the ability and motivation to learn the task, and second the skill, ability, and motivation to perform it. Different methods of measuring performance, hands-on tests, written tests, and ratings, capture slightly different aspects of performance. Some of these relationships are apparent from the data presented above.

What remains for us is to understand which kinds of tasks are most appropriately measured by which methods. The research reported here, while open to several interpretations, presents a method and several examples of a way to do this. Clearly, more research needs to be conducted into the content of the tasks themselves and their relationships to method factors across several more occupations than are included here.

References

Borman, W. C., White, L. A., Gast, I. F., & Pulakos, E. D. (1985, August). Performance Ratings as Criteria: What is being Measured. Paper presented at the meeting of the American Psychological Association, Los Angeles, CA.

Campbell, J. P., & Harris, J. H. (1985, August). Criterion Reduction and Combination via a Participative Decision Making Panel. Paper presented at the meeting of the American Psychological Association, Los Angeles, CA.

Eaton, N. K., Hanser, L. M., & Shields, J. L. (in press). Validating Selection Tests Against Job Performance. In J. Zeidner (ED.), Human Productivity Enhancement. New York: Praeger.

Rumsey, M. G., Osborn, W. C., & Ford, P. (1985, August). Comparing Work Sample and Job Knowledge Measures. Paper presented at the meeting of the American Psychological Association, Los Angeles, CA.

WALK-THROUGH PERFORMANCE TEST DEVELOPMENT: LESSONS LEARNED

Carl J. Taylor
Jack L. Blackhurst
Rodger D. Ballentine

Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

The Air Force Human Resources Laboratory (AFHRL) is involved in a multi-year effort investigating the feasibility of measuring and linking job performance to Armed Services Vocational Aptitude Battery (ASVAB) scores. The major focus of this work is the development of a technology for systematically obtaining job performance data as criteria for validating enlisted, officer, and civilian selection systems, and evaluating Air Force training programs.

Planning for the Air Force's program of research in performance assessment began several years ago as the result of three primary requirements. Operational military and civilian program managers in the manpower, personnel, and training communities asked AFHRL to develop an approach for measuring job performance so that the measures could be used to assist in the evaluation of their training and selection programs. Secondly, the manpower, personnel, and training research community needed performance measures to serve as criteria in their research projects. Plans for the Air Force performance measurement effort to meet these requirements were already under development when a third requirement for these measures came with the congressional mandate to test the feasibility of validating the ASVAB against job performance measures.

The cornerstone of this criterion development effort is a work sample testing approach known as Walk-Through Performance Testing (WTPT). The WTPT process is being developed to expand the range of job tasks measured to include tasks which do not lend themselves to hands-on testing because of cost, time, and/or safety considerations. WTPT combines hands-on task performance and interview procedures to provide a high fidelity measure of individual technical job competence.

The walk-through procedure involves taking the job incumbent to the work site and administering a combination of performance tasks and interview questions. The interview testing component will be evaluated both as supplement to hands-on data collection to ensure adequate sampling of the domain of tasks in a job and as a more cost-effective substitute for hands-on testing. A wide range of alternative job performance measures will be developed in addition to the walk-through testing methodology. These include peer, supervisor, and self-performance ratings, at the task, dimension, and global levels.

Alternative measures will be developed for the same specialties used to develop walk-through testing techniques. When job sample, interview, and alternative forms have been developed for each of the Air Force specialties selected for this study, their relative ility will be determined. Existing performance measures such as technical training scores, Airman Performance Report ratings, skill-level advancement indices, and Specialty Knowledge Test scores will also be evaluated as possible alternative measures.

Over the past two years AFHRL has developed WTPTs for four Air Force Specialties (AFS). The Jet Engine Mechanic career field was selected as the AFS for WTPT prototype development. As the jet engine mechanic WTPT was developed, the process was cocumented (Alba and Wilcox, 1985) to serve as a set of procedural guidelines. These guidelines were then applied to three additional career fields: Air Traffic Control Operator, Avionic Communications Specialist, and Ground Radio Operator. The desire was to apply the guidelines directly to the next three career fields; however, it was anticipated that modifications might be needed to account for the dissimilarities between mechanical and non-mechanical career fields. The purpose of this paper is to highlight the lessons learned as the WTPT Procedural Guidelines were applied to the other career fields.

TASK SELECTION/TASK ANALYSIS

This step of WTPT development was d rectly transferable to the three additional career fields. That is, computerized occupational analysis data was used to develop a task selection plan. This plan specified the guidelines for identifying Phase I tasks (tasks performed by 30% or more of first-term incumbents) and Phase II tasks (tasks performed by 40% or more of first-term incumbents in a functional area and not included in Phase I).

The selected tasks were then reviewed/validated by subject matter experts (SME) during task selection workshops. These workshops resulted in a number of tasks being discarded or moved from one phase to another due to mission requirement changes, technology changes, or low task difficulty levels. The inconsistencies between occupational analysis data and SME input can be attributed to the age of the data. Even though the occupational survey reports (OSR) were the most recent ones for the three career fields, they were three to four years old. This resulted in several false starts. For instance, based on OSR data pertaining to the numbers of first-term personnel in subareas of the ground radio operator career field, the initial focus of task selection was on the Military Affiliated Radio System (MARS) area and not on the Mobile Communications (MOB) area. After spending some time with SMEs, it was discovered that, since the last OSR, mission emphasis had been increasing for the MOBs and decreasing for the MARS.

This type of problem can best be alleviated in the future by ensuring that OSR data is current on career fields selected for WTPT development. Communication with the major command (MAJCOM) career field functional manager very early in the task selection plan development is also essential so that recent or upcoming modifications in technology or mission can be a built into the WTPT.

Once tasks had been selected, a task selection workshop was held. Participants included the career field functional manager and SMEs having extensive experience in the career field. During the task selection workshop, emphasis was placed on ensuring the tasks were currently performed and that they were performed in a similar manner by everyone in the applicable functional area (i.e., flightline, shop, operations, maintenance). If workshop participants suggest deleting a task or moving it from one phase to another, detailed documentation should be provided on the justification. For each task, documentation should also be provided on equipment-related similarities/differences in how a task is performed, reliance on local operating procedures and Air Force or MAJCOM regulations, whether ermers only perform parts of or all of a task, and if a task is performed in the same manner in all units. Increasing the emphasis in these areas during the workshop will reduce confusion during the task analysis stage.

The objective of our task analysis is to gather information essential to WTPT item development. Relevant information includes beginning and ending points for each specific task, critical steps for task accomplishment, logistical requirements for task completion, required configuration of equipment, time critical and safety steps, effects of local operating procedures on task performance, and representativeness of the task. This information is gathered by referencing applicable regulations, technical orders, and local operating instructions as well as discussions with SMEs in the field. The desired result of the analysis is a comprehensive list of steps required for successful task completion. These steps should be generalizable to any situation in which the task might be observed and should be used to objectively evaluate an individual's performance on the task.

For the ground radio and avionic communications career fields this process worked very well; however, problems were encountered in the air traffic control area. One problem area is related to the differences in the work environment for radar approach control (RAPCON) personnel and tower personnel. RAPCON controllers depend solely on various radar scopes to separate and sequence aircraft. Tower controllers are concerned with line of sight traffic separation and use radar very infrequently. The result is that there are only a very small number of Phase I tasks (i.e., those accomplished in both facilities). Those tasks which are performed in both areas, typically are performed differently in the two areas. This resulted in a "separate but equal" approach. Problems encountered in constructing the actual WTPT will be discussed in the next section.

Test Development

The WTPT procedural guidelines also served as the basis for test development on all three additional specialties. The guidelines transferred easily to the new specialties; however, problems were encountered in developing standardized tests once the tasks were identified.

The first problem area is test security. Because of the physical arrangement of most radio operations rooms, avionic maintenance shops, and air traffic control facilities, it will be difficult to administer the WTPT without allowing others to observe the tasks being evaluated. One possible solution is to temporarily partition off the test area; however, this may impact the unit's ability to carry out their mission. Additionally, some radio facilities require security clearances for facility access due to their mission. This may present problems during data collection as prospective contractor evaluators have not obtained clearances. In an attempt to alleviate this problem, the AFHRL recently conducted a study comparing performance ratings given by contractors with those given by active duty personnel; however, results are unavailable at this time.

The air traffic control field presents unique problems in terms of developing standardized testing situations. For instance, the radio, avionics, and jet engine mechanic jobs involve heavy reliance on technical orders, are labor intensive, and are characterized by completion of tasks in a routine, standardized manner. Most tasks can only be done one way and only rarely does the performance of one task have any effect on subsequent tasks. These characteristics make it relatively easy to construct standardized testing environments. In addition, jet engines or radio equipment can actually be used for the test and the test administrator can ensure the equipment is configured exactly the same for each incumbent.

Conversely, the air traffic controller's job is characterized by many correct ways to deal with a situation, is intensely interactive, and often has a time critical component. For test administration purposes it cannot be requested that aircraft fly repeatedly in preset patterns for purposes of presenting standardized scenarios to the incumbents being evaluated.

Another characteristic of the air traffic control area is that individual tasks for this career field require few measureable steps for completion. This is a function of the air traffic controller's job. For instance, tuning a radio or troubleshooting avionics equipment requires a series of specific steps which can be readily observed and objectively evaluated. In contrast, the observable portions of the air traffic control job consists of keying a microphone, operating runway lights, or using the correct phraseology. These tasks cannot be further dissected into measureable substeps.

To address these problems, AFHRL developed a "job module approach" for air traffic control operators involved in RAPCON duties. The job module makes use of RAPCON simulators and combines a number of tasks into standardized scenarios. The incumbent is seated at the simulator radar screen and is required to "work" typical traffic problems of varying levels of complexity. Individual tasks are scored as they occur in these scenarios. Due to the absence of tower simulators, a number of approaches for developing standardized job modules for tower personnel were considered. These approaches will be outlined in the following section.

Due to the absence of tower simulators, AFHRL considered a number of approaches to standardized hands-on testing for tower personnel (e.g., using the tower simulator, video taping scenarios from the simulator, developing a new simulator, computer games, slide presentations, and video taping live situations). Each of these approaches was evaluated with respect to test standardization, objectivity, stimulus fidelity, cost, discrimination, and time to develop. As a result of these comparisons, the use of video taped live traffic is being pursued.

Summary

The AFHRL has developed a set of procedural guidelines which can be used to design valid, reliable, and standardized hands-on performance assessment instruments for Air Force enlisted career fields. The guidelines were developed around a mechanical career field and then applied to career fields representing the remaining three ASVAB Aptitude Index areas (administrative, electrical, and general). The major lesson learned with respect to applying the procedural guidelines concerns the OSR information used as the starting point for task selection. Every effort should be made to ensure that the OSRs are current for future AFSs. The other lessons learned involve the ability to develop standardized testing environments and test administration logistics issues. Being aware of these issues at the beginning of WTPT development will result in a much smoother development.

References

Alba, P. A. & Wilcox, T. (1985). Walk-through performance testing procedural guidelines manual. Report submitted to AFHRL, Brooks AFB, TX: Air Force Human Resources Laboratory.

Uniform Guidelines for Employment Selection Procedures. From Federal Register 43(166), August 15, 1978.

ON THE CONTENT AND MEASUREMENT VALIDITY OF HANDS-ON JOB PERFORMANCE TESTS

PROBLEM

The justification for using aptitude tests to help select enlisted recruits and assign them to occupational specialties is that aptitude tests are valid predictors of performance. The aptitude tests used by the military services have been extensively validated as predictors of performance in occupational specialty training courses. Their usefulness as predictors of performance on the job, however, is less well documented. The Job Performance Measurement Project has been initiated to validate the Armed Services Vocational Aptitude Battery (ASVAB) as a predictor of job performance.

The question then arises of how job performance should be measured. The measures favored by the Joint Service Job Performance Measurement Working Group are hands-on job performance tests. These tests have intrinsic validity because of their high fidelity to job behavior. Hands-on performance tests, however, are susceptible to poor content and measurement validity.

- Poor content validity may arise because the tests focus on skills easy to test in the hands-on mode without including the full range of job requirements.
- Poor measurement validity may arise because the scoring standards of test administrators are not calibrated and because test security is difficult to maintain (examinees can find out what is being tested and practice beforehand).

The purpose of this analysis is to examine the content and measurement validity of prototype hands-on tests for three Marine Corps specialties – Ground Radio Repair, Automotive Mechanic, and Infantry Rifleman – used in a feasibility study to evaluate ASVAB qualification standards.

FINDINGS

The findings pertain to the two technical specialties, Ground Radio Repair and Automotive Mechanic. Because the infantry riflemen in the sample had limited job experience, the results for them are inconclusive.

- Hands-on test scores were only weakly related to amount of job experience, as measured by months in the Marine Corps (figure I). Test scores were expected to increase with experience, and the lack of relationship raises questions about how well the tests represent the full range of job requirements.

- The ASVAB is a valid predictor of hands-on test scores for people with 2 years or less of service in the Marine Corps, but not for people with more than 2 years of service (table I).

- Hands-on test scores did increase with experience for people with low aptitude, but not for people with high aptitude (figure II).

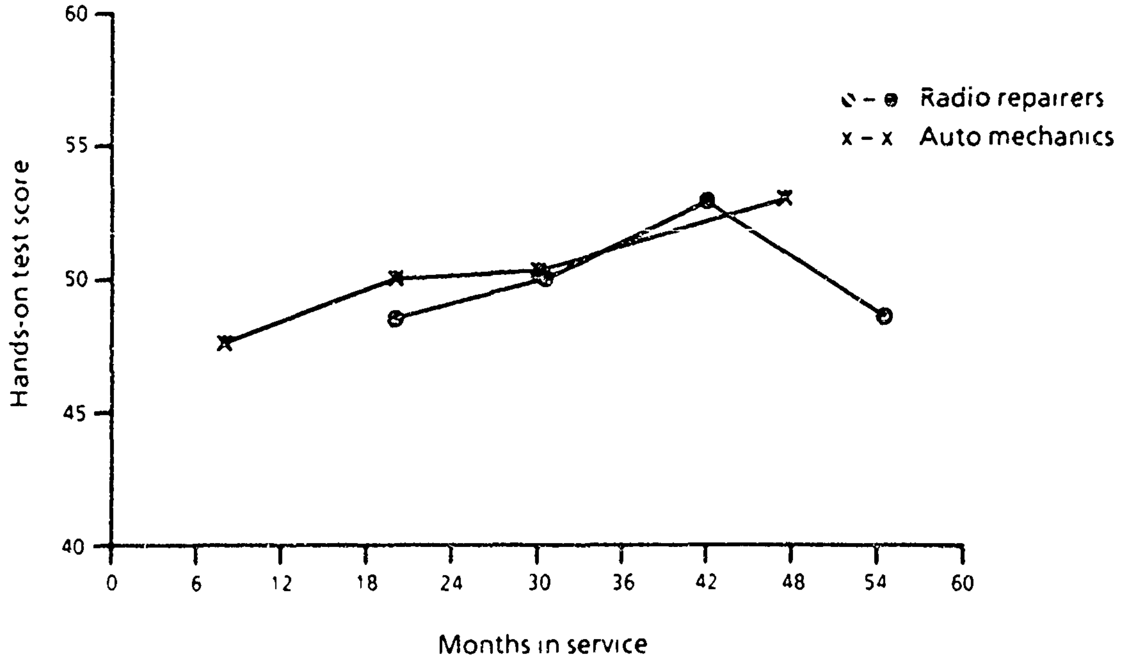


FIG. I: JOB PERFORMANCE RELATED TO TIME IN THE MARINE CORPS

These results suggest that the hands-on test content was appropriate for people recentl" assigned to their first duty station, but less appropriate for people with more experience, who perform job tasks not reflected in the tests.

The findings on measurement validity bear on institutionalizing of hands-on job performance tests:

- The test administrators used different scoring standards, and the same administrators changed their scoring standards across time (see figure III).

- Maintaining test security is difficult.

TABLE I

VALIDITY OF THE ASVAB FOR PREDICTING HANDS-ON TEST SCORES

Months in service	Validity[a]	Number of cases
	Ground Radio Repair	
15-25	69	38
26-35	00	53
36-48	00	37
Total	37	128
	Automotive Mechanic	
2-14	.72	57
15-25	52	56
26-34	15	53
35-60	- 07	54
Total	37	220

a Population-wide estimate of validity coefficient

CONCLUSIONS

- The ASVAB is a valid predictor of job performance, as measured by hands-on tests.

- But hands-on tests lack robustness:

- Content validity is sensitive to job experience.

- Measurement validity is sensitive to the calibration, or scoring standards, of test administrators.

- Institutionalizing hands-on job performance tests would be difficult.

Milton H. Maier
Center for Naval Analyses
Alexandria, Virginia 22302-0268

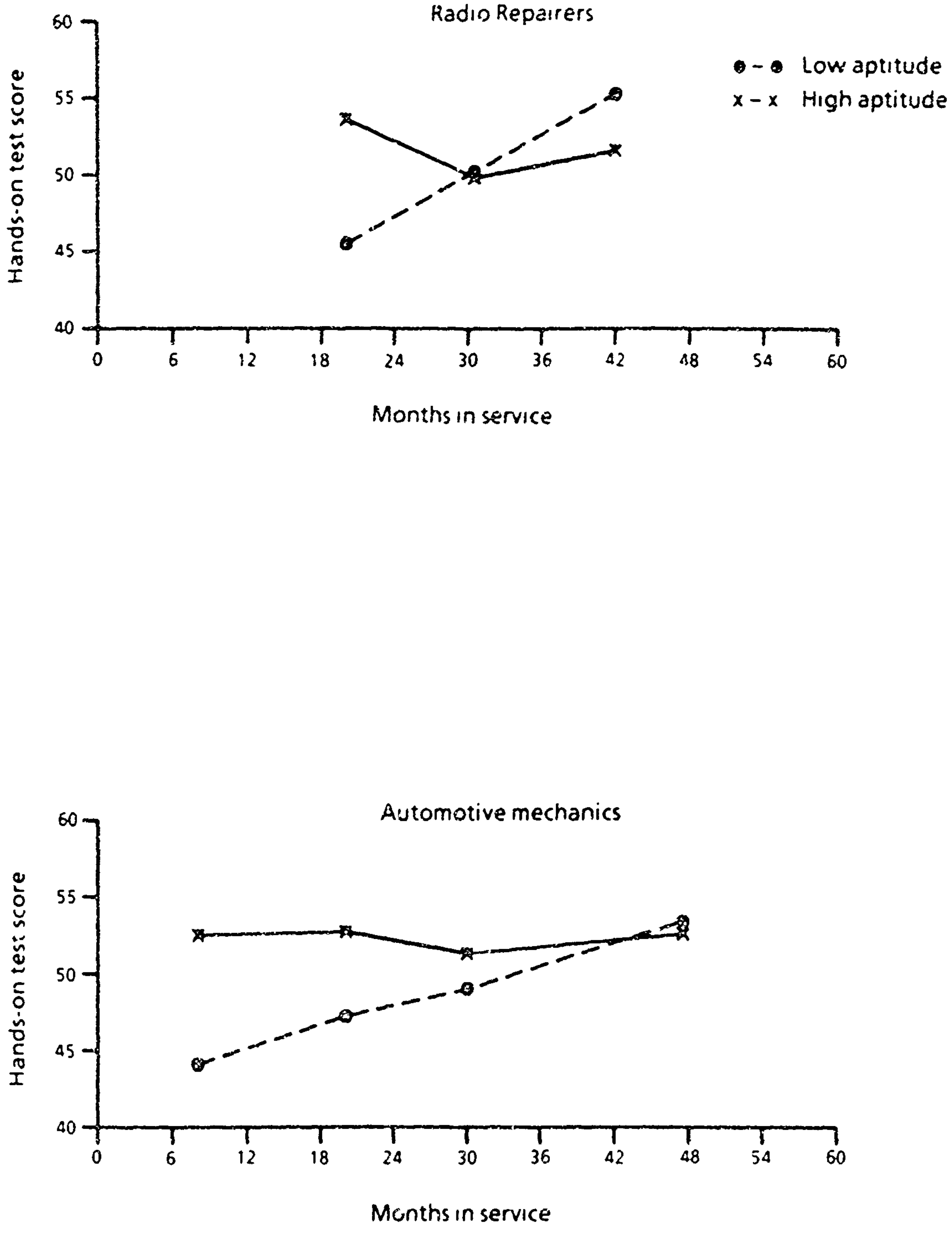


FIG. II: PERFORMANCE RELATED TO APTITUDE AND EXPERIENCE

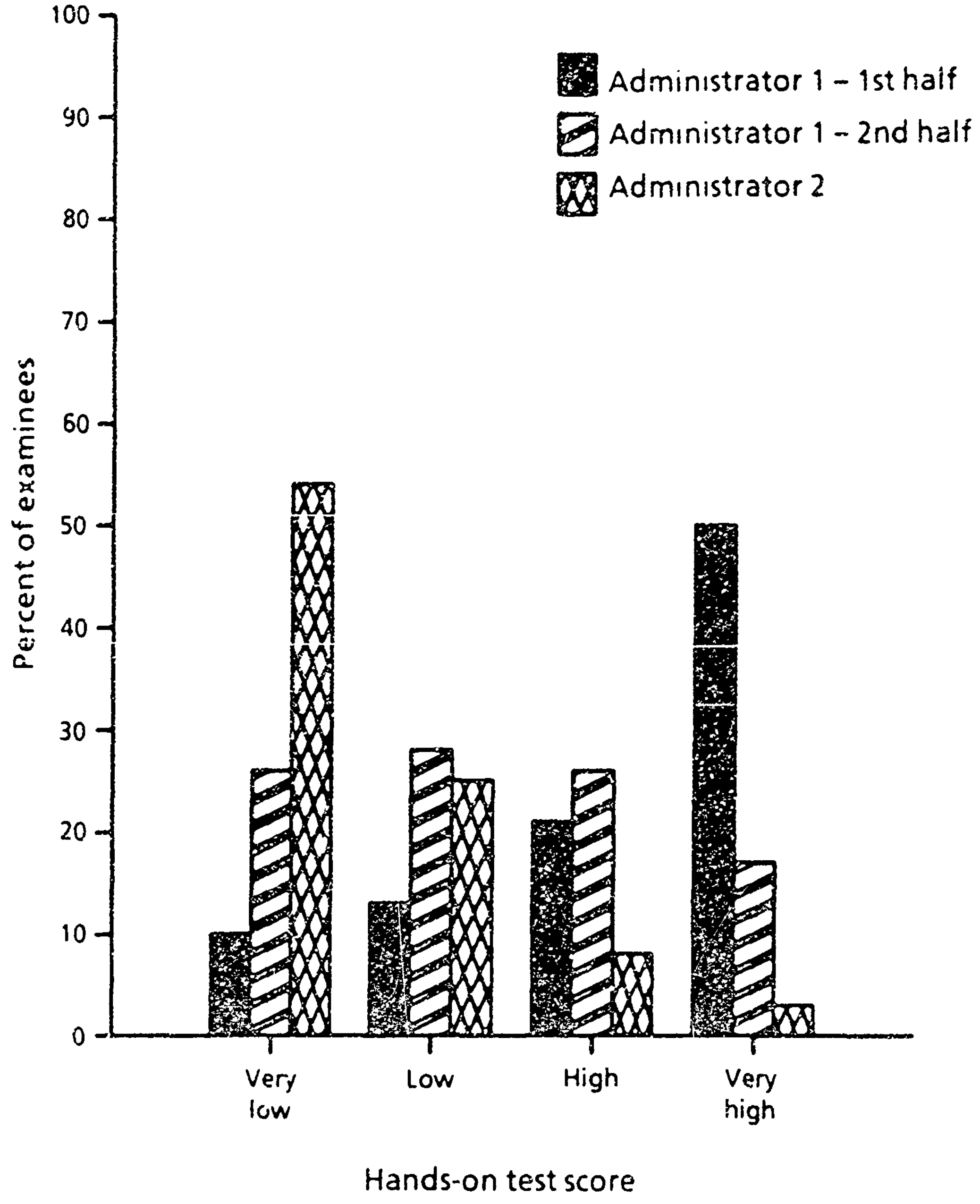


FIG. III: HANDS-ON TEST SCORES ASSIGNED BY TEST ADMINISTRATORS TO AUTOMOTIVE MECHANICS

DEVELOPING AND EVALUATING
A HANDS-ON PERFORMANCE TEST

L. P. Kroeker, R. M. Bearden, & G. J. Laabs
Navy Personnel Research and Development Center
San Diego, CA 92152

The first step in reaching the Navy's goal to improve enlisted personnel classification through the use of job performance information includes the development and evaluation of hands-on performance tests. This paper describes the developmental procedures employed in the first rating to be covered in a Congressionally mandated, Joint Service project aimed at linking job performance to enlistment standards.

Background

In 198[illegible], Congress formally required the Armed Services to establish methods for measuring job performance and validating selection standards against them. The Navy's contribution to this coordinated effort is entitled Performance-Based Personnel Classification. It's objectives are to investigate measurement approaches that can be used to assess on-the-job performance and to improve the Navy's automated classification and assignment system or CLASP (Kroeker and Rafacz, 1983) by including job performance information.

The Navy's approach focuses on direct measurement of technical proficiency, which follows the research strategy of the Joint Service project. The purpose of this large scale effort is to develop job performance measures for first-term enlistees with four or fewer years of service and demonstrate their use as criteria for predictor validation (Office of the Secretary of Defense, 1984).

In the Joint Service project the hands on job sample test has been adopted as a high fidelity benchmark measure against which other less costly and more easily administered measures will be compared. Therefore, it is essential to construct valid, reliable, and objective hands on performance tests.

Because of the extensive resources required for this large scale validation effort, performance measures will be developed only for a small number of Navy ratings. The Machinist's Mate (MM) rating is the first one to be covered.

Machinist Mate Performance Test Development

To ensure that the job sample test for MM's adequately represented the technical proficiency content domain, a sequence of task selection steps were taken (Guion, 1979). These steps, which involved the orderly reduction of the job content universe to the job sample, resulted in the identification of the critical tasks to be included in the test.

The job content universe was defined by a comprehensive job task analysis taken from the Navy Occupational Task Analysis Program (NOTAP) data base. The latter was supplemented with information from: (1) Occupational Standards; (2) Personnel Advancement Requirement (PAR); (3) Personnel Qualification Standards (PQS); (4) standard operating procedures; (5) standard maintenance procedures; (6) technical manuals; and (7) A-school learning objectives.

The job content domain was determined by subject matter expert (SME) judgments. The original task list was reduced to only those tasks involving technical proficiency for MM's employed on 1052 class frigates.

The next step involved the definition of the test content universe. This universe consisted of those job tasks from the technical proficiency content domain that might be observed in a hands on test, in addition to all conditions that might be imposed in the testing situation and the procedures for observing and recording responses. At this stage SME judgments were gathered on the criticalness of each task to the operation of a ship's propulsion plant.

Finally, the test content domain was defined. Tasks were drawn from the test content universe based on the task-critical judgments obtained in the preceding step. They cover the two main areas of work behavior for MM's, namely, maintenance and watch-standing. The procedure ultimately yielded the following tasks, which were approved as a representative job sample by the MM training community:

A. Maintenance Tasks

1. Planned Maintenance
   - (i) Sample oil
   - (ii) Inspect oil
   - (iii) Tag out

2. Corrective Maintenance
   - (i) Make gasket
   - (ii) Repair valve
   - (iii) Repair Valve

I. Watchstanding Tasks

1. Normal Watchstanding
   - (i) Shift and inspect water strainer
   - (ii) Shift and inspect oil strainer
   - (iii) Operate eductor
   - (iv) Operate fire pump
   - (v) Record gauge readings
   - (vi) Clean and inspect oil filter or operate oil pump

2. Casualty Control
   - (i) Loss of oil pressure
   - (ii) Major oil leak
   - (iii) Loss of vacuum
   - (iv) Hot bearing

In completing the task selection procedures the following [illegible] procedure was followed. First, the behavioral elements of each task were identified. This was relatively easy because MM's have fully proceduralized job performance aids that specify the step-by step requirements of watchstanding tasks. These task steps are outlined in the Engineering Operational Procedures (EOP) and Engineering Operational Casualty Control (EOCC) documents. The specification of the behavioral elements for the maintenance tasks were obtained from fleet SME's and [illegible] the MM training community.

Following the delineation of the behavioral elements associated with each task, MM's aboard two frigates were observed as they performed the tasks. A few tasks were eliminated because of the lack of procedural standardization across ships. It was at this point that the set of tasks were presented to the MM training community for approval.

After the performance of the tasks was observed aboard the two ships, a structured observation form was generated. Draft versions were reviewed for accuracy and comformity to conventional operating procedures by SME's. Next an experimental version of the instrument was administered and revised.

The experimental version was administered to apprentice MM's on board the frigate. As a result of this try out some modifications of the test were made. One of our discoveries was that there was informal job specialization even on this relatively small ship, whereas we had expected all first term MM's to perform almost all of the watchstanding tasks connected with the MM propulsion engineering spaces. This led to the development of parallel forms for the two major MM work spaces: one for the Engine Room and one for the Auxiliary (Generator) Room. The common maintenance tasks were developed for administration in a pierside building or van rather than on board a ship.

Finally, a scoring key was prepared in which each behavioral element is dichotomously scored depending upon the presence or absence of the required response.

Observer Training

To ensure that results from hands-on job sample testing are reliable, examiners must be trained to act as unbiased observers of test performance. Therefore, it was necessary to develop a training program for administrators and scorers of the tests.

A training package using a role-playing procedure was developed. The procedure requires that each person to be trained takes a turn playing the role of scorer, examinee, and observer. The role of observer differs from that of scorer in that the former scores the test but does not otherwise interact with the examinee. The rationale underlying this procedure was the expected convergence of scorer behavior as scorers and observers compare and discuss points at which their observations and actions diverge.

Two three-member teams (four contractor and two NPRDC personnel) participated in the training program on board several San Diego based ships. The first training day was devoted to orientation and the second to reviewing standardization procedures, administration instructions and scoring processes. The next six days were spent in role playing and discussion sessions on board a frigate. The training steps were repeated for all job tasks at both work stations as well as at the pierside maintainance test station.

Data Collection

Testing is currently in progress and is being done under contract with a team of four test observers, all of whom are former MM's. Each has approximately twenty years of experience and has attained the minimum paygrade of E7. The remainder of the team, who provide periodic quality control checks, consists of two military members of NPRDC.

Ultimately, we will have scores on the hands-on test for about [illegible] first-term MM's who are serving aboard twenty eight 1052 class frigates. These ships are located at seven test sites including San Diego, Long Beach, Charleston, Norfolk, Pearl Harbor, Newport RI, and Mayport FL.

Performance Test Evaluation

Various analyses will be applied to the data collected. These will include the derivation of distributional characteristics and descriptive statistics, item analyses, and the calculation of various reliability and validity indices. Initially, we are placing a great deal of emphasis on methodology arising from Generalizability Theory.

Performance measures may contain the the same sources of error as traditional paper and pencil measures; namely, instability of responses from one occasion to another, non-equivalence of supposedly parallel forms, and heterogeneous subtest responses. Generalizability Theory (Shavelson and Webb, 1981) will be used to estimate the magnitude of each of these error sources, individually and in combinations.

Scores on the hands-on test may result on the one hand from true performance, skill, and ability factors and on the other hand from extraneous features of the testing situation. These features may reflect differences in MM job experience, differences in testing conditions, and differences in MM qualifications across the different testing locations.

By making extraneous variance sources explicit through a generalizability framework, it is possible to identify where error in the criterion measurement is occurring. Consequently, the generalizability framework can act as a conceptual guide to job performance research and increase awareness of potential error sources that need to be controlled or monitored during test administration.

The main questions of interest within the context of a generalizability design involve different possible sources of error variance. The following sources appear most plausible as reliability attenuators:

1. Amount of experience
2. Type of watch station
3. Type of entry level training
4. Location of duty station
5. State of equipment
6. Distractions during testing
7. Test equipment source
8. Scorer differences

These are the facets of the generalizability design for which variance estimates will made.

We are hopeful that we have constructed a sufficiently parallel test that will be objectively observed so that sources of error variance from the administration of the instrument will be negligible. The potential sources of error variance associated with site and equipment remain to be evaluated.

Summary

The MM hands-on performance test is the first one to be developed within the Navy's job performance research effort. After SME's selected the set of critical tasks to assess technical proficiency of first term MM's, a hands on test was developed.

The discovery of informal job specialization during test development led to the construction of parallel test forms for a number of job tasks. A training package for test observers was designed and was used to reduce the variation in test administration and scoring.

Training of the test observers has been completed and data collection has begun on the pre-test ships with two observers at each of three test stations. In order to gather data from 300 MM's on 28 ships in 7 different homeports we will be testing for about one year.

Finally, generalizability models will be used to estimate potential error sources associated with hands on tests in the overall sample. These sources will include observer disagreement, nonequivalence of parallel forms, unstandardized testing conditions across ships, and the lack of test security. The results of these analyses will be used to guide subsequent analyses.

References

Guion, R. M. (April 1979). Principles of work sample testing: III. Construction and evaluation of work sample tests (ARI TR 79-A10). Alexandria: Army Research Institute.

Kroeker, L. P., & Rafacz, B. A. (November 1983). Classification and assignment within PRIDE (CLASP): A recruit assignment model (NPRDC Tech. Rep. 84-9). San Diego: Navy Personnel Research and Development Center. (AD A136 907)

Office of the Assistant Secretary of Defense (M,I, & L). (December 1984). Joint Service efforts to link Enlistment Standards to Job Performance: Third Annual Report to the House Committee on Appropriations. Washington, D.C.

Shavelson, R. J. and Webb, N. M. Generalizability theory: 1973-1980. British Journal of Mathematical and Statistical Psychology, 34, 133-166, 1981.

SIMULATION OF HANDS-ON TESTING
FOR NAVY MACHINIST'S MATES

Robert Vineberg & John N. Joyner

Human Resources Research Organization
27857 Berwick Drive
Carmel, CA 93923

This paper concerns the substitution of one type of job sample test, a so-called abstracted measure, for another type, a direct measure, in evaluating proficiency. Direct job samples maintain the materials, the responses, and the appearance of actual job tasks. Such samples often encompass entire tasks -- all of the elements of performance that a person might execute together naturally, as in changing a tire or boiling an egg. Even when direct job samples consist only of segments of tasks, each segment generally maintains its integrity and its appearance as one part of a larger unit of performance.

Abstracted job samples, on the other hand, are representations of a process or components of performance that have been disassociated from th elements that naturally accompany them. The appearance of the abstracted component, therefore, is inevitably altered, sometimes very slightly but sometimes radically. An abstracted sample almost always looks more like a test than a job activity, whether or not it is recognizable in terms of its parent task.

An abstracted job sample can be derived from any component or process of task performance that discriminates among performers, regardless of whether the resulting measure has a "hands-on," joblike appearance. An abstracted sample may measure skill or physical ability or job knowledge. Soldering skill required to repair electronic equipment has been evaluated by having persons make solder joints on a "breadboard"; one of the physical requirements for sanitation workers and firemen has been tested by having them lift and carry heavy weights. Perhaps the most familiar and frequently used abstracted job sample is the multiple-choice test of job information.

For evaluating job proficiency, abstracted measures of job knowledge are usually preferred to direct measures because of their efficiency. It is generally accepted that for many tasks a capability to perform can reasonably be inferred if the acquisition and retention of task knowledge have been demonstrated. Yet, direct measures, despite their inefficiency, still have considerable appeal because their use avoids the analysis and alteration of performance inherent in developing an abstracted sample. In the process of translating a task into an abstracted sample, discriminating requirements may be lost and artificial requirements must generally be introduced.

However, this is the problem that arises from the use of multiple-choice and other selected-response formats to measure abstracted knowledge: much of actual performance calls for the recall of information, but the multiple-choice format requires only recognition. Ellis, Wulfeck, and their co-workers have designed procedures for checking the adequacy of test formats for meeting the performance requirements in knowledge measurement (Ellis, Wulfeck, & Fredericks, 1979; Ellis & Wulfeck, 1982).

Abstractions of the information-gathering and decision-making components of electronic and mechanical troubleshooting was a popular research area some years ago. The Tab Test (Cornell, Damrin, Saupe, & Crowder, 1954) and like instruments were used as abstract representations of troubleshooting processes. Steinemann (1966) demonstrated, however, that the behaviors elicited in such simulations are quite different from those seen in a direct job sample. Correlations between a hands-on test of troubleshooting and the abstracted measures ranged from about -.5 to +.1. The abstracted measures appear both to invite information gathering that doesn't take place in actual performance (where searching requires much more effort) and to provide considerable structure to a task that is, in actuality, often ambiguous.

Osborn & Ford (1977) reported that pictorial tests provided a poorer measure of knowledge of manual procedures than did multiple-choice tests. They compared performance on four different types of knowledge tests with hands-on performance on tasks like installing a field telephone and disassembling a rifle. In the pictorial test pictures had to be arranged to demonstrate a correct sequence of steps. This seems to have required persons to form a mental image of the entire procedure before arranging the pictures. It appears that the skills required to sort and arrange pictures in the abstracted sample were not part of the actual tasks: average correlations with the hands-on test dropped from the .80s to the high .50s. It is also of interest that persons with lower aptitudes showed their lowest scores when the knowledge tests were based on the picture-arranging format.

I would like to report a different kind of problem encountered in the construction of an abstracted measure for evaluating the proficiency of machinist's mates on Navy frigates (Vineberg, Joyner, & Zimmerman, 1985). In this instance, the abstracted measure was a pictorial simulation of an existing job sample test.

A machinist's mate engages in a variety of sometimes lengthy procedural tasks in the alignment, checking, operating, and troubleshooting of pumps, strainers, condensers, and generators. The number and variety of valves and gauges distributed through the engineering spaces of a frigate present a potentially bewildering array of options to the examinee. In a direct job sample developed earlier, tasks were performed on actual equipment in a ship's engine and generator rooms. In the abstracted job sample, a scenario (identical to that used in the direct job sample) was presented together with photographs of equipment taken in engine and generator rooms.

The examinee was required both to describe actions he would take and to mark the particular equipment components on the photographs that he would observe and manipulate. The photographs had been taken with a wide-angle lens to capture both the equipment relevant to a given task and a good deal of irrelevant equipment in the surround. The pictured equipment provided both a context for recall of task procedures and distractors that could elicit incorrect responses.

In the course of developing the abstracted job measure, it was discovered that the appearance and sometimes the makeup and location of equipment varied among ships of even the same type and class (Knox class frigates) for which the test was developed. Such variability in ship construction is apparently common when ships are built or overhauled at different times, by different manufacturers, in different shipyards. (The arrangement of equipment on a particular ship may sometimes be documented by ad hoc survey, but we are not aware of any compilation of this information across ships.)

The result, of course, is that examinees will vary in familiarity with the pictures of equipment appearing on such a test taken on a particular ship, solely as a consequence of their experiences on other ships with differently configured equipment. The impact of such variation on test performance is likely to be greatest among apprentice job incumbents, for whom the present test was intended, since such persons are least familiar with such variation.

Notice, by contrast, that these variations among ships do not necessarily cause problems in measuring proficiency by direct job sample. In the latter, variations in equipment and work materials in the normal work setting need not lead to differences among examinees in familiarity with a particular set of equipment selected to be represented on the test. Direct job sample tests are often administered in a person's own work situation and use the very same equipment and materials that he or she works with on the job, and standardized score sheets for recording performance can be prepared in a general enough way to be independent of any variations in equipment appearance in the different work settings where the test is administered.

If an abstracted measure, on the other hand, attempts to faithfully render shipboard equipment, as by photograph, the features of the equipment must necessarily be those of particular machinery on a particular ship. Although the consequences of differences in equipment have not been investigated here, the question of test fairness naturally arises when some examinees are presented with familiar materials and others are not.

The effects of such situational variations on abstracted measurement can be avoided by (1) generalizing the abstracted measure so that situational variation is irrelevant or (2) restricting the abstracted measure solely to tasks that use equipment and materials common to all work settings. It may be, of course, that neither of these options is acceptable for making valid inferences about proficiency and that only a direct job sample may remain viable.

REFERENCES

Cornell, F.G., Damrin, D.L., Saupe, J.L., & Crowder, N.A. (1952). Proficiency of Q-24 radar mechanics: III. The Tab Test--a group test of trouble-shooting proficiency (AFPTRC-TR-54-52). Lackland Air Force Base, TX: Air Force Personnel and Training Research Center.

Ellis, J.A., & Wulfeck, W.H. II (1982). Handbook for testing in Navy schools (NPRDC Special Report 83-2). San Diego: Navy Personnel Research and Development Center.

Ellis, J.A., Wulfeck, W.H. II, & Fredericks, P.S. (1979). The Instructional Quality Inventory: II. User's Manual (NPRDC Special Report 79-24). San Diego: Navy Personnel Research and Development Center.

Osborn, W., & Ford, P. (1977). Knowledge tests of manual task procedures. Proceedings of the Annual Military Testing Association Conference, 1977, 634-649.

Steinemann, J.H. (1966). Comparison of performance on analogous simulated and actual troubleshooting tasks (RM SRM 67-1). San Diego: Navy Personnel Research and Development Center.

Vineberg, R., Joyner, J.N., & Zimmerman, R. (1985). Development of a Hands-On Job Sample Test and Scorer-Training Materials for Apprentices in the Machinist's Mate Rating. San Diego: Navy Personnel Research and Development Center.

Inter-Service Transfer of Job Performance Measurement Technology

Capt Jack L. Blackhurst, USAF
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

Herbert George Baker, PhD
Navy Personnel Research and Development Center
San Diego, California 92152-6800

The Department of Defense is coordinating a Joint-Service Job Performance Measurement (JPM) Project in response to a congressional mandate to establish linkages between job performance and enlistment standards. The objectives of this Joint-Service Project are to: (1) develop prototype methodologies for the measurement of job performance; and (2) if feasible, link enlistment standards to on-the-job performance. The overall program will develop measurement techniques for the collection of valid, accurate, and reliable hands-on job performance information that can be related to recruit capabilities. These measures, in turn, will be used as benchmarks against which surrogate indices of performance (less expensive, easier to administer tests and/or existing performance information) will be evaluated as substitutes for the more expensive, labor intensive, hands-on performance measures. The long-term goal of the research and development program is for each Service to establish an operational performance measurement program so that job performance data will be available for use in evaluating personnel and training policies and practices.

Each of the Services is responsible for developing performance measurement technologies on occupational specialties which are comparable across Services. This approach permits the Services to share the technology for similar specialties. A part of the basic strategy of the Joint-Service JPM project is to determine if the technologies developed by one Service can be utilized successfully in other Services. The Air Force is responsible for taking the lead in inter-Service technology transfer. This paper will discuss the transfer to the Navy of a technology developed by the Air Force.

Air Force Testing Technology

The Air Force has developed a comprehensive performance assessment system for the Jet Engine Mechanic Specialty using the combination of Walk-Through Performance Testing (WTPT), rating forms, and related questionnaires. WTPT is a task-level job performance measurement system that combines hands-on task performance and interview procedures to provide a high fidelity measure of an individual's technical job competence. The hands-on component resembles a traditional work sample designed to measure performance on a sample of tasks that have survived the imposition of essential measurement constraints such as testing time/cost or risk of personal injury/equipment damage. The interview component has been added as a means of assessing those tasks that would have been eliminated because of these constraints. Interview testing takes place in the work setting and requires the evaluator to assess an incumbent's proficiency on a task by asking questions designed to uncover knowledge and procedural strengths and weaknesses related to the performance of that task. The incumbent can answer the questions by a combination of verbal responses, gestures, and demonstration. The interview testing component will be evaluated both as a more cost-effective surrogate and supplement to hands-on measures.

In addition to the WTPT, the Air Force has developed a wide range of rating forms as potential surrogate job performance measures. These include peer, supervisor, and self ratings at four different levels of measurement specificity: task, dimension, global, and Air Force-wide. In addition, a set of questionnaires were also developed to assess job experience and level of motivation. A detailed discussion of the development of the Air Force performance assessment system can be found in Gould and Hedge (1983) and Hedge (1984).

Results to Date

The Air Force developed performance measures for the jet engine mechanic (AFS 426X2) on the three most representative engine types (J-79, J-57, TF-33) currently used by the Air Force. The Air Force, Navy, and Marines all use the J-79 engine in F-4 (fighter) aircraft and have first-term jet engine mechanics who maintain this engine. The Air Force and Navy began discussions in the Summer of 1984 regarding the feasibility of transferring the performance measures developed for the Air Force J-79 jet engine mechanic to the Navy J-79 mechanic. These discussions led to a transfer plan outlining individual Service responsiblities and a three-phase effort to transfer the technology. Phase I was a feasibility study to determine if the transfer could successfully be accomplished. Phase II was the modification of the instruments and procedures, including a pilot test. Phase III is the actual data collection and analysis. The Air Force has responsibility for Phases I and II, with the Navy responsible for Phase III.

To begin the effort, representatives from both Service research laboratories, contractor personnel, Navy training personnel, and two Marine J-79 jet engine mechanic experts attended a workshop at the Air Force Human Resources Laboratory in April 1985. The workshop laid the ground work for the completion of the task. The Marine subject matter experts (SMEs) reviewed the Air Force instruments and test procedures to determine the extent of change needed. Their review indicated that the measures could be transferred with minimal modifications. Following the workshop, several field visits with SMEs at Navy sites were made to confirm the necessary changes, as well as to examine the feasibility of testing Navy mechanics. In addition, appropriate Navy documents (Occupational Survey Report, Training Outlines, etc.) were reviewed to ensure that the tasks represented in the Air Force tests were appropriate for testing Navy and Marine personnel.

A second workshop was held in July 1985 at the Navy Personnel Research and Development Center to receive the contractor's report on the feasibility study, and to determine if the effort should continue. The report indicated that the transfer of performance measurement technology from the Air Force to the Navy and Marine Corps was feasible. The study found that most of the tasks in the Air Force instruments could be used with only minor modifications and that only two of the tasks were not performed by Navy or Marine personnel due to an equipment difference. For example, the Air Force J-79 engine requires the installation of a starter; however, the Navy or Marine J-79 engine uses an air start and does not have a starter. Therefore, two tasks related to the starter were deleted from the performance test. The rating forms and experience and motivation questionnaires required only minor Service terminology changes and the test logistics did not appear to be a problem. A decision was made to continue the technology transfer effort.

However, a major finding of the feasibility study was that the Navy is phasing out the J-79 engine, which affects the number of first term Navy mechanics that will be available for testing. Because of the limited number of Navy incumbents available, the Navy suggested the inclusion of Marine J-79 jet engine mechanics since, in addition to undergoing identical skill training, they often work side by side with Navy jet engine mechanics on the same engines. Thus, approval has been requested to collect data on Marines, transforming this study into a Tri-Service JPM technology transfer effort.

Another recommendation from the feasibility study was to incorporate the development of Navy job knowledge tests into the study design. The Navy's job knowledge test is different than the typical knowledge test in that it makes extensive use of photographs that the incumbent can use to answer questions. The photographs enable the incumbent to reference the equipment, forms, etc. that he/she would normally use on the job. Inclusion of this additional measurement technique would allow direct comparison of surrogate performance measures developed by different Services on the same sample of incumbents, something which is currently not available in the Joint-Service project. The recommendation was adopted and the measures will be developed for administration with the other performance tests.

Progress has been made in the study. The instruments have been developd or modified as needed and are ready to be pilot tested. Following the pilot test, the measures will be administered to approximately 100 Navy and Marine jet engine mechanics, using test administration procedures similar to those used by the Air Force to collect data on their jet engine mechanics. Results of the data analyses will be available by the Fall of 1986.

Research Benefits

This technology transfer effort will have four major research benefits: (1) This study will serve as a prototype for future attempts to transfer performance measurement technologies. Experience gained from this effort will be invaluable to any future inter-Service performance measurement technology transfers. (2) This is the first attempt at direct comparisons of surrogate techniques, thus underscoring the joint-Service nature of this project. Such comparison of useful Service surrogates provides an expanded assessment of more cost-effective performance measures. The primary surrogate measures from two Services will be directly comparable on the same sample of job incumbents. (3) It allows the Services to gather performance information on additional specialties at significant cost savings because much of the design work has been completed. As a result, the transfer can occur relatively quickly and at a much lower cost than to do the specialty separately by Service. (4) This effort enhances the total Joint-Service JPM effort. It allows the Services opportunity to share new techniques while enriching their own individual research programs. Tremendous opportunity exists for generation of research ideas and for additional analyses of alternative performance measurement methods. Significant contributions will be made to performance measurement data bases and research.

Summary and Future Research Possibilities

As part of the Joint-Service JPM project, the Air Force and Navy are transferring performance measurement technology developed by the Air Force for jet engine mechanics to the Navy and Marine Corps. The feasibility study and test construction phases have been completed. Data collection will be conducted during the current year. The effort is the first of its kind and will enhance future inter-Service transfer of performance technology. It will provide additional insight into the comparison of performance measurement methods. One very viable research option, a follow-on to this effort, would be for the Services to select a common specialty (e.g., security police or personnel) and develop multiple surrogate measurement techniques for the same sample of incumbents in one Service. To avoid overloading the test incumbent with performance tests, a sample of the various techniques could be used for comparison purposes. This study would allow for direct comparison of all the Service measurement techniques and provide a tremendous data base from which to explore additional research in performance measurement. Such research might include a cost-effective/utility analysis of the various surrogate measurement techniques to determine which technique gives the maximum payoff for the least cost or how surrogates can be combined to provide the greatest amount of performance information.

In summary, benefits derived from this study will have a significant impact on performance measurement research for the individual Services, for the Joint-Service project, and for the field of industrial psychology.

References

Gould, R. B. & Hedge, J. W. (1983). Air Force Job Performance Criterion Development. Paper presented at the annual meeting of the American Psychological Association, Anaheim, CA.

Hedge, J. W. (1984). The Methodology of Walk-Through Performance Testing. Paper presented at the annual meeting of the American Psychological Association, Toronto, Canada.

SIMULATION OF INSTRUCTOR AND GROUP PROCESS ROLES WITH MICROCOMPUTER TECHNOLOGY

Barbara L. McCombs
Denver Research Institute

Many students entering military technical training not only are deficient in basic reading skills, study skills, and cognitive strategies, but they also are deficient in motivational skills (McCombs & Dobrovolny, 1982). These motivational deficiencies are reflected in trainees' inability to positively adjust to technical training requirements and implement necessary self-management, personal responsibility, and positive self-control strategies related to self-motivation (McCombs, 1984). Specific deficiencies related to unsatisfactory technical training performance include inadequate goal setting and problem solving skills, self-evaluation and planning skills, strategies for dealing with anxiety and stress, and communication skills (McCombs & Dobrovolny, 1982). A program for remedying these deficiencies, entitled the Motivational Skills Training Program, was developed by McCombs and Dobrovolny (1982) and evaluated with Air Force trainees. The program includes seven self-instructional, printed modules that have been implemented in an instructor-led, small-group format which provides trainees with the opportunity to practice new strategies and skills, share experiences, and develop feelings of rapport with their instructors and peers. Evaluation data indicated that trainees liked the program and found it helpful in their course work and personal lives. Trainees participating in the program also had significantly higher test scores and lower test failure rates than control group trainees (McCombs & Dobrovolny, 1982). Although these evaluation findings with the motivational program pointed to its success, several questions remained. One set of questions concerned training format and whether program cost effectiveness could be enhanced by reducing instructor and/or group interaction requirements through the use of computer-assisted instruction (CAI) for selected portions of the training. These questions were addressed in the reported research, undertaken for the Army Research Institute.

Background. In discussing the use of computer-based media versus conventional media such as instructors or group instruction, Clark (1983, 1984) makes the cogent point that it is not the media *per se* that influence learning. Rather, it is the content and method of instruction that are critical and the medium is merely an alternative delivery vehicle. Clark argues that in using the medium of computers, one must focus on available instructional theory in finding the necessary instructional methods for fostering the desired learning outcome. He also argues that decisions to use computers are more a matter of implementation issues such as cost, practicality, resources, and equity of access and that this medium can be maximized to address particular implementation problems by focusing on the computer's special features (Clark, 1984). Therefore, in using CAI to simulate critical instructor or group process functions, special delivery features must be carefully matched to content and function requirements.

Some features of CAI that are potentially useful to students who have responded poorly to traditional methods include individualization, mastery learning, and self-pacing. Recent research with these features, particularly the mastery learning model, however, has raised some question about their effectiveness for the disadvantaged background, low ability student (Covington & Omelich, 1981; Federico, 1981; Stinard & Dolphin, 1981; Thompson, 1980). Covington and Omelich (1981) question whether instructional features which include repeated test trials and grading against absolute standards perpetuate a negative failure cycle for low ability/low self-confidence students. On the practical side, however, Siegel and Simutis (1979) point out that within the Army there are many problems associated with providing basic skills training to large numbers of individuals at many different locations (e.g., inconsistent content quality, inconvenient training times, inappropriate matches of skill levels and basic skills curriculum). Problems such as these led ARI to explore the use of CAI for basic skills training in the Army. In initial studies to evaluate CAI for various types of skills training, Siegel and Simutis (1979) report that CAI was at least as effective as traditional instruction, particularly if instructors are given training in the roles required by the new technology.

An issue of concern in the successful implementation of CAI with a skills training curriculum, then, is the role instructors play in the learning process. At a minimum, it has been suggested that instructors have input into how CAI is used, be given short inservice workshops wherein CAI applications are explained and demonstrated, and receive meaningful role training (e.g., Bloom, 1984; McCombs & Dobrovolny, 1980, 1982; McCombs, Dobrovolny, & Lockhart, 1983; Stasz, Winkler, Shavelson, Robyn, & Feibel, 1984; Swing & Peterson, 1981). In addition, Jernstedt (1983) has argued that individualized computer technologies not be used as replacements for teachers and for group learning, and stresses the need for successful combinations of interpersonal relations and computer technologies. Interpersonal or human functions seen as important include a focus on peer relationships and cooperative goals, and defining leadership roles for students and instructors such that high task engagement results. Computer functions cited as important include frequent and varied active student interaction and the use of visual and other sensory feedback to maintain student attention.

Lubin (1984) argues that for computer-based instruction to be effective, a learning environment has to be created that mirrors the teaching/learning characteristics of live instruction. Similarly, Podenski (1984) argues that this technology should simulate ideal student-teacher interactions and free teachers for more complex tasks, such as diagnosing learning problems, helping students develop appropriate learning strategies, and monitoring instructional effects. Recent advancements now make it possible to include a rich array of audio and visual capabilities within interactive CAI lessons. Ginther (1983) discusses advances in the area of audio/speech devices that can be connected to a variety of common microcomputers. Benefits of these devices include the reduction of reading requirements, the provision of multisensory exploration of new information, and the personalization of materials. Implications for how CAI might be used to simulate group interactions can be drawn from the work of Bloom (1984), Bouton and Garth (1983), Cubberly, Omizo, & Longano (1984), Michaelson (1983), and Neale (1983). These include the use of multiple context case histories of meaningful peer problems in which students can interactively engage in identifying the problem, consequences, and alternative solutions through computer-guided inquiry, imagery, and explanations.

In summary, this selective review has identified features of CAI that can be used to simulate instructor and group process functions. In particular, careful selection of training content, careful design of CAI strategies, the incorporation of personalization through an integration of audio and visual capabilities, the identification of meaningful roles for instructors, and use of the inherently motivating qualities of this medium should contribute to the effectiveness of the CAI enhancements.

Method

Design and development of CAI/audio segments. CAI introductory and practice segments were designed and developed for each of the seven motivational skills modules. A simple computer-controlled audio capability was developed to achieve the personalization desired in the simulation of instructor and group functions. The character "PC," created to simulate instructor functions, was designed to enact three primary roles: facilitator, modeler, and motivator. In the facilitator role, PC helped students acquire new concepts, skills, and strategies via introductory explanations and practice exercises. In the modeler role, PC demonstrated the application of new concepts, skills, and strategies in guided practice segments. In the motivator role, PC coached and encouraged students to apply new concepts, skills, and strategies in both introductory and practice segments. To provide the identified group functions of peer identification, opportunities for shared problem solving, and peer modeling and feedback, a set of military characters was defined. Case studies and audio scripts were prepared for each character and for the introductory and practice CAI segments. The characters were designed to "grow" as a result of their skill training from an initial inability to solve particular problems to competent problem solvers and self-managers. This transition occurred between PC's guided CAI introductions and CAI practice sessions for each module.

Much of the modeling provided in the CAI segments is accomplished via the audio enhancements, while most of the actual skill practice is provided through the CAI exercises. These two functions are integrated into a single unified presentation via the audio interface in the following ways: (a) the CAI screen reinforces audio information, (b) the audio information precedes a CAI segment or frame; (c) the audio information follows a CAI segment or frame; or (d) the audio is an integral part of a CAI segment or frame. Once the audio tapes had been recorded, pulses were added at points that coincided with CAI screen changes. The contractor- developed audio interface consists of a specially designed interface card which plugs into the Apple IIe game I/O port. The interface receives the pulses from a standard slide-sync audio cassete player. These pulses trigger screen changes and, in turn, allow the CAI software (in this case, the Apple SuperPILOT Authoring System) to control the on/off function of the audio player. This capability allows for computer control of a linear sequence of audio messages that coincide with particular CAI frame sequences, as well as provides for the personalization of skill training introductions and practices, at about one-eighth the cost of videodisc technology.

Experimental design. Six experimental conditions were defined: an historical control group (HC), current control group (CC), a CAI introduction and practice group (CAI), a CAI introduction and instructor practice group (CAII), an instructor introduction and CAI practice group (ICAI), and an instructor introduction and practice group (II). Other independent variables included student scores on the General, Electrical, and Clerical subscales of the Armed Services Vocational Aptitude Battery (ASVAB); military rank; sex; initial judgments of self-efficacy; and initial indices of anxiety and ability to cope with stress. Dependent variables included time to complete the first and second course segments; test failure rates in the first and second course segments; progress index for the entire course; and whether students attrited or graduated. (See McCombs et al, in press, for a description of measures used.)

Subjects. Participants in the study were male and female students in the Electronic Communications (EC) school at Ft. Sill. Students were assigned to one of the five current experimental conditions by designated Ft. Sill personnel, using guidelines and procedures specified by the contractor. At the conclusion of the study, data on a total of 479 students were available for analysis. The number of students in each condition were as follows: 53 in the HC condition, 253 in the CC condition, 55 in the CAI conditon, 61 in the CAII condition, 53 in the ICAI condition, and 57 in the II condition.

Procedures. A 16-hour training program for three Ft. Sill EC course instructors acquainted them with the purpose of the evaluation and the training program, provided guidelines for introducing each module and conducting the small group practice sessions, and described procedures for implementing each experimental condition. Instructors were also trained in the content of the CAI materials and operation of the CAI equipment. A workshop format was used to ensure comparability between the CAI and instructor conditions, instructors were provided with scripts of the case studies applying to each CAI character which they could use as part of their introductions and practice sessions for each module. The pretest measures were administered to students on the day they began the EC course. Following completion of the pretests, students were assigned to control or experimental groups per contractor procedures. The number of students assigned to experimental conditions in any given class varied from 10 to 15, depending on the class size. The length of training averaged 25 hours and was conducted for 6 hours a day for 4 to 5 days, depending on the rate at which students completed the training. Five Apple IIc systems were available for the CAI conditions. Following the completion of the motivational training, experimental students began their regular EC course work. The EC course is divided into two portions, the first of which is approximately 4 weeks long and the second of which is approximately 5 weeks long, for an average total length of approximately 9 weeks. The course is implemented in a self-paced mode wherein times to complete vary as a function of student ability and motivation. Postmeasures were administered by Ft. Sill personnel in the EC course at the end of the first course portion.

Results and Discussion

Findings. Study results indicated that students who received the motivational skills training with instructor introductions/group practice (II condition) performed better during the subsequent EC course than either students who received no motivational skills training (CC condition) or students who received the motivational skills training via CAI with no instructor practice (CAI and ICAI conditions). This better performance was manifested in significantly fewer test failures and less training time. In addition, as compared to the CC condition, students receiving either the CAI or II conditions tended to have lower attrition rates. Although these differences were not statistically significant, the 8.2 percent reduction for the CAI condition and the 4.4 percent reduction for the II condition may have some practical significance in terms of training costs. For similar students going through the EC course a year earlier (the HC condition) attrition rates were 5.2 percent higher than for students in the current study and progress indices were approximately 12 percent higher. These findings for the HC as compared with current groups were at least in part a function of new procedures implemented in the student battery (housing area) that required mandatory study periods for students not progressing through the EC course at a satisfactory rate. Thus, it is likely that findings with the Motivational Skills Training Program were attenuated in the present study.

Since study findings indicated no overall superiority of the CAI vs. II conditions, some exploratory analyses of potential individual dfferences in subsequent student performance as a function of treatment condition were conducted. These analyses generally indicated that the CAI condition was at least equally effective for approximately half of the students (i.e., those students of high general ASVAB abilit, and those students with low perception of competence or self-esteem). These findings imply that there may be systematic and reliable individual differences that could be used in differential treatment format assignments, thereby reducing some of the instructor support requirements for the motivational program as well as capitalizing on the use of microcomputer technology for this type of skill training.

Implications of study results. A major assumption of this study was that well designed CAI introductory and practice segments, implemented via a rich microcomputer/audio technology mix, could provide the degree of personalization and simulation of critical instructor and group functions to offset or partially reduce instructor and group process requirements for this type of motivational skills training. Main effect findings with various combinations of CAI and instructor/group supported conditions did not bear out this assumption. This raises the possibility that, in line with Clark's (1983, 1984) arguments concerning the influence of media on learning, that the CAI/audio combination was not sufficiently matched to the content and method of instruction required in this training context. The main effect findings also reinforce

Jernstedt's (1983) point that CAI technologies cannot be used as replacements for teachers and group learning, and that there is a need for a synergetic combination of the human and computer functions to achieve maximum instructional effectiveness.

The exploratory individual difference analyses have suggested, however, that the CAI enhanced version of the motivational training was at least equally effective for some types of students. Of the individual difference variables available for inclusion in these exploratory analyses, the findings with the general ability measure (ASVAB General) are not surprising. That is, a number of studies have found CAI or other multimedia treatments to be as effective as traditional instructor/group methods for high ability students (e.g., Clark, 1984; Kulik, Bangert, & Williams, 1983). On the other hand, the findings that students low in perceptions of competence subsequently perform better if they received the CAI enhanced version vs. the instructor/group version are somewhat puzzling. A plausible explanation, however, may be derived from Bandura's (1982) theory of self-efficacy which suggests that individuals low in perceived competence do not judge themselves as capable of handling particular situations, including interpersonal situations. Because of their low feelings of personal adequacy, they are often threatened in interpersonal situations and fearful of having their perceived inadequacies exposed. For these individuals, then, it is reasonable that the nonhuman medium of CAI may provide a less threatening learning environment, particularly for this type of self-development training that requires considerable self-analysis and self-exposure. As Bowman (1982) has argued, CAI advantages include (a) freedom from fear of reprisal, ridicule, or rejection; and (b) provision for active involvement in tasks that are based on a high probability of success. Thus, it may be that the medium of CAI is a more optimal treatment for those students whose initial perceptions of self-efficacy are low.

The preceding speculations need to be verified by further research. As noted by Siegel and Simutis (1979), CAI's potential lies in its ability to provide individualized, standardized, and efficient instruction, particularly to adult learners who require remedial training. Individualization issues with microcomputer technology thus need to be systematically explored, such that differential assignment to this medium can improve training performance and lead to more cost-effective use of human personnel. There is little question based on the research reported here that instructors and the group process play a critical role in the success of motivational training and the success of individualized computer-based approaches in general (McCombs, in press). Continued research with microcomputer/audio technologies that can simulate instructor and group requirements promises to contribute to a realization of the full benefits of this technology.

References

Bandura, A. (1982). Self-efficacy mechanism in human agency. American Psychologist, 37(2), 122-147.

Bloom, B. S. (1984). The search for methods of group instruction as effective as one-to-one tutoring. Educational Leadership, 41, 4-17.

Bouton, C., & Garth, R. Y. (1983). Students in learning groups: Active learning through conversation. In C. Bouton and R. Y. Garth (Eds.), Learning in groups. San Francisco: Josey-Bass Inc., Publishers.

Bowman, R. F., Jr. (1982). A "Pac-Man" theory of motivation: Tactical implications for classroom instruction. Educational Technology, 22(9), 14-16.

Clark, R. E. (1983). Reconsidering research on learning from media. Review of Educational Research, 53(4), 445-459.

Clark, R. E. (1984, April). Learning from computers: Theoretical problems. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Covington, M. V., & Omelich, C. L. (1981). As failures mount: Affective and cognitive consequences of ability demotion in the classroom. Journal of Educational Psychology, 73(6), 796-808

Cubberly, W. E., Omizo, M. M., & Longano, D. M. (1984, January). The effects of group counseling on self-concept and locus of control among learning disabled children. Paper presented at the annual meeting of the Southwestern Educational Research Association, Dallas.

Federico, P. A. (1981, April). Individual differences and mastery learning in computer-managed instruction. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Ginther, D. W. (1983). Micro's are talking back in the classroom: The promise of speech technology in education. Technological Horizons in Education Journal, 11(2), 105-107.

Jernstedt, G. C. (1983). Computer enhanced collaborative learning: A new technology for education. Technological Horizons in Education Journal, 10(7), 96-101.

Kulik, J., Bangert, R., & Williams, G. (1983). Effects of computer-based teaching on secondary school students. Journal of Educational Psychology, 75(1), 19-26.

Lubin, D. A. (1984). What's right with CBT? Data Training, 3(7), 18-19, 22.

McCombs, B. L. (in press). Instructor and group process roles in computer-based training. Educational Communication and Technology Journal.

McCombs, B. L. (1984). Processes and skills underlying continuing intrinsic motivation to learn: Toward a definition of motivational skills training interventions. Educational Psychologist, 19(4), 199-218.

McCombs, B. L., Bruce, K. L., & Lockhart, K. A. (in press). Enhancements to motivational skill training for military technical training students: Phase I evaluation study report. Alexandria, VA: Army Research Institute.

McCombs, B. L., & Dobrovolny, J. L. (1982, December). Student motivational skill training package: Evaluation for Air Force technical training. (AFHRL-TP-82-31). Lowry AFB, CO: Air Force Human Resources Laboratory.

McCombs, B. L., Dobrovolny, J. L., & Lockhart, K. A. (1983, June). Evaluation of the CMI instructor role training program in the Navy and Air Force. (NPRDC-SR-83-43). San Diego, CA: Navy Personnel Research and Development Center.

Michaelson, L. K. (1983). Team learning in large classes. In C. Bouton and R. Y. Garth (Eds.), Learning in groups. San Francisco: Josey-Bass Inc., Publishers.

Neale, D. C. (1983, April). Specifications for small group activities in instructional designs. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Podemski, R. S. (1984). Implications of electronic learning technology: The future is now! Technological Horizons in Education Journal, 11(8), 118-121.

Siegel, M. A., & Simutis, Z. M. (1979, February). CAI for adult basic skills training. Two applications. Computer based education: Mission of the future, Volume 3, Proceedings of the Annual Convention of the Association for the Development of Computer-Based Instructional Systems, San Diego, CA.

Stasz, C., Winkler, J. D., Shavelson, R. J., Robyn, A. E., & Feibel, W. (1984, April). Staff development for instructional uses of microcomputers: The teacher's perspective. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Stinard, T. A., & Dolphin, W. D. (1981). Which students benefit from self-paced mastery instruction and why. Journal of Educational Psychology, 73(5), 754-763.

Swing, S. R., & Peterson, P. L. (1982). The relationship of student ability and small group interaction to student achievement. American Educational Research Journal, 19(2), 259-274.

Thompson, S. B. (1980). Do individualized mastery and traditional instructional systems yield different course effects in college calculus? American Educational Research Journal, 17(3), 361-375.

ASSESSING M1 TANK COMMANDERS WITH A COMPUTERIZED HAND-HELD TUTOR

Brent Bridgeman
Educational Testing Services
and
Theodore Post
Essex Corporation

Tank commanders on the M1 Abrams tank must be trained to very quickly evaluate a battlefield situation, identify the target which should be engaged first, choose the appropriate weapon and/or ammunition (from among three machine guns and two types of main gun ammunition). issue the appropriate fire command to the loader and gunner, direct the driver to move or change direction if necessary, and maintain communication with other tanks in the unit. For such complex tasks, realistic hands-on training is clearly essential to reach a level of effectiveness, or even survival, on the modern battlefield. But hands-on training is extremely expensive because of high costs of equipment and ammunition as well as the costs of transportation to the few areas that can support full-scale field exercises. In order to make better use of the very limited time for hands-on training, the commander trainees should have already mastered the basic prerequisite skills before going into the field. Thus, for example, if the trainee in the class practices basic elements of a fire command until they are virtually automatic, he is more likely to be able to use them correctly under the multiple pressures in the field environment.

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) has been developing a number of approaches to making traditioinal training more effective. Several projects employ microcomputers combined with videodisks. Such systems are reasonably realistic and are relatively inexpensive compared with hands-on training, but they are still costly enough that the amount of time a given soldier can spend working with the system is limited. In addition, such systems are not easily portable so that soldiers must come to a fixed location at a fixed time to work with the system. Therefore, as a supplement to microcomputer/videodisk technologies, ARI sponsored the development of a low-cost, hand-held computerized Tutor. The intention was to make a device that was low enough in cost (under $150) and small enough in size (no larger than a notebook) that it could be used by soldiers in much the same way that they could use a textbook or manual. The Tutor that was developed as a result of this initiative is a 10" x 11" x 2" device with a 32 character dot matrix display screen on the top, a keyboard on the bottom (numbers 0-9, letters A-E and three operational keys: SAY, ERASE and GO) and an indentation in the center that holds an open 5" x 5" booklet. The use of a printed booklet for the display of test questions, instructional text, and graphics permitted substantial cost savings compared to systems that store this type of textual and graphical information in the computer's memory and display it on a CRT. The Tutor also contains a digitized speech system. (For a more complete description of the Tutor, see Fertner and Bridgeman, 1985).

The Tutor was originally intended to teach technical vocabulary. Its three independent (but mutually supporting) courseware components are — (a) an instructional sequence including a pretest and explanatory text with embedded questions, (b) a drill and practice session (called Word War) in which items answered incorrectly initially are presented again after just one other item has been presented, again after three more items have been presented, etc., (c) a game (called Picture Battle) requiring recognition of an appropriate picture (or portion of a picture) given a spoken stimulus. Because of the success of this approach for vocabulary instruction (Bridgeman and Wisher, 1985), ARI focused attention on other areas where this technology could be applied. One of these areas was instruction in issuing fire commands for M1 tank commanders. A contract was awarded to Educational Testing Service (and its subcontractors Advanced Technology Laboratories and BioTechnology Inc.) to adapt a set of existing instructional booklets for presentation on the Tutor. The remainder of this paper describes how the Tutor's three operational modes (pretest and explanation, Word War, and Picture Battle) were used for fire commands instruction.

Pretest and Explanations

Each of the 28 instructional units begins with a brief multiple-choice pretest. The questions and answer choices are printed in the book, and the soldier responds by pushing one of the A-E keys on the keyboard. After the pretest, the Tutor instantly evaluates the soldiers performance , and allows soldiers with no errors to proceed immediately to the next unit. For soldiers who made errors, the Tutor displays the answer choice they selected followed by the correct response. Thus, this test review becomes the first step in the instructional process.

Next, soldiers who made errors on the pretest are directed to begin reading the explanatory text. Frequent questions are sprinkled throughout the text to ensure that attention and comprehension are maintained. The Tutor provides immediate corrective feedback on these items, but errors should be rare if the soldier is reading carefully.

In the vocabulary module, the "SAY" key was used to make the Tutor pronounce target words that were underlined in the text with code numbers under them. The soldier pushed "SAY" and then the number under the word that he wanted to hear. In the fire commands instruction the use of the "SAY" key is quite different; it is used to provide a brief explanation of incorrect answers. Instead of entering responses on the A-E keys, the soldier makes a selection by pushing "SAY" then 1, 2, 3, or 4. This activates the Tutor's digitized voice system. For example, a page in the book shows a picture of a battlefield scene with several potential targets labeled 1 to 4, and the text asks the student to identify the most dangerous threat. For one incorrect answer. The Tutor says "No, out of range, try again" and for another incorrect answer it says "No, can't kill you, try again." Through this type of oral feedback the soldier learns not only which answers are incorrect, but why they are incorrect. Although this information could be presented through the display screen instead of the voice system, it would be considerably more distracting, as the soldier would have to repeatedly shift attention from the picture of the battlefield scene to the display screen. Thus, while the use of voice technology in this application was not as crucial as it was for word pronunciation in vocabulary instruction, it still adds a different, valuable dimension to the instructional process.

Word War

The logic behind the increasing ratio review used on Word War (Siegel and DiBello, 1980) applies to many different rote memorization tasks, not just vocabulary learning. In the fire commands module it is used to provide practice in weapon/ammunition selection, and in the identification of the name of threat weapons. For the weapon/ammunition selection routine, the screen first presents a situation (e.g., T-72 at 1000 meters). Followed by three answer choices, presented one at a time (e.g., SABOT, HEAT, M-240). The soldier is instructed to push "GO" when the correct answer appears on the screen. The Word Wars for threat weapon identification were added when data from a preliminary field trial indicated that some soldiers had difficulty reading the text because they did not know, for example, that an SPG-9 is a recoilless anti-tank gun.

Picture Battle

In this game-like activity the Tutor's voice system asks a question based on a picture in the book, and the soldier responds on the keyboard. The display screen is used to keep score. With each correct answer a "projectile" formed by the dots on the display screen moves one step from left to right across the screen. For each incorrect answer an "enemy" projectile moves across the screen in the opposite direction. The object of the game is to destroy the enemy target before the enemy destroys you. Hitting the enemy target is accompanied by sound effects of a shell exploding. For vocabulary instruction, Picture Battle was used to reinforce an association between the spoken name of an object and a pictorial representation of the object (e.g. the Tutor would say "equilibrator" and the solider would find the picture of an equilibrator), eliminating entirely the need to read anything. Although fire commands instruction did not require a task with no reading requirement, the game-like score keeping features of Picture Battle could still be used to good advantage. In one implementation, a battlefield scene is pictured and a situation described in the text; when the Tutor's voice asks "Initial Fire Command" the soldier must select the appropriate fire command for the pictured situation. Immediate corrective feedback is provided and the appropriate projectile advances across the screen, then the soldier is asked to turn to the next battlefield scene and the process is repeated.

Conclusions

As an aid to assessment and training of tank commanders, the Tutor falls on a continuum between traditional paper-based materials and microcomputer videodisk systems. The Tutor lacks the flexibility and moving graphics capabilities of interactive videodisks, but it is a fraction of the cost of such systems and is easily portable. Although the Tutor is more costly than purely paper-based materials. it provides a much more interesting and interactive environment for assessment and instruction: soldiers may be immediately branched to more difficult material based on pretest scores, visual and auditory explanatory feedback is provided, drill and practice exercises in which the item presentation order is varied depending on student performance are available, and interactions with game-like visual and auditory scoring features are included. Every training technology has advantages and disadvantages, and the hand-held computerized Tutor appears to have a place in the instructional arsenal.

REFERENCES

Bridgeman, B. & Wisher, R. Development of a hand-held computerized vocabulary tutor. Machine-Mediated Learning, 1985, 1(3).

Fertner, K. & Bridgeman, B. Increasing the effectiveness of machine mediated tutoring using embedded testing. Proceedings of the 26th Annual Conference of the Miltary Testing Association, 1984, 75-80.

Siegel, M. & DiBello, L. Optimization of computerized drills: An instructional approach. Paper presented at the Annual Meeting of the American Educational Research Association, April 1980.

Disclaimer

The views expressed in this paper are those of the authors and do not necessarily reflect the official policy of the Department of Defense, the Army Research Institute for the Behavioral and Social Sciences, or the U.S. Government.

MICROFORM IN TRAINING

Lieutenant Commander K Jones, BSc., Royal Navy, Staff of the Royal Naval School of Educational and Training Technology.

Lieutenant Commander A E Mizen, BPhil(Ed)., Royal Navy, Staff Officer to Commander in Chief Naval Home Command.

BACKGROUND

1. A major percentage of the technical training of officers and ratings conducted by Naval Establishments has traditionally used hard copy Books of Reference (BRs) and other Technical Publications, supplemented by a series of professionally made training aids. However, the increasing complexity of HM ships has led to a large increase in the documentation required to operate, maintain and repair ships equipments and systems such that a modern frigate carries the equivalent of some quarter million A4 pages weighing some 1½ tons. Microform, having already demonstrated considerable advantages in many areas, has been chosen as the most suitable media for more general naval use, and in particular for technical BRs and publications. Consequently, for training, the withdrawal of hard copy BRs and the introduction of microform requires the conversion of microform to readable print with the aid of readers, projectors and printers.

2. The Royal Naval School of Educational and Training Technology (RNSETT), and the Royal Naval Submarine School (RNSMS), were tasked to identify one or more microform projectors that could operate under the same conditions and to the same standards as conventional overhead projectors, and thereby enable the establishment of a teaching strategy for microform in the classroom. (1)

SCOPE OF THE INVESTIGATION

3. This investigation considered:

a. The type of microform projectors available (forward or rear projection).

b. The applicability of the microform projector to the classroom environment.

(1) Investigation into Microform Projectors, March 19[illegible], Lt Cdr P C Tovey B Ed., RN, and Maj R K Morrison M Ed., RAEC.

c. The most appropriate teaching strategy for use with microform.

d. The range of subject matter to be taught using microform.

CLASSROOM TRIALS

4. Trials were conducted, in the classroom environment, on a range of commercially available microform projectors. Additionally some desk-top readers were also trialled.

5. Each of the projectors was evaluated under similar conditions using an instructor to give a lesson which was observed by:

a. A subject matter expert.

b. An Instructional Techniques officer.

c. A Training Design officer.

d. A Quality Control officer.

e. An officer from the RNSETT Instructional Techniques (IT) Group.

MICROFORM PROJECTION

6. Forward Projection. Each of the microform projectors was evaluated to assess its capability to project an image onto a screen (as a conventional OHP). In all cases the image projected was of poor quality and could only be read by students close to the screen. The edges of the image were blurred, definition was poor and key-stoning presented a problem. After a few minutes there was evidence of eye strain and the instructor found it increasingly difficult to maintain student concentration. This was aggravated by the fact that the projectors could only operate in a darkened room. Used in this mode the image projected was unacceptable to both students and instructor.

7. Rear Projection. At the time of the trial only one projector had a rear projection capability. When used with its A2-size screen it could operate under normal classroom conditions with up to 6 students, comfortably, around the screen at any one time. In this mode the image was clear and eye strain no longer appeared to be a problem. However, the instructor could not teach using the strategy of the lesson and a tutorial style had to be adopted.

8. Compatible Desk-Top Readers. It was found that if a rear projection microform projector was to be used in the classroom, students needed desk-top readers for back-up and consolidation. These readers were operated under the same conditions as the projector.

APPLICABILITY TO THE CLASSROOM

9. The investigation confirmed that microform projectors cannot replace the conventional OHP and should not be considered as such. However, in the rear projection mode they are a most useful training aid and, in this mode, rather than replace the OHP they should be used to complement it.

10. The microform projector should be just another training aid available to the instructor. To be used successfully, however, class numbers would have to be small.

TEACHING STRATEGY

11. As already indicated the rear projection microform projector is suitable only for small class teaching and the tutorial style of instruction would appear to be the most appropriate. This may be an entirely new way of teaching for many instructors. Used in this way the microform projector has the added advantage that the instructor can focus the attention of the students onto the material being presented and as a result have more control of the learning situation (which could never be guaranteed using hard-copy publications).

12. Consequently it is clear that teaching from BRs will no longer be possible, it will be teaching with BRs. Used correctly, therefore, microform may well improve the quality of instruction. Instructors will require training in the use of microform in the classroom and the tutorial method. Properly trained in the correct techniques instructors may overcome some initial qualms about using microform.

SUBJECT MATTER RANGE

13. Microform projectors enable a number of students simultaneous access to BR text and diagrams. It should be remembered, however, that these are photographed pages and as viewgraphs are of the very worst type; as such they are unacceptable. Similarly microform projectors should not be used to present material that can already be done so by using the OHP and viewgraph.

14. The use of microform BRs in training will require the instructor to direct the attention of students to relevant pages of BR text and diagrams. In this situation the instructor is teaching students how to use the microform BRs and as such could be considered a skill. The microform projector is suited to this task although the question of indexing pages, cross-referencing and the requirement to view more than one page at a time will require careful consideration. This may be alleviated by the use of two microform projectors. Other aspects of instruction should be conducted in the traditional manner, emphasising the need for microform projectors to be able to operace under the same conditions as other training aids.

FUTURE DEVELOPMENT

15. Microform projectors are being continually improved. There are now available larger screen rear projection models than the one trialled. There has also been an increase in the number and quality of desk-top readers.

SUMMARY

16. In sum the investigation concluded that:

a. Microform projectors cannot operate under the same conditions or to the same standards as the conventional OHP.

b. Of the microform projectors evaluated, only the rear projection model could operate satisfactorily in the classroom environment.

c. To be used successfully microform projectors must be employed as additional training aids and must be able to operate under the same conditions as OHPs and orher aids.

d. The microform projector should only be used by the instructor to direct the attention of students to the relevant pages of the BR text and diagrams. It should not be used to present material that is normally presented by the OHP and viewgraph.

e. Used appropriately, microform projectors could improve the quality of instruction currently undertaken using hard-copy publications.

f. Microform should be used to teach small groups of students and is more suited to the tutorial mode of instruction. For many instructors this will be an entirely new way of teaching and they will require training in the use of microform in the classroom. The RNSETF has already designed an instructional module for 'teaching with microform' as part of their Instructional Techniques Course.

Selling Job Analysis

Stanley D. Stephenson
School of Business
Southwest Texas State University
San Marcos, Texas 78666

To the job analysis practitioner, the value of job analysis is patently obvious. Moreover, the value of job analysis data is more and more becoming recognized by others; e. g., witness the increase in articles about various personnel functions that are based on having an accurate job and/or position description, a situation that can only occur when detailed position information is available. There has also been a corresponding increase in the number of books published on the topic of job analysis.

With this growing awareness of the value of job analysis data, one would expect there to be a general eagerness to engage in job analysis projects. However, such is not necessarily the case. When it comes down to releasing the monies to conduct a job analysis (and a job analysis can be expensive), many managers fail to grasp the overall value of job analysis and elect to spend their funds on projects that are more immediate, and perhaps better understood.

The primary problem is that those who truly want the job analysis done (e. g., the trainers, the job evaluators, etc) are not the ones who control the funds; i. e., the manager. On the other hand, the outside consultants, who are contacted to submit proposals to perform a job analysis, initially talk primarily with those who want the work done and not with the managers. However, when it appears that funds are not to be forthcoming, those desiring the project will often arrange a meeting between their manager and the job analysis consultants. Although the meeting is advertised as a meeting of three groups (the manager, the job analysts, and those wanting the project), the meeting often turns out to be a direct discussion between the consultants and the manager whom they have only recently met. Moreover, the manager, since he or she has already indicated that funds may not be available for the project, will be somewhat defensive and will have erected a barrier to counter what is anticipated to be a strong sales pitch. It soon becomes apparent that support from the third party in the meeting (i. e., those who want the project) is non-existent; these individuals are hoping that something (which will not harm their careers within the organization) will happen to change their manager's mind. Consequently, the job analysts not only have to overcome a negative predisposition on the part of the manager, they also have to do it alone.

The underlying problem is that managers usually have but a limited grasp of the value of job analyst data. For instance, they may know about its value in designing a training program but be uninformed about its usefulness in designing a performance evaluation system. Such a situation can be easily understood when it is remembered that usually one function within personnel training is doing the pushing behind the job analysis request. Their views are presented to the manager, and these views become the sole basis for the manager's perceptions of the value of job analysis data. Consequently, the overwhelming objective for any job analyst making a sales pitch is to break these preconceived but limited notions and to make the manager aware of the tremendous wealth of information that is available from the collection of job analysis data. Once a manager realizes the indispensable and varied aid that job analysis data can be to the organization, the manager may be more willing to release funds for the project.

This paper briefly reviews the variety of ways in which job analysis data can be of use to an organization. These uses will be discussed with regard to both military and non-military organizations. Next, a suggested approach for expanding an organization's awareness of the value of job analysis data will be demonstrated using actual data.

Uses of Job Analysis Data

Job analysis data are useful in the broad areas of selection, training, production, performance evaluation, promotion, compensation, termination, and termination/retirement. These areas will discussed in the following sections. Not initially obvious military applications will be highlighted.

Selection

Job analysis data allows a detailed job description to be prepared. Such a description is essential to the proper hiring of new employees; if the nature of the job is not known, how can a good hiring decision be made? For instance, research has shown that realistic job previews (RJP) are beneficial in terms of initial hiring and retention: job analysis data lends itself to the preparing of RJPs.

Managers in the military do hire/select personnel to come into their units. In fact many military managers spend considerable time reviewing the files of military personnel. Moreover, military managers go through lengthy interview sessions to hire civilian replacements. The military manager's decision would be greatly enhanced if he or she spend as much time coming to grips with the nature of the job being filled as with the nature of the applicant. However, such information is frequently lacking, and therefore managers spend their time on information that is available; e. g., applicants' personnel folders.

Training

Training can have several purposes. First, it can be used to prepare an individual for entry into the work force. Second, it can be used to maintain proficiency. Third, it can be used to prepare an individual for advancement. In all cases, knowing the job that is currently being done and/or the job to which an individual aspires are crucial to designing and implementing a good training program.

Most military managers realize the first two purposes listed above. However, many managers do not give proper emphasis to preparing their personnel for advancement, often because the "next" job is not understood. Related to this discussion of training is the concept of backup capability. A good manager insures that his personnel have the capability to backup key personnel during periods of illness, vacation, etc. Job analysis data can provide a basis for determining which other position in the organization is the most likely candidate for the backup function. Coming to grips with the concepts of both advancement and backup also permits the military manager to better understand job progression, and in the long term the manager is better able to improve the morale and productivity of his workers.

Production

Job analysis data can permit the manager to improve productivity in two, general ways. First, it can possibly isolate the differences in task performance of high performers versus low performers. Second, it can provide a picture of how the entire work force "fits" together; i. e., identify the gaps or overlaps in accomplishing the assigned mission. A military manager could make much better work design decisions with job analysis data versus having to rely on direct observations alone.

Performance Evaluation, Promotion, and Compensation

These three factors are directly linked. First, performance must be evaluated. Obviously, a proper evaluation can not be accomplished without knowing the details of the job, yet research has frequently shown that supervisors do not know the details of their subordinates work. Second, promotion is a direct end product of merging the performance in a current job and a knowledge of the next higher level job.

The subtleties of the military promotion system are well known. The military manager who understands the job in question, and has a feel for the best next assignment, is in a much better position to aid his or her subordinates in terms of job progression which in the military directly relates to promotion.

Compensation is a direct function of logically tying salary to worth to the organization; i. e., a classification procedure. The basic method for doing this is to conduct a job evaluation. The need for job analysis data as the basis for any job evaluation is well documented, and most civilian managers are aware of this fact. However, although informal job evaluation does occur, compensation issues do not appear to be a relevant issue for the military.

This brief review demonstrates that job analysis data really is an organization-wide management tool and not just a specific personnel tool. Other aspects not discussed are morale, affirmative action, comparable worth, labor negotiations, and legal aspects in general; job analysis data are also of benefit in these areas.

The bottom line is that when job analysts meet with a manager, he/she needs to be shown how job analysis data can be useful in a larger context than originally thought. It is the "big" picture that will convince a manager that a job analysis project should be funded. Once a project is funded, of course, the specific goals of different personnel/ training functions will also be met.

The above argument should not be dismissed as being appropriate just for the non-military audience   The military manager shares many similarities with his civilian counterpart. Both groups have a vested interest in morale, productivity, training, etc. Moreover, with a steady trend toward converting military to civilian positions. military managers are increasing becoming involved with managing a civilian work force. Consequently, many of the points made in this paper apply to military as well as civilian managers.

Increasing Job Analysis Awareness; An Applied Example

Recently, a job analysis was done of clerical workers employed by a small state agency in Texas; small in the sense that there were only 10 clerical positions surveyed. Job titles were: File Clerk, Receptionist, Clerk, Secretary III, Secretary II, Secretary I, and Office Services Supervisor. This small group of clerical positions has become very useful in demonstrating to others how job analysis can be used in a variety of ways.

Prior to meeting with the management of a different organization, a copy of the clerical task inventory is sent to the organization with a request that a mid-level clerical worker complete it. The completed inventory is processed, and the results are compared to the "population" of 10 workers described in the preceding paragraph. Examples of the utility of job analysis data are demonstrated and discussed with management. An illustration of how this procedure actually works is presented below.

In this case, the inventory was completed by a job incumbent working in a position with the title of Administrative Secretary in the target organization. Task responses produced the following results.

Table 1
Task Cluster Percent Time Spent Results

Task Cluster (Duty)	Percent Time Spent
Written Communications	15.12
Verbal Communications	14.40
Distribution	17.28
Financial	3.60
General Filing	3.60
Use of office machines	3.60
Process general paperwork	0.72
Typing	15.12
Management Assistance	1.44
Miscellaneous	3.60
Office Management	11.52
Forms Management	0.72
Special	3.60
Special - Customer Service	0.72
Special - Dictation	0.72
Special - Notary Public	3.60
Supervision	0.72
Total	100.08

These results could assist management in all of the areas discussed earlier. However, I have found that in a selling session managers like to be shown how job analysis data can be immediately used. Consequently, for the purpose of this paper, two immediate uses are highlighted: backup capability and training for advancement.

Backup Capability

Job analysis data can immediately inform a manager whether or not a key worker's responsibilities can be met when he/she is absent. Comparing this worker's tasks performed with the hypothetical population reveals that for the most part this worker's duties can be covered by other workers should he or she be absent. In some task cluster areas, multiple coverage is available; i. e., more that one other position is performing the identical tasks. Such a situation is convenient for it allows a manager to efficiently spread the workload should the worker in question be absent. The major selling point, however, is that a manager would know exactly where to turn to insure that an absent worker's total responsibilities are covered.

A more detailed analysis reveals that the worker in question performs some unique tasks. In this case, specific task coverage can also be aralyzed; analysis results are presented in Table 2.

Table 2
Specific Administrative Secretary
Task Coverage by Other Positions

Task	Covered by:
Edit Administrative Material	Sec II, Supervisor
Write Draft Minutes	Supervisor
Answer Inquiries	Supervisor
Draft Communications	Clerk, Sec III
Process Payment Vouchers	Clerk
Acknowledge Invitations	Sec III, Receptionist
Conduct Tours	NOT COVERED
Transcribe Dictation	Receptionist
Perform Notary Public Duties	Sec III

Obviously no one other worker can cover these unique and/or critical tasks. Simply put, a manager would have to delegate these tasks to different workers in order to have the work accomplished. Interestingly, for some tasks the best coverage is provided by a Clerk while for others coverage is provided by an Office Services Supervisor. More importantly, there is no one in this population who presently conducts tours. This one piece of information should forewarn a manager about a possible lack of coverage should a key worker be absent. Such information would be perceived as being very useful by any manager.

Training for Advancement

Good managers always look out for their employees, even to the point of insuring that they are prepared for advancement. Job analysis data permits managers to immediately determine the skills needed to perform at a higher job level. For instance, in the example being used, the position in question (Administrative Secretary) would be equivalent to a Secretary II or Secretary III in the comparison population. This comparison was based on a similarity of tasks performed analysis. The next higher job in this population would be Office Services Supervisor. Table 3 presents a summary of the major time spent differences between the present position and the next higher position.

Table 3
Percent Time Spent Differences
Office Services Supervisor Versus Administrative Secretary

Task Cluster (Duty)	Percent Time Spent	
	Administrative Secretary	Office Services Supervisor
Written Communications	15.12	1.01
Verbal Communications	14.40	4.28
Distribution	17.28	3.52
Education/Training	0.00	4.03
Financial	3.60	0.75
General Filing	3.60	0.50
Use of office machines	3.60	1.51
Process general paperwork	0.72	1.01
Typing	15.12	6.55
* Management Assistance	1.44	7.78
Miscellaneous	3.60	2.51
* Office Management	11.52	18.86
* Plan/Organize	0.00	8.55
Forms Management	0.72	8.79
Social	3.60	4.47
Special - Customer Service	0.72	0.75
Special - Dictation	0.72	0.00
Special - Notary Public	3.60	0.00
* Supervision	0.70	20.87
Other:		
Review/prepare documents		1.76
Assist in posting payroll		1.26
Retrieve stored information		1.51
Total	100.08	99.56

* Skills not used in the present position but used in the higher level position.

These results highlight the skills that the administrative secretary should be acquiring while still in his/her present position. Then when an opening occurs, the individual would have the necessary training record to compete for the new position. However, unless someone is aware of the differences between the two positions, a proper on-the-job training program can not be created. Job analysis data is that awareness.

These two examples, backup capability and advancement, illustrate how job analysis data can be put to immediate use by management. Similar examples of how the same data could be used in the other areas discussed earlier could be presented. For instance, a very realistic, realistic job preview could be quickly and accurately created if job analysis data were available. But, the point being made is that job analysis data is not just for training, not just for creating job description, not just for conducting job evaluation, etc. Job analysis data is for managers to use in a wide variety of ways. The better you able to deliver this message, the better will be your chances of having the job analysis program accepted. The approach outlined here will help you do just that.

LINKING WORK, TRAINING, AND PROMOTION IN THE COAST GUARD

Karen N. Jones and John A. Burt
U. S. Coast Guard Institute, Oklahoma City, Oklahoma

The Coast Guard has begun correcting a long-standing problem in its enlisted personnel system -- the mismatch between the work performed by Coast Guard personnel and the content of our training courses and promotion examinations. This mismatch was caused by the lack of an accurate description of the work performed by Coast Guard enlisted personnel.

The work performed by enlisted personnel is described by the enlisted qualifications for advancement (quals). The quals are a series of occupational duty statements describing the jobs performed in each rating (job occupation) at each paygrade. The quals are the basis for the Coast Guard's promotion examinations and rating-specific training. As the primary user of the quals, the training community was aware the quals for many ratings were incomplete and out-of-date. However, the training community did not have the responsibility and authority to correct the problem -- the rating manager has the responsibility for ensuring the quals reflect the minimum occupational standards currently required for the rating.

The rating manager (called rating program manager or program manager in the Coast Guard) for each rating is in the headquarters office responsible for managing the operational program or programs that rating supports. For example, the rating manager for Machinery Technician 's in the Office of Engineering. In most instances, a general duty officer is assigned the responsibility for managing a rating or group of ratings as a collateral duty. Since these officers are usually generalists, they often do not have prior training and experience in agency-wide personnel management. In addition, a large established personnel research and development support system (e.g., an on-going occupational analysis program) is not available to assist them.

In 1983, the training community, in conjunction with the rating managers, started a project to correct the mismatch between the work performed by Coast Guard enlisted personnel and the content of the quals, promotion examinations, and training courses. The vehicle chosen to correct the problem was improvement of the promotion examinations (called servicewide examinations within the Coast Guard) which are administered to all eligible enlisted personnel as a part of the promotion competition.

The promotion examinations are designed to test the job knowledge required by personnel in each rating. With the wide diversity of jobs and types of duty stations within each rating, it is feasible to test only a sample of the job knowledge required in a rating. Therefore, identification of what should be tested on an examination is very important. Historically, the content of each examination has been selected by a single subject matter specialist who develops the examination based upon his experience and perception of what was critical job knowledge for the personnel in that rating. After the Coast Guard changed the promotion examinations to pass/fail examinations, it became even more important to identify the knowledges for each rating which should be tested on the promotion examinations and the knowledges which should be omitted. In this project these knowledges were identified by a panel of subject matter specialists who prioritized the quals for testing on each rating's promotion examination at each paygrade.

Prioritization of the quals was designed to enable the Coast Guard to achieve three goals:

> Strengthen the link between the content of the promotion examinations and the work performed by Coast Guard enlisted personnel.
>
> Develop an accurate description of the work performed.
>
> Develop promotion examinations and training courses reflecting the accurate description of the work performed.

Prioritizing the quals would strengthen the link between the content of the promotion examinations and the work performed. Since meaningful quals were required to get meaningful prioritizations, we incorporated review and revision of the quals as a part of the process. This review and revision of the quals would enable the rating managers to begin developing an accurate description of the work performed (i.e., to revise the quals). Existence of the revised quals would then enable the training community to develop courses and examinations reflecting the actual work performed.

PROCEDURE

During 1983 and 1984, panels of subject matter specialists in 25 ratings met for one week each to review, revise, and prioritize their quals. At the start of the planning for each panel, the rating manager provided a representative for the panel meeting and selected three subject matter specialists from the field. In addition to the rating manager's representative and the three subject matter specialists from the field, the following personnel were on each panel: one subject matter specialist from the resident school, the Institute's subject matter specialist (who developed the promotion examinations and nonresident courses), and a facilitator. Within the panel structure, the program manager's representative chaired the meeting and provided policy guidance, the subject matter specialists provided the rating-specific subject matter expertise, and the facilitator provided testing/training expertise.

Since there is wide diversity within each rating, the subject matter specialists were carefully selected to ensure broad coverage of each rating. Also, efforts were made to include both junior and senior personnel so input from personnel actually doing the work today would be available. The products produced by the panels therefore would reflect both the broad range of experience possessed by senior personnel and the current working knowledge of the first-line supervisors.

Each panel evaluated each qual for currency and adequacy as written, recommended revisions to individual quals, drafted or recommended development of additional quals, and evaluated the adequacy of sections (or content areas) of the quals. Then the panel prioritized each qual for testing on each paygrade's promotion examination. In this prioritization, the panel assigned one of the following evaluations to each qual at each applicable paygrade: (1) essential to the rating, (2) necessary to the rating, or (3) enhancing to the rating. Then the panel recommended testing emphasis for each examination, which equates to number of questions, for each section of the quals. Throughout the process, the panel members were encouraged to identify problems and recommend solutions to the program manager and training community.

Two reports of each panel's evaluations and recommendations were prepared -- one for the Institute and one for the rating manager. The report to the Institute was designed to meet the project's first goal:

> strengthen the link between the content of the promotion examinations and the work performed by Coast Guard enlisted personnel.

The report for the Institute contained the prioritization of the individual quals for testing on each paygrade's promotion examination. Since this report contained testing guidelines which could be misinterpreted by personnel in the field, it was not given wide distribution.

The report to the rating manager was designed to help the rating manager meet the second goal:

> Develop an accurate description of the work performed.

The rating manager's report contained the problems identified by the panels, the proposed solutions, the overall evaluation of each section of the quals, and the recommendations for testing emphasis on each paygrade's promotion examination. The report also contained the panel's evaluation of the currency and adequacy of individual quals, recommended action for each qual (i.e., retain as written, revise, or delete), and recommendations for additional quals.

IMPACT OF PROJECT

The project's first goal has been achieved and the Coast Guard is working toward achieving the others. The panels prioritized the quals and thus strengthened the link between the content of the promotion examinations and the work performed. The rating managers are reviewing the panels' recommendations and many of them have already revised the quals thus providing the Coast Guard with an accurate description of the work performed. As revised quals become available, the training community is revising the training courses and promotion examinations to reflect this accurate description of the work performed by Coast Guard enlisted personnel.

There have been other tangible and intangible benefits as a result of this project. For example, as expected the subject matter specialists who develop the promotion examinations agreed the prioritizations and testing emphasis reflected the work performed. However, a surprise benefit was unsolicited comments from the field that "the tests were testing what they should be testing." These comments indicated the strengthened link between the content of the examinations and the work performed had been established to an extent where it could be noticed by personnel who were not aware of our efforts to improve the promotion examinations.

Another result was several rating managers using the panels as an opportunity to get additional information. For example:

Some rating managers used the panels to refine proposed quals.

One rating manager adapted the panel meeting to include extra subject matter specialists and an extra task -- generation of a task list for an occupational analysis.

Since the rating managers were involved in planning and conducting the panels, most of them were ready to implement the panels' recommendations upon receipt of the formal report. For many ratings, the panels were able to provide the information needed for the rating manager to revise the quals so additional work was not required. In other cases, additional work was required and the rating managers have been coordinating and funding the work. Examples of this work include:

Occupational analyses using task lists generated by subject matter specialists.

Follow-up panel meeting to refine the quals revised by this project's panel and assist the rating manager perform long-range planning for the rating.

Top-down occupational analysis of all functions performed at Captain of the Port and Marine Safety Offices in enforcing laws, regulations, and Coast Guard policy.

The rating managers are reviewing the problems identified by the panels and the panels' recommendations. Some of the problems identified have already been corrected. For example, in one rating the personnel did not have access to all of the manuals needed to prepare for the promotion examinations. Within a few months, the rating manager had compiled, printed, and distributed an informal publication with extracts from the relevant manuals.

Other tangible and intangible benefits are continually appearing. As mentioned previously, the field is beginning to perceive the strengthened link between the content of the promotion examinations and the work performed. The impact of other intangible benefits (e.g., increased coordination between the resident and nonresident school and greater understanding of the testing/training system) are expected to become apparent over time.

RECOMMENDATIONS FOR SIMILAR PROJECTS

As a result of this project, we have several observations and recommendations which can be useful in planning and conducting this type of project within the Coast Guard or in other organizations.

Panel procedures and materials. Keep your instructions, procedures, definitions, etc. simple. Get the information to the panel members as soon as possible to allow sufficient time for them to review the panel materials. (Our not doing this was the major complaint voiced by the panels.) Word all written materials in terms which are familiar to the reader. This may require different documents for different groups of people (e.g., management or subject matter specialists), but it will save time at the panel meeting and improve the caliber of the products produced. Consider whether or not it is necessary to specify everytning in the instructions. We allowed some degree of flexibility for our panels and they had little trouble using the procedures and refining the procedures and definitions to meet the needs of their individual ratings.

Subject matter specialists' perceptions. The subject matter specialists stated the project was exceptionally worthwhile and should be repeated. Their primary concern was that management would not listen to them -- that their recommendations would "get lost" in the system. This concern was handled by assuring them a formal report would be sent to the rating manager. After completion of the panel meetings, an area of frustration was the slowness in implementing the recommended revisions to the quals. If you anticipate a similar problem, minimize the frustration at the beginning by explaining how long it should take to implement the recommendations and why (e.g., the quals manual is revised infrequently and actual printing takes months). In addition, be aware of problems or projects which may interact with the panel's work. As might be expected, when we incorporated previous work or consideration of current problems in the rating into the agenda for the panel meeting, the panel members were more motivated than when previous work (e.g., proposed changes to the quals) or current problems were not included.

Limitations of the process. We did find limits to the process in terms of the type of quals the subject matter specialists could develop. Sometimes the subject matter specialists had difficulty generalizing across different situations. One of the most common responses was: "It varies from unit to unit, district to district, etc." This problem was particularly pronounced when the panels tried to develop quals at the E-7 through E-9 levels. At these levels Coast Guard enlisted personnel are moving from positions with technical emphasis to positions where leadership and management skills are emphasized. The problem we found may be limited to organizations similar to the Coast Guard where duties for a group of people are unit- or even billet-specific.

Facilitator. Most of our panel members were selected because of their position (e.g., Institute's subject matter specialist) or rating-specific experience. The exception to this was the facilitator who could be selected based upon other factors. For this project, the facilitator needed interviewing skills similar to the skills required of a job analyst to: apply interviewing skills in a group setting; obtain, summarize, and record the consensus of the panel on revisions, recommendations, and evaluations; and ensure participation from all panel members. It was also useful for the facilitator to be familiar with the personnel system, the project's purpose, and the relationship between the products the panel produced and the personnel system. In those instances where we knew the facilitators might not have this type of knowledge, we were able to provide the necessary information in a short orientation session.

Followup. We are in the last part of the third year on this project. During the time required for this type of project, you can expect changes in priorities, funding, and personnel. These changes do not have to stop the project or stop followup activities. The Coast Guard has handled these changes by increased coordination between the rating manager and the training community and among the members of the training community. This has been very important since the rate of change in the quals (and thus changes in our examinations and courses) is much higher than it was a few years ago. To make it easier for your organization, we recommend having one project coordinator throughout the project or, if that is not feasible, maintain clear and concise documentation to ease the transition across personnel.

DETERMINING THE BEST SET OF TRAINING FACTORS FOR USE WITH ARMY OCCUPATIONAL SURVEYS

LAWRENCE A. GOLDMAN, Ph.D.   USA Soldier Support Center - NCR

Background. The Army Occupational Survey Program (AOSP) has collected, on a routine basis, Training Factor (TF) data from senior supervisors and managers within enlisted Military Occupational Specialties (MOS) since 1981. The primary purpose of collecting TF data, supplementing information obtained from MOS incumbents, is to assist training course developers in deciding which tasks should be trained - either by the school which has responsibility for providing formal training or by supervised on-the-job training. Based upon a formal written request to the AOSP by the proponent training school for a given MOS, one or more of the TF included within the Instructional Systems Development (ISD) 8 Factor Model may be used for a specific MOS survey.

These factors are as follows:

(1) Percent of members performing
(2) Average percent of time spent by members performing
(3) Task Learning Difficulty (TLD)
(4) Consequences of Inadequate Performance (COIP)
(5) Task Delay Tolerance (TDT)
(6) Probability of Deficient Performance (PDP)
(7) Immediacy of Performance (IM)
(8) Relative Frequency (RF)

In addition, a ninth factor, Training Emphasis (TE), although not part of the ISD 8 Factor model, has been generally used in all AOSP TF surveys.

Despite the wide-spread use of these nine TF, there have been few, if any, previous U.S. Army studies indicating the extent to which these TF are related and the extent to which each of these factors could aid in "critical" task selection. Therefore, this study addressed the following areas:

(1) The degree of commonality between these nine factors; and
(2) Identification of those factors which could best isolate "critical" tasks from non-"critical" tasks.

"Critical" tasks are those found in the Soldier's Manual for each MOS and may be defined as those collective or individual tasks which are required for mission performance in combat or for survival on the battlefield. While there are "critical" tasks for each skill level within a MOS (a skill level corresponds to an authorized paygrade or cluster of authorized paygrades), the focus of this effort was on "critical" tasks required to be performed successfully by entry-level personnel since this is generally the level at which most formal training is provided.

Methodology. The results reported in this phase of the study were based on an Army-wide sample survey of job incumbents and supervisors/managers in the following enlisted MOS: 11B (Infantryman); 63H (Tracked Vehicle Repairer); 91C (Practical Nurse); 76Y (Unit Supply Specialist); 95B (Military Police); and 12B (Combat Engineer). These six MOS were selected for this study primarily because they represent widely different types of work performed by U.S. Army enlisted personnel. Data on percentages of members performing each task and average percent time spent by members performing each task (comprising the first two TF) were collected from job incumbents holding these MOS. The average percent of time spent by all members is based on a 7-point Relative Time Spent scale ranging from "1" (Very much below average) to "7" (Very much above average). Using this scale, each job incumbent rates the total amount of time spent performing each task in his/her present job relative to the time he/she spends performing every other task.

The second three TF, for which ratings were provided by supervisory/managerial senior Non-commissioned Officers (NCO's) in each of these MOS, may be defined as follows:

(1) TE - The amount of emphasis that should be given to each task requiring some type of systematic training, e.g., formal training school, supervised on-the-job training. The scale used ranges from "1" (Very low emphasis) to "7" (Very high emphasis). No response to a task by raters (a value of "0") was presumed to indicate "do not train" and was included in the computation of mean values per task.

(2) LD - Learning difficulty reflects the amount of time required to learn to perform the task satisfactorily. The more time required, the higher the learning difficulty. The scale used ranges from "1" (Extremely low learning difficulty) to "7" (Extremely high learning difficulty).

(3) CIP - This factor relates to the need for identifying tasks essential to job performance, when needed, even if they are seldom performed. The consequence of inadequate performance of certain tasks could result in injury to personnel, loss of life, or damage to equipment. The scale used ranges from "1" (Extremely low consequence - negligible effect on people/equipment/mission) to "7" (Extremely high consequence - may result in injury/death, serious damage to equipment, or failure to accomplish critical mission).

(4) TDT - Task Delay Tolerance relates to how much delay can be tolerated between the time the need for task performance becomes evident and the time that actual performance begins. The scale used ranges from "1" (Extremely low - there is virtually no requirement that an individual be able to perform the task immediately) to "7" (Extremely high - an individual must be able to perform the activity immediately whenever it is encountered).

(5) PDP - This factor relates to insuring that training is given in those essential job skills in which job incumbents frequently perform poorly. It is assumed that for any job, some tasks are more difficult to accomplish than others. The highest ratings on this factor would reflect the most poorly performed tasks. The scale used ranges from "1" (Never performed) to "7" (Very frequently deficiently performed).

(8) IM - This factor is associated with the interval of time between completion of training and the first performance of the task on the job. The scale utilized ranges from "1" (Never performed) to "7" (Initially performed less than months after Advanced Individual Training).

(9) FF - This factor relates to the frequency with which tasks are performed by job incumbents. It ranges from "1" (Very seldom) to "7" (Very frequently).

The sample sizes for the entry skill level personnel (job incumbents) in each of these six MOS follow: (1) MOS 11B - 1424 (skill level 1 corresponding authorized paygrades E-3 and E-4); (2) MOS 63H - 332 (skill level 1); (3) MOS 91C - 161 (skill level 2 corresponding to authorized paygrade E-5); (4) MOS 76Y - 45 (skill level 1); (5) MOS 95B - 1042 (skill level 1); and (6) MOS 12B - 879 (skill level 1). The number of raters by MOS for each of the seven TF ranged as follows: (1) MOS 11B - 56 for COIP to 104 for IM; (2) MOS 63H - 28 for COIP to 53 for TF; and (3) MOS 91C - 31 for TDT to 43 for PDF; (4) MOS 76Y - 29 for COIP to 8[illegible] for RF; (5) MOS 95B - 66 for TDT to 95 for IM; and (6) MOS 12B - 19 for COIP to 30 for TF and IM.

The Comprehensive Occupational Data Analysis Programs (CODAP) were used to obtain data files for the six MOS consisting of computed mean values of the nine TF for each task in the MOS questionnaires. The Statistical Package for the Social Sciences (SPSS) was then used for each MOS. To examine the degree of inter- correlation among the nine TF, a Pearson correlation coefficient matrix was generated followed by factor analysis with varimax rotation of the principal factors. To determine which factors could best descriminate "critical" tasks from non-"critical" tasks, step-wise discriminant function analysis was utilized. For this latter analysis, only the seven TF based on ratings provided by senior supervisory/managerial personnel were examined. This was due to the fact that the senior raters, who could have responded to each task in the questionnaire for each of the seven rating TF, theoretically would have attached greater importance to each "critical" task impacting on successful mission performance. On the other hand, job incumbents would have responded, by direction, only to those tasks which they had performed or had been trained to perform in their current duty position. Also, regardless of the outcome of this study, the two TF obtained from job incumbents would continue to be collected for each MOS to be surveyed. Conversely, focusing on the seven TF rated by senior personnel would facilitate reduction of the number of factors which training development personnel need to examine for "critical" task selection.

Findings.

A. TF Reliability. To determine internal consistency of the data, two types of reliability coefficients were computed for each of the seven TF for these six MOS obtained from senior raters: (1) the average inter-rater reliability of a single rater (r11); and (2) the stepped up reliability coefficient reflecting the overall group of raters for a particular TF (rkk). In general, the r11 values were moderate while the rkk values were consistently high across all three MOS. With respect to MOS 11B, the r11 values ranged from .20 for COIP and TDT to .39 for IM; the rkk values ranged from .91 for COIP to .98 for IM.

With respect to MOS 63B, the r11 values varied from .[illegible] for COIP to .[illegible] for [illegible]I; the rkk values ranged from .[illegible] for COIP to .91 for [illegible]I. The r11 values for MOS 91[illegible] ranged from .15 for [illegible]I to .36 for [illegible]D; the rkk values ranged from .8[illegible] for [illegible] to .9[illegible] for IM. The r11 values for MOS 76Y varied from .15 for I[illegible] to .31 for COIP; the rkk values ranged from .8[illegible] for IF to .93 for PDP. The r11 values for MOS [illegible] varied from .17 for ID to .39 for IM; the rkk values varied from .86 for ID to .96 for IM. Lastly, the r11 values for MOS 12B varied from .16 for IF and COIP to .35 for [illegible]I; the rkk values ranged from .7[illegible] for COIP to .9[illegible] for IM.

B. Inter-correlations. The Pearson correlation coefficients among the nine factors were all statistically significant for these four MOS: 12B, 63B, [illegible] and [illegible]. With respect to MOS 91C, significant correlations were observed for all factors except for ID with IDT and for COIP with the following factors: IDT; RF; [illegible] (Skill Level 2); and Average Percent Time Spent by all Members (Skill Level [illegible]). With respect to MOS 76Y, significant correlations were also noted for all factors with the exception of COIP with RF and Average percent time spent (Skill Level 1). Table 1 shows the inter-correlation matrices for MOS 12B, 63B and 91C. Similarly, Table 2 shows the matrices for MOS 76Y, 9[illegible] and 76X.

TABLE I - INTER-CORRELATION MATRICES FOR MOS 11B, 63H AND 91C

MOS 11B

	TE	LD	COIP	TDT	PDF	IM	RF	%Do	AVG % TIME
TE	1.0	-.45	.81	.78	.86	.91	.77	.86	.76
LD		1.0	-.25	-.35	-.43	-.58	-.39	-.57	-.55
COIP			1.0	.90	.36	.79	.81	.64	.61
TDT				1.0	.93	.84	.92	.65	.58
PDF					1.0	.92	.94	.76	.76
IM						1.0	.84	.91	.87
RF							1.0	.67	.71
%Do								1.0	.94
AVG % TIME									1.0

MOS 63H

	TE	LD	COIP	TDT	PDF	IM	RF	%Do	AVG % TIME
TE	1.0	.68	.77	.84	.90	.76	.87	.62	.59
LD		1.0	.73	.84	.67	.36	.82	.13	.22
COIP			1.0	.75	.80	.57	.81	.36	.38
TDT				1.0	.82	.56	.90	.37	.42
PDF					1.0	.75	.87	.56	.56
IM						1.0	.66	.79	.75
RF							1.0	.44	.49
%Do								1.0	.94
AVG % TIME									1.0

MOS 91C

	TE	LD	COIP	TDT	PDF	IM	RF	%Do	AVG % TIME
TE	1.0	-.27	.27	.82	.73	.78	.76	.83	.76
LD		1.0	.55	.04	-.40	-.61	-.53	-.47	-.42
COIP			1.0	.50	-.04	-.14	-.08	-.04	-.02
TDT				1.0	.68	.61	.60	.68	.65
PDF					1.0	.81	.83	.82	.77
IM						1.0	.81	.90	.83
RF							1.0	.87	.86
%Do								1.0	.96
AVG % TIME									1.0

TABLE 2 - INTER-CORRELATION MATRICES FOR MOS 76Y, 95B AND 12B

MOS 76Y

	TF	LD	COIP	TDT	PDF	IM	RF	%Do	AVG % TIME
TF	1.0	-.50	.25	.71	.85	.75	.77	.82	.73
LD		1.0	-.20	-.76	-.59	-.78	-.48	-.64	-.63
COIP			1.0	.46	.29	.36	.06	.18	.09
TDT				1.0	.79	.67	.61	.72	.68
PDF					1.0	.84	.76	.82	.76
IM						1.0	.57	.85	.79
RF							1.0	.69	.70
%Do								1.0	.96
AVG % TIME									1.0

MOS 95B

	TF	LD	COIP	TDT	PDF	IM	RF	%Do	AVG % TIME
TF	1.0	-.57	.77	.87	.91	.87	.85	.85	.70
LD		1.0	-.43	-.59	-.61	-.71	-.71	-.57	-.58
COIP			1.0	.89	.68	.60	.66	.54	.42
TDT				1.0	.86	.78	.84	.67	.56
PDF					1.0	.90	.92	.79	.69
IM						1.0	.90	.89	.84
RF							1.0	.79	.76
%Do								1.0	.93
AVG % TIME									1.0

MOS 12B

	TF	LD	COIP	TDT	PDF	IM	RF	%Do	AVG % TIME
TF	1.0	-.52	.71	.83	.78	.74	.64	.71	.64
LD		1.0	-.50	-.58	-.65	-.67	-.63	-.53	-.53
COIP			1.0	.64	.68	.61	.58	.62	.56
TDT				1.0	.80	.77	.72	.56	.52
PDF					1.0	.82	.88	.63	.62
IM						1.0	.77	.77	.73
RF							1.0	.55	.58
%Do								1.0	.96
AVG % TIME									1.0

Based on the generally high inter-correlations of all nine factors for each of these six MOS, it was not altogether surprising that through factor analysis (with varimax rotation of the factor matrices) only two factors emerged for each of the six MOS. As shown in Table 3, the principal factor for MOS 11B (accounting for 90 percent of the total variance) reflected significantly high factor loadings for all factors with the exception of LD. Similarly, the principal factor for MOS 63H (representing 80 percent of the total variance) reflected significantly high loadings for all of the seven TF - provided by senior raters; of some interest was the fact that the two TF produced by job incumbents (% Do and Average percent time spent) had insignificant loadings. The primary factor for MOS 91C (accounting for 78 percent of the total variance) had factor loadings above 0.8 for all factors except LD and COIP. Similar results were also obtained for MOS 76Y, 95B and 12B. As noted in Table 4, the principal factors for MOS 76Y and MOS 95B (each accounting for 90 percent of the variance) reflected significantly high loadings for the nine TF with the exception of COIP for MOS 76Y. The principal factor for MOS 12B (accounting for 89 percent of the variance) also revealed substantially high loadings for each of the nine TF. What these findings indicate is that there is apparently just one underlying general TF, rather than nine different TF.

TABLE 3 - VARIMAX ROTATED FACTOR LOADINGS OF NINE TRAINING FACTORS ON PRINCIPAL FACTORS - MOS 11B, 63H, AND 91C

INDIVIDUAL FACTORS	FACTOR LOADINGS ON PRINCIPAL FACTOR MOS 11B	 MOS 63H	 MOS 91C
TF	.67	.79	.89
LD	-.16	.89	-.43
COIP	.87	.81	.07
TDT	.92	.89	.80
PDF	.85	.80	.87
IM	.66	.45	.91
RF	.84	.91	.90
% Do	.41	.13	.96
AVG % Time	.44	.20	.91
PERCENT OF TOTAL VARIANCE ACCOUNTED FOR	90	80	78

TABLE 4 - VARIMAX ROTATED FACTOR LOADINGS OF NINE TRAINING FACTORS ON PRINCIPAL FACTORS - MOS 76Y, 95B, AND 12B

INDIVIDUAL FACTORS	FACTOR LOADINGS ON PRINCIPAL FACTOR MOS 76Y	 MOS 95B	 MOS 12B
TF	.81	.62	.72
LD	-.54	-.55	-.60
COIP	.03	.24	.61
TDT	.58	.38	.84
PDF	.81	.65	.89
IM	.71	.83	.73
RF	.79	.70	.81
% Do	.91	.87	.37
AVG % Time	.91	.91	.36
PERCENT OF TOTAL VARIANCE ACCOUNTED FOR	90	90	89

C. Prediction of "Critical" Tasks. In predicting "critical" tasks from the total task inventory, only the seven TF derived from senior raters for each of these MOS were used as independent (predictor) variables for the reasons noted previously. The objective of the use of stepwise discriminant analysis was to determine the fewest number of TF which best predicted "correct" group membership. That is, it was desired to isolate those predictors which significantly increased the percent of tasks classfied correctly. The first factor entered was that which had the largest value representing the greatest power of discrimination as measured by Rao's V. Subsequent variables selected were those which, when added to the previously selected predictors, measurably increased the percentage of tasks classified correctly.

Table 5 displays those factors which best achieved this objective for MOS 11B, 63H and 91C. Table 6 shows those factors for MOS 76Y, 95B and 12B. For each factor selected in these six MOS, the concomitant change in Rao's V is displayed together with the cumulative percentage of tasks classified correctly.

TABLE 5 - PREDICTION OF "CRITICAL" TASKS -
MOS 11B, 63H, AND 91C

STEP ENTERED	INDIVIDUAL FACTOR SELECTED			CHANGE IN RAO'S V			PERCENT OF TASKS CORRECTLY CLASSIFIED		
	11B	63H	91C	11B	63H	91C	11B	63H	91C
1	TE	LD	TE	38.0	195.1	123.5	63.7	76.6	69.0
2	COIP	TE	LD	37.9	61.8	35.9	68.5	80.1	74.1
3	IM	COIP	*	5.9	26.2	*	70.8	82.2	*
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
TOTAL	-	-	-	-	-	-	72.4	83.9	75.2

* The variable for MOS 91C which was entered at step 3 (RF) caused a statistically insignificant change in Rao's V.

TABLE 6 - PREDICTION OF "CRITICAL" TASKS -
MOS 76Y, 95B, AND 12B

STEP ENTRED	INDIVIDUAL FACTOR SELECTED			CHANGE IN RAO'S V			PERCENT OF TASKS CORRECTLY CLASSIFIED		
	76Y	95B	12B	76Y	95B	12B	76Y	95B	12B
1	TE	TE	TE	64.9	221.2	48.7	74.0	77.4	66.6
2	COIP	*	TDT	18.5	*	13.2	76.6	*	72.7
3	IM	**	COIP	19.2	**	8.5	80.0	**	73.6
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
TOTAL	-	-	-	-	-	-	80.0	75.6	75.9

* The variable for MOS 95B which was entered at step 2 (PDP) is not shown because when it was added to TF which was previously selected the percent of tasks correctly classified decreased. ** The variable for MOS 95B which was entered at step 3 (RF) caused a statistically insignificant change in Rao's V.

As indicated in Tables 5 and 6, TE was the predominant factor, being entered at either step 1 for five of the six MOS and entered at step 2 for the other MOS (63H). The second best predictor of correctly classfied tasks was COIP, appearing among the top three best predictors in four of the six MOS. Two other predictors, ID and IM, were noted in two of these MOS.

It was observed that the percentage of tasks classfied correctly for MOS 11B, 63H and 95B was higher for "critical" tasks than for non-"critical" tasks. For MOS 11B, these percentages were 80.8 percent vs. 71.4 percent. For MOS 63H, these percentages were 92.0 percent vs. 79.9 percent. Similarly, for MOS 95B, they were 80.2 percent vs. 75.8 percent. On the other hand, the percentage of tasks classified correctly for MOS 91C and MOS 12B was higher for non-"critical" tasks than for "critical" tasks - 79.1 percent vs. 62.4 percent for MOS 91C and 76.2 percent vs. 71.9 percent for MOS 12B, respectively. For MOS 76Y, these percentages were the same - 80 percent.

Conclusions and Implications.

It was evident that there is a remarkably high degree of correlation between the seven TF rated by senior supervisory/managerial personnel and the two TF relating to job incumbent information (reflecting the percent of members performing at the entry skill level and the average percent time spent by these members). What these findings suggest is that rather than nine separate factors there exists, in reality, only one clearly defined factor. Similarly, in terms of predicting "critical" vs. non'"critical" tasks, rather than nine individual factors there is essentially only TF (and probably COIP) which could be thought of as generally consistent significant predictors. These finding could provide a basis for training developers to decide which training factor(s) is (are) most beneficial for "critical" task selection and thus facilitate their efforts to identify accurately, entry level training requirements. In turn, the AOSP could collect the most useful amount of TF data, primarily to aid these training developers, from a minimum number of factors. This would enable the AOSP to significantly improve its administration of occupational surveys.

EXAMINATION OF ENVIRONMENTAL DETERMINANTS OF ARMY PERFORMANCE CRITERIA

Darlene M. Olson
U.S. Army Research Institute[1]

Walter C. Borman
Personnel Decisions Research Institute

Job performance has been conceptualized as a product of individual attributes, abilities, and skills which are measurable at the time an individual first enters the organization, of environmental/organizational variables which impact on the individual after job-entry and of the person's motivation to perform. Previous empirical research has investigated work performance in terms of taxonomies of human abilities, values, and personality characteristics (Dunnette, 1976). However, until recently little research has focused on developing taxonomies of environmental/organizational variables or examining relationships between these factors and work-related outcomes.

The major purpose of this research was to examine relationships among individual, organizational/environmental factors, job characteristic variables, and measures of both maximal (e.g., hands-on and job knowledge tests) and typical (e.g., supervisory and peer ratings of performance) performance criteria for first-term soldiers in the Army. This paper discusses results from administering a 110-item Army Work Environment Questionnaire (AWEQ) to 800 first-term enlisted personnel from five military occupational specialties (MOS).

A major impetus for research on environmental variables was the work of Schneider (1978), who proposed that such situational influences as job/task characteristics, organizational practices (e.g., reward system) and climate variables could either directly influence performance or moderate the relationship between cognitive abilities and performance. During the early 1980's several research projects were initiated to develop empirically validated taxonomies of environmental variables (e.g., Peters & O'Connor, 1980; Olson, Borman, Robertson, & Rose, 1984). Results from the development of situational/environmental taxonomies have suggested that situational variables can be identified, categorized and reliably measured. In a series of laboratory studies conducted by Peters and O'Connor, and their colleagues (for a review see Eulberg, O'Connor, Peters & Watson, 1984), results have demonstrated that situational constraints are significantly related to ineffective task performance, job dissatisfaction, and increased frustration.

Although correlational field studies have supported the relationships between environmental/situational variables and affective reactions to the job (e.g., satisfaction), associations between these factors and ratings of performance effectiveness have been inconsistent. For example,

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the view of the U.S. Army Research Institute or the Department of the Army.

O'Connor et al. (1984), reported a weak, but significant relationship between an overall measure of situational constraints and both performance (r = [illegible]) and turnover (r = [illegible]) criteria in a national sample of convenience store managers. In another study (O'Connor, Peters, Eulberg & Watson, 1984), no significant correlations were observed between situational constraint factors and either performance or reenlistment intentions of Air Force personnel. In contrast, findings from other research (Olson, et al., 1984; Olson & Borman, in press) which used different constructs/measures have demonstrated significant relationships (rs ranged from [illegible] to [illegible] across criteria) between environmental factors and performance criteria.

In general, the mixed results found for relationships between environmental factors and performance suggest that the magnitude of the correlations obtained are dependent on the level of inhibitors/facilitators actually present in the work environment. Further, the ways situational variables are conceptualized, the kinds of jobs investigated, and the types of performance criteria examined may impact on the observed relationships.

METHOD

Subjects. The research sample contained [illegible] first-term enlisted personnel from five Army MOS. There were [illegible] infantrymen (11B MOS), 169 armor crewmen (19E MOS), 144 radio teletype operators (31C MOS), 155 light wheel vehicle mechanics (63B MOS), and [illegible] medical care specialists (91A MOS). These soldiers were sampled at four continental United States and two European Army installations.

Measures. An assessment battery containing an environmental questionnaire and a comprehensive set of typical (e.g., supervisory ratings) and maximal (e.g., job knowledge test) performance measures was used in this research.

Army Work Environment Questionnaire (AWEQ). The Army Work Environment Questionnaire is a 110-item multiple choice instrument that measures 14 dimensions of the Army work environment. The AWEQ was constructed in a two-stage process (Olson, et al. 1984). Briefly, in Stage I, a taxonomy of important environmental influences on soldier performance was derived through application of a critical incident methodology. A total of 282 critical incidents, generated by Army experts (N = 67) and independently content-analyzed by six psychologists, identified environmental/organizational influences beyond the control of the soldier that had a significant impact on performance, either inhibiting or facilitating that performance. The Army work environment taxonomy contains the following nine "job-oriented" factors: (1) Resources tools/equipment, (2) Workload/Time Availability, (3) Training, (4) Physical Working Conditions, (5) Job-Relevant Information, (6) Job Relevant Authority, (7) Perceived Job Importance, (8) Work Assignment, and (9) Changes in Job Procedures/ Equipment, as well as, the remaining five "climate-oriented" dimensions: (10) Reward System, (11) Discipline, (12) Individual Support, (13) Job Support/Guidance and (14) Role Models. In Stage II, items were written to cover the content of the 14 environmental dimensions.

Items on the AWEQ are descriptive in nature and respondents are asked to indicate on a 5-point rating scale (e.g., 1 = Very Seldom or Never to 5 = Very Often or Always) how often each environmental situation described by the items occurs on their present job.

Job Performance Measures. The set of typical and maximal performance criteria used in this study was developed as a component of a broader research program conducted under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This comprehensive nine year research effort was initiated to help the Army access, assign, and retain quality personnel.

The typical performance criteria included supervisory and peer job performance ratings. Separate behaviorally-anchored rating scales (BARS), derived from a critical incident job analysis procedure, were used to measure both the MOS (job)-specific and Army-wide components of soldier performance and effectiveness on a 7-point behavior rating format. For each research participant in the five MOS, an Army-wide and MOS-specific rating was computed by averaging the performance ratings across all individual dimensions for supervisors and peers separately.

The maximal performance criteria included hands-on (work sample) tests and job knowledge measures. The hands-on tests for each MOS consisted of 15 tasks identified for the MOS. The individual performance components of each task were scored by trained raters on a pass-fail basis and an overall hands-on score was computed for each soldier by averaging the proportions passed across the tasks tested. Multiple-choice tests were developed to assess job knowledge relevant to each important task for an MOS. An overall job knowledge test score for each research participant was derived as a percentage of the number of items answered correctly

Procedures. After the supervisor and peer raters were trained to use the Army-wide and MOS-specific BARS, they evaluated the job performance of soldiers in the research sample. Concurrently with these assessments, first-tour soldiers participating in the research were administered: (a) the Army Work Environment Questionnaire and (b) the appropriate job knowledge and hands-on test. For all respondents, scores on the environmental measure were merged with scores from the maximal and typical performance criteria for analyses.

RESULTS AND DISCUSSION

For the total sample, Table 1 presents the means, standard deviations, and reliability coefficients for the research measures. When mean ratings on the AWEQ scale dimensions are collapsed across MOS and installation, results suggest that a complex set of both facilitating and inhibiting influences describe the Army work environment. For instance, the mean ratings for such AWEQ scales as Training (*M* = -3.02), Work Assignment (*M* = -1.90), Reward System (*M* = -1.75), and Job Support (*M* = -1.42) were described somewhat negatively. In contrast, such environmental variables as Perceived Job Importance (*M* = 1.76), Discipline Practices (*M* = 1.10), Individual Support (*M* = .79), and adequacy of Role Models (*M* = .74) were generally described more positively. Uncorrected reliability estimates displayed in Table 1 show that the job knowledge tests tend to be the most reliable of the maximal performance criteria and the Army-wide BARS (supervisors) have the largest coefficients of the typical performance measures. Generally, the AWEQ scale scores, with coefficients ranging from .[illegible] to .78, have adequate reliabilities for a research instrument.

Table 2 presents the intercorrelation matrix for the AWEQ scales. Intercorrelations among the 14 AWEQ scales show that the climate-oriented dimensions are more highly related than the job-oriented factors. For

Table 1

Means, Standard Deviations, and Reliability Coefficients for

Selected Measures Across MOS.

Measures	N	X̄	SD	r[1]
Army-Wide BARS (Peers)	727	4.52	.71	.78-.86
Army-Wide BARS (Supervisors)	722	4.50	.84	.81-.86
MOS-Specific BARS (Peers)	727	4.60	.56	.76-.86
MOS-Specific BARS (Supervisors)	718	4.62	.77	.78-.87
Hands-on Test	685	71.72	16.11	.35-.56
Job Knowledge Test	745	62.47	10.63	.84-.91
AVEQ Scales (N of items):[2]				
Resources (n=[illegible]	734	-.99	4.96	.75
Workload (n=8	752	-.67	4.34	.58
Training (n=11)	736	-3.02	5.91	.64
Physical Working Conditions (n=6)	741	.67	3.85	.57
Job Authority (n=6)	760	-.25	3.65	.57
Job Information (n=8)	726	.45	4.60	.67
Job Importance (n=7)	725	1.76	4.65	.67
Work Assignment (n=9)	731	-1.90	6.80	.70
Changes in Job Procedures (n=8)	745	-.89	4.21	.58
Reward System (n=7)	736	-1.75	5.14	.78
Discipline (n=6)	751	1.10	4.07	.65
Individual/Support (n=9)	727	.79	5.46	.73
Job Support (n=8)	734	-1.42	5.12	.72
Role Models (n=10)	731	.74	5.98	.71

Note. 1). For performance ratings, the range of interrater reliabilities across MOS are reported.

For Hands-on and Job Knowledge tests, the range of split-half reliabilities across MOS are reported.

For the Environmental scales, Cronbach's alpha coefficients are used as measures of internal consistency.

2). Mean scale scores were computed such that "0" is a neutral environment. Positive mean values indicate positive descriptions of the environment for that scale. Negative scale means indicate the opposite.

Table 2

Scale Intercorrelations for the AVEQ.

AVEQ Scales*	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Resources	-													
2. Workload	.52	-												
3. Training	.29	.26	-											
4. Working Conditions	.55	.48	.23	-										
5. Job Authority	.47	.52	.39	.51	-									
6. Job Information	.52	.50	.42	.50	.60	-								
7. Job Importance	.2[illegible]	.20	.30	.22	.32	.33	-							
8. Work Assignment	.26	.24	.66	.18	.35	.36	.43	-						
9. Job Procedures	.49	.51	.44	.43	.50	.51	.24	.40	-					
10. Reward System	.38	.40	.40	.37	.58	.56	.29	.33	.45	-				
11. Discipline	.31	.31	.18	.37	.47	.48	.30	.14	.36	.45	-			
12. Individual Support	.31	.32	.35	.36	.56	.60	.34	.27	.41	.62	.54	-		
13. Job Support	.39	.40	.44	.38	.64	.62	.34	.37	.48	.73	.48	.72	-	
14. Role Models	.41	.46	.44	.42	.61	.60	.34	.35	.50	.56	.48	.58	.65	-

Note. All AVEQ scale intercorrelations are significant at $p < .05$.

*Correlations significant at $p < .05$.

1). Scales 1-9 are more job-oriented and scales 10-14 are more climate-oriented.

example, Job Support is strongly associated with the Reward System (r = .[illegible]), Individual Support (r = .72), and Role Models (r = .65). The Discipline and Authority Dimension, conceptualized as a job-related factor, had its strongest associations with the climate scales of Role Models ([illegible]) and Job-Support (r = .[illegible]). Subsequent test development work on the AWEQ, which has included an item-analysis and a principle component factor analysis with a varimax rotation, has been conducted to identify a subset of the original [illegible] items that best define the factor structure of the AWEQ. Although findings from these analyses corroborate the redundancy displayed in Table [illegible] for some of the AWEQ scales and tentatively suggest that a five factor solution with [illegible] items may permit a more parsimonious explanation of the underlying Army work environment constructs, results based on the revised-AWEQ have not been sufficiently cross-validated.

The results presented in Table [illegible] focus on the relationships between the [illegible] scores from the conceptual taxonomy, and a comprehensive set of [illegible] ratings of job performance and more objective performance indices.

Table [illegible] presents the correlation coefficients between the 14 AWEQ scale scores and the set of performance criteria for the total sample. Several interesting findings emerged. First, the largest correlations were found between environmental variables and typical performance measures, specifically the Army-wide BARS. In terms of the number of significant effects, 46.4% of the correlation coefficients between environmental variables and typical measures, as compared with 28.6% of the correlations for maximal criteria, were statistically significant. This difference cannot be attributed to sampling error, since differences in sample sizes for the correlational values shown in Table 3 were relatively minor.

Second, generally the environmental dimensions of (a) Perceived Job Importance, (b) Discipline practices, (c) Individual Support, and (d) the Reward System tended to be significantly correlated with performance criteria for the total sample. In contrast, the AWEQ scale scores on (a) Resources Tools/Equipment, (b) Workload/Time Availability, (c) Physical Working Conditions, and (d) Changes in Job Procedures/Equipment were not significantly associated with scores on the performance measures. Although the magnitude of these environment-performance relationships are lower than those previously reported with Army field test data from Project A (see Olson et al., [illegible]), fairly consistent trends have been observed in the pattern of significant relationships between climated-oriented AWEQ scales and performance ratings.

Third, when relationships between typical performance measures and environmental factors were examined, 60% of the correlations between climate-related dimensions and 58.3% of the correlations with job-oriented factors were significantly related to performance ratings. Further, a similar pattern of significant relationships was found between the environmental variables and maximal performance criteria, specifically 50% of the observed correlation coefficients for climate dimensions and 16.7% of the correlations for job dimensions were significantly associated with scores on maximal performance measures. It was predicted that job-oriented environmental factors should have more significant relationships with the objective, maximal performance measures, than the supervisory and peer ratings of overall soldier effectiveness. However, these findings did not support this contention, because a larger percentage of climate-oriented factors than job-oriented factors were significantly correlated with both types of performance indices.

Table 3

Correlations Between AWEQ Scale Scores and Performance Criteria.

Performance Criteria	Scale Scores on Army Work Environment Questionnaire													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
							Typical Performance Measures							
Army-wide BARS (Peers)	.05	.02	.07	.08*	.13*	.09*	.23*	.07	.04	.11*	.14*	.18*	.14*	.11*
Army-wide BARS (Supervisors)	.01	-.03	.08*	.06	.15*	.11*	.17*	.12*	.01	.11*	.13*	.14*	.13*	.09*
MOS-specific BARS Peers	.01	-.04	.06	.05	07	.04	.16*	.04	0	.05	.10*	.09*	.07	.05
MOS-specific BARS (Supervisors)	-.01	-.08*	.08*	.03	.06	.06	.13*	.11*	-.01	.03	.06	.05	.03	.03
							Maximal Performance Measures							
Hands-on Test	-.02	0	-.06	-.01	-.04	.02	.09*	-.02	-.02	-.08*	.08*	.04	-.07	.04
Job Knowledge Test	-.05	-.05	-.05	.03	0	.03	.13*	-.07	-.08*	-.09*	.13*	.11*	-.03	.01

Note AWEQ SCALES. 1= Resources, 2=Workload, 3=Training, 4=Physical Working Conditions, 5=Job Relevant Authority, 6=Job Relevant Information, 7=Perceived Job Importance, 8=Work Assignment, 9=Changes in Job Procedures, 10=Reward System, 11=Discipline, 12=Individual Support, 13=Job-Related Support, 14=Role Models.

*Correlations which are significant at $p < .05$.

Fourth, consistent relationships were observed between environmental variables and the typical performance measures, specifically the Army-wide BARS, regardless of whether performance was evaluated by supervisors or peers. This finding indicates the existence of some convergence across types of performance criteria with respect to the influence of environmental factors.

Finally, since sampling error may explain many of the observed differences between MOS, only a few potentially meaningful environment-performance relationships will be discussed for the various Army jobs. In the Infantry (11B MOS) and the Medical Specialist (91A MOS) jobs, soldiers' performance on nearly all the typical measures was significantly correlated with Perceived Job Importance (significant rs ranged from .17 to .30). Significant relationships were observed between Work Assignment and the performance of Armor Crewmen (19E MOS) on the MOS-specific BARS and the Hands-on measure. In the 31C MOS, the performance of radio-teletype operators was significantly correlated with Job Relevant Authority (r = .31 observed with Army-wide BARS for supervisors). Performance of mechanics (63B MOS) on the Army-wide BARS (peers) was significantly related (r = .30) to scores on the Individual Support dimension.

Although these specific MOS findings suggest potentially interesting relationships between environmental variables and performance, these correlations are based on substantially smaller sample sizes than those reported in Table 3 and cannot, without cross-validation, be assumed to represent stable estimates of the true correlational values.

CONCLUSIONS

This research examined correlations between 14 scale scores on an Army Work Environment Questionnaire and measures of both typical and maximal performance. Prior to this applied research in an Army setting, inconsistent findings were reported in the empirical literature with respect to relationships between organizational/environmental variables and performance.

Results from this applied Army research indicated that significant relationships exist between job-oriented and climate-related environmental variables and both job performance ratings (typical measures) and more maximal, objective criteria-job knowledge and hands-on tests. Further, these findings suggest that: (1) environmental factors have their strongest correlations with more typical performance measures such as Army-wide BARS, (2) climate-oriented environmental variables have a larger number of significant effects on maximal performance criteria than job-related environmental dimensions, and (3) generally such job-oriented environmental variables as Resources/Tools/Equipment, Physical Working Conditions, and Changes in Job Procedures/Equipment are not significantly correlated with the comprehensive set of performance criteria.

Perhaps, the weak but significant correlations observed between environmental dimensions and performance may be related to: (1) a lack of sufficiently constraining or facilitating conditions on the part of the environmental variables themselves or (2) contextual factors such as raters adjusting their performance evaluations to compensate for the negative/positive effects of specific work environments.

REFERENCES

Dunnette, M. D. (Ed.) (1976). Handbook of industrial and organizational psychology. Chicago, IL.: Rand McNally.

Eulberg, J. R., O'Connor, E. J., Peters, L. H., & Watson, T. W. (1984). Performance constraints: A selective review of relevant literature. Psychological Documents.

O'Connor, E. J., Peters, L. H., Eulberg, J. R., & Watson, T. W. (1984). Situational constraints in Air Force work settings: Effects on performance, affective reactions and reenlistment plans. Paper presented at the annual meeting of the Academy of Management, Boston, MA.

O'Connor, E. J., Peters, L. H., Pooyan, A., Weekley, J., Frank, B., & Erenkrantz, B. (1984). Situational constraint effects on performance, affective reactions, and turnover: A field replication and extension. Journal of Applied Psychology, 69(4), 663-672.

Olson, D. M., Borman, W. C., Roberson, L., & Rose, S. R. (1984). Relationship between scales on an Army work environment questionnaire and measures of performance. Paper presented at the 92nd annual meeting of the American Psychological Association, Toronto, Canada.

Peters, L. H., & O'Connor, F. J. (1980). Situational constraints and work outcomes: The influence of a frequently overlooked construct. *Academy of Management Review*. *5*, 391-397.

Schneider, B. (1978). Person-situation selection: A review of some ability-situation interaction research. *Personnel Psychology*, *31*, 381-397.

Steel, R. P., & Mento, A. J. (in press). Opportunity knocks: The impact of situational constraints on relationships between job performance criteria. *Organizational Behavior and Human Decision Processes*.

Initial Standardization of an Air Force Organizational Assessment Survey Instrument*

Lawrence O. Short, Lt Colonel, USAF
Air Force Human Resources Laboratory (AFHRL)

James K. Lowe, Captain, USAF
1605th Civil Engineering Squadron

Janice M. Hightower, Captain, USAF
Detachment 5, Air Force Operational Test and Evaluation Center

In 19[illegible], the Directorate of Research and Analysis of the Leadership and Management Development Center (LMDC) began work on a revision of the primary data gathering instrument used in the LMDC consulting process, the Organizational Assessment Package (OAP). The purpose of this report is to discuss the results of initial standardization research performed on the second generation OAP, now called the Organizational Assessment Survey (OAS). More specifically, the report deals with deriving the OAS factor structure and testing each obtained factor's internal consistency reliability.

In its present form, the OAP survey consists of a computer-scored response sheet and a 109-item (93 attitudinal and 16 demographic) booklet. Responses use a scale of one to seven, with a value of "1" generally indicating strong disagreement or dissatisfaction with the question or statement, and a "7" indicating strong agreement or satisfaction. Through factor analysis, the 93 attitudinal items are combined into factors as presented in Table 1.

Table 1. OAP Factors

Skill Variety
Task Identity
Task Significance
Job Feedback
Performance Barriers and Blockages
Need for Enrichment (Job Desires)
Job Performance Goals
Pride
Task Characteristics
Work Repetition
Desired Repetitive Easy Tasks
Advancement/Recognition
Management-Supervision
Supervisory Communications Climate
Organizational Communications Climate
Work Group Effectiveness
Job Satisfaction
Job Related Training
General Organizational Climate

Administration of the survey is the first step in the consultation process. The survey is given to a stratified random sample of the organization to which LMDC has been invited. The results of the survey are an important feature in the assessment of task, supervision, climate, and productivity in an organization. The results are handled in a confidential

*All work was completed while authors were members of the Leadership and Management Development Center, Research and Analysis Directorate.

manner between LMDC and the client. After approximately five to six weeks for analysis, consultants return to the organization to provide feedback of data to commanders and supervisors.

When organizational problems are encountered, a consultant and supervisor develop a management action plan designed to resolve the problem at that level of the organization. Within six to nine months, the consulting team returns to readminister the survey instrument as a means to help assess the impact of the consulting process.

The data from each OAP administration effort are stored in a cumulative data base currently containing over 200,000 records for research purposes. These data are aggregated by work group codes developed for this instrument. The data may be recalled by demographics such as personnel category, age, sex, Air Force Specialty Code, pay grade, time in service, and educational level.

The OAP was developed jointly by LMDC and AFHRL at Brooks AFB, Texas (Hendrix & Halverson, 1979a; 1979b). More recently, additional standardization work has been done with the OAP. Short and Hamilton (1981) provided evidence of the factor-by-factor reliability of the instrument considering both internal consistency and test-retest (stability) aspects. In addition, Short and Wilkerson (1981) provided evidence in support of the group differences aspect of OAP construct validity. Webster (1982) also studied construct validity of the leadership and organizational climate areas of the OAP by using a modified multi-trait, multi-method approach, favorably comparing the OAP to the Survey of Organizations (Taylor & Bowers, 1972). Finally, the stability of the OAP factor structure was studied across selected functional area and demographic groups (Hightower & Short, 1982) and across intervals of time (Hightower & Short, 1983). These studies yielded a slightly different factor structure than that currently in use, but showed the revised structure to be extremely stable across all comparison groups. These studies combined with several years of experience using the instrument in the consulting process pointed out ways the OAP could be revised to enhance the process and the accuracy and precision of organizational diagnoses.

Method

Overview and Subjects

The subjects of the OAP-OAS data gathering came from three operational bases in the Continental United States. Since the surveys were administered by LMDC consultants, the test bases were preselected via the consultant's schedule. The data were gathered by LMDC consultants both as part of the consulting process and as an effort to test the new OAS instrument. The subjects responded to both the OAP and the OAS and were then provided the opportunity to verbally express their thoughts concerning the face validity of the OAS instrument. The responses from each person on the two surveys were linked using special coding.

A total of 1763 personnel took the back to back OAP-OAS surveys. The sample consisted of 85.2% males and 14.8% females, compared to the 1984 Air Force ratio of 89% male to 11% female. Ages ranged from 17 to 63, although

95% were younger than 47. The sample consisted of 13% officers, 79% enlisted, 7% General Schedule (GS) civilians, and 1% Wage Board (WB) civilians. The Air Force officer/enlisted ratio is 17% officer, 83% enlisted. Note that if the civilians were removed, the officer/enlisted ratio of the OAS sample would be 14% officer to 86% enlisted. Fifty-four percent (54%) had made either one or two PCS moves, and 24% had been on at least one unaccompanied PCS tour. Over 50% of the personnel had daytime work schedules. Racially, 76% of the sample were whites, 13% were blacks (compared to the Air Force's 15%), and 14% were others. Over 27% had more than 12 years in the Air Force, while 21% had less than 2 years of total service time.

Instrumentation

The version of the OAS used for the present study contained 104 attitudinal items for test purposes and ten demographic items in addition to those in the OAP. The OAS survey items were selected to meet two criteria. First, previous factor analyses had shown that the supervisory-related factors and the climate related factors on the present OAP are not separate. Therefore, one criterion was to test new items to create separate factors while eliminating some of the seemingly ambiguous items which presently load with the composite supervisor-climate factor. Second, consulting experience has shown the need for some additional factors not represented in the original OAP. Included in this group are factors pertaining to stress management and intergroup cooperation.

Procedure

Administration. During each survey administration, the OAP was administered prior to the OAS to prevent contamination of OAP results. After completing both surveys, participants were asked to complete a short questionnaire about the OAS. In addition, 364 respondents were randomly selected and verbally polled to see if they had additional comments about the survey. If so, those comments were recorded on the questionnaire.

Factor Structure. Derivation of the factors was accomplished by use of a principal components analysis with a varimax rotation using pairwise deletion. For factor solutions, the "eigenvalue greater than one" criterion was used. In addition, a Scree-test was used to help determine the optimum number of factors to extract. Following Ware, Snyder, and Wright (1976), the Factored Homogeneous Item Dimensions (FHID) criteria were used to assign items to factors. Under these criteria, it was required that all items in a factor have high loadings (± .40 or greater) only on that factor and low loadings (± .39 or less) on all other factors in the matrix. This method was a useful check not only for item homogeneity but also for item discriminant validity. An additional requirement imposed was that there be an absolute difference of at least .10 between the primary loading and the item's highest secondary loading. Numeric scores for negatively worded items were reflexed before the factor mean was calculated so that numerically higher factor responses always indicated more favorable responses.

Internal Consistency Reliability. The method of choice here was Cronbach's alpha. Generally, the most popular of the internal consistency methods, alpha can be obtained from a single survey administration and eliminates the inconsistency of splitting items. Its calculation is based on the number of items in a scale or factor and the mean interitem correlation for that same scale or factor. Usually, therefore, as the average interitem correlation or the number of items increases, so does the value of alpha. These procedures must be balanced, however. For example, there is an upper bound on significant increases in alpha from adding items. In addition, adding items that reduce the interitem correlation will not increase alpha. It should also be noted that alpha is often considered the lower bound of internal consistency reliability. Thus, alpha may generally be considered a conservative estimate of the true reliability of a scale or factor (Carmines & Zeller, 1979). Since it is difficult to attach significance levels to alpha, a more direct standard of comparison was used. For purposes of this study, alpha coefficients were considered acceptable at .50 or above, good at .70 or above, and high if .90 or above (Ware, Davies-Avery, & Stewart, 1978; Carmines & Zeller, 1979; Hendrix & Halverson, 1979a).

Results

Factor Structure

The initial factor analysis included 104 attitudinal items of which 47 were new items, 34 were reworded versions of OAP items, 14 were nearly identically worded to OAP items, and 9 were identical to OAP items. Each item was included to help measure one of the factors. Demographic items were not included in the factor analysis.

The initial factor analysis was actually two factor analysis problems, as the SPSS system utilized limited the number of items per run to 100. The initial analysis extracted 15 factors, 13 of which were expected; the remaining two factors accounted for less than 2.5% of the variance and had no items with loadings of 0.40 or higher. Based upon the results of the initial factor analysis, items which did not satisfy the item-loading criteria mentioned in the Procedure section of the report were eliminated from the OAS. Selecting the final items to be included in the OAS was an iterative process to ensure that each item loaded uniquely to its expected factor and to insure that no item that loaded uniquely onto an expected factor was left off the OAS.

The final factor analysis contained only items which satisfied all the item-loading criteria, insuring item "dimensionality." This reduced the OAS from 104 to 77 attitudinal items. Five items from the OAP which dealt with the Need for Job Enrichment were added to the items to be included in the OAS. The Need for Job Enrichment items were included in the final factor analysis even though they had not been field-tested as part of the OAS instrument. We expect these items to continue to load into the same factor once they are included in the OAS. The OAS survey contained five Combat Readiness items which, even though they loaded strongly into one factor, were removed from the OAS survey to be included instead within LMDC's Combat Attitude Survey. Only 11 of the final 77 attitudinal items had means above 5.5.

The 13 final factors are listed in Table 2. The factors accounted for 66.5% of the variance. We had hoped that the items relating to Recognition would load into a separate factor; however, these items loaded with the Organizational Climate items. Two of the extracted factors, Advancement and Intergroup Cooperation, consisted of three items and the Work Conditions factor consisted of two items. The factor loadings of these items onto their respective factors were all above 0.54.

Table 2. Derived OAS Factors

Job Goals
Job Characteristics
Task Autonomy
Training
Work Support
Work Conditions
Effective Stress Management

Supervisor
Advancement
Organizational Climate
Intergroup Cooperation
Work Group Effectiveness
Need for Job Enrichment

Internal Consistency Reliability

The 77 attitudinal items which loaded onto each of the 13 factors were used to determine Cronbach's alpha for each factor. Values of Cronbach's alpha ranged from .71 to .96, with 10 of the 13 factors having alphas of at least .84. The factors with three or fewer items, Advancement, Intergroup Cooperation, and Work Conditions had alpha values of .75, .71, and .89, respectively. Even though the three small factors had fewer tems than desired, their reliability scores were acceptable.

A Final Comment

The present research, then, shows the OAS factor structure to be statistically sound and generally consistent with the literature as well as previous experience with similar items on the OAP. The factors also showed very high internal consistency reliability. The combination of these two findings provides initial support for the OAS as either a consulting or survey research instrument. The OAS seems appropriate to replace the OAP as it provides a more solid, replicable factor structure with fewer items and factors than the OAP.

References

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment.* Beverly Hills, CA: Sage.

Hendrix, W. H., & Halverson, V. B. (1979a). *Organizational survey assessment package for Air Force organizations* (AFHRL-TR-78-93). Brooks AFB, TX: Air Force Human Resources Laboratory.

Hendrix, W. H., & Halverson, V. B. (1979b). *Situational factor identification in Air Force organizations* (AFHRL-TR-79-10). Brooks AFB, TX: Air Force Human Resources Laboratory.

Hightower, J. M., & Short, L. O. (1982). Stability of the Organizational Assessment Package factorial validity across groups. Paper presented at the 90th Annual Convention of the American Psychological Association, Washington, DC.

Hightower, J. M., & Short, L. O. (1983). Temporal stability of the Organizational Assessment Package factor structure. Paper presented at the 91st Annual Convention of the American Psychological Association, Anaheim, CA.

Mahr, T. A. (1982). Manual for the Organizational Assessment Package survey (Report Number 82-2560). Maxwell AFB, AL: Air Command and Staff College.

Short, L. O., & Hamilton, K. L. (1981). An examination of the reliability of the Organizational Assessment Package (LMDC-TR-81-2). Maxwell AFB, AL: Leadership and Management Development Center.

Short, L. O., & Wilkerson, D. A. (1981). An examination of the group differences aspect of the construct validity of the Organizational Assessment Package. Proceedings of the 23d Annual Conference of the Military Testing Association. Washington, DC: Military Testing Association, 1981.

Taylor, J. C., & Bowers, D. G. (1972). Survey of organizations. Ann Arbor: Institute for Social Research, University of Michigan.

Ware, J. E., Jr., Snyder, M. K., & Wright, W. R. (1976). Development and validation of scales to measure patient satisfaction with health care services: Volume I of a final report. Part A: Review of the literature, overview of methods, and results regarding construction of scales. Publication No. PB-288-329. Springfield, VA: National Technical Information Service.

Ware, J. E., Jr., Davies-Avery, A., & Stewart, A. L. (1978). The measurement and meaning of patient satisfaction. Health and Medical Care Services Review, 1(1), 1-15.

Webster, A. H. (1982). An assessment of the construct validity of the Leadership and Management Development Center's Organizational Assessment Package (Report number 82-2640). Maxwell AFB, AL: Air Command and Staff College.

PRODUCTIVITY MEASUREMENT: ISSUES AND CHALLENGES

Paul van Rijn, Ph.D.

US Army Research Institute[1]
5001 Eisenhower Avenue
Alexandria, Virginia 22333-5600

This paper describes some of the lessons learned during the Army Research Institute's (ARI) efforts to conduct an independent outside evaluation of the outcome of a macro-level sociotechnical system (STS) intervention at a large Army maintenance Depot. This particular intervention focused on a major division of the Depot which consisted of about 900 workers, mostly Army civilians in skilled blue-collar occupations, and which had primary responsibility for the repair and overhaul of the airframe (as opposed to the engine and transmission) of Army helicopters.

The STS analysis and the development of the design recommendations were conducted by two outside STS experts working closely with a team of 12 Depot volunteers who had been carefully selected to represent both the different levels and occupations within the Depot. This phase of the intervention lasted nearly one year and involved: (1) the development of a Depot Philosophy Statement, (2) the explicit articulation of the mission of the Depot, (3) the identification of the key variances (deviations from the norm) affecting the work of the Depot and their controls, (4) an analysis of the social (people-related) system of the Depot, and finally, (5) the development of 17 specific recommendations for change.

Although ARI's preliminary efforts to evaluate the intervention outcome were initiated from the beginning of the STS analysis, these efforts were not a prominent component in the discussions of the STS analysis or in the development of the recommendations. Goals and expectations for increased productivity and enhanced quality of working life were expressed, but seldom in operational or measurable terms. ARI's role was that of independent outside evaluator.

Toward the end of the STS analysis, when the change recommendations had been developed, this researcher first became responsible for ARI's evaluation of the intervention. Numerous archival measures had been identified previously as potential indicators of increased productivity and a survey instrument had

[1] The views expressed in this paper are those of the author and do not necessarily reflect the views of the US Army Research Institute or the Department of the Army.

been develcped to assess key organizational variables. Implementation of the recommendations had not yet been begun and the conditions for tracking the intervention before, during, and after implementation seemed optimal. However, appearances were deceptive.

First, it became apparent that the archival measures identified were much more complex than had previously been assumed. Interpretations of the simplest measures, such as "Number of aircraft produced," surfaced a long list of questions. Did the count include all aircraft or only some models? Were crash-damaged aircraft with their special requirements included in the counts? And to what extent did the overhaul of non-Army aircraft and special projects affect the measure?

It was also apparent that, with few exceptions, the "number of aircraft produced" matched exactly the number that had been contracted each month. If 30 aircraft were required, 30 were produced. If 28, then 28. Overproduction would result in the aircraft being "carried over" into the next month or would result in diverting work to neglected shopfloor maintenance or other support tasks. If production was behind schedule, the remedy might be to authorize more overtime or schedule extra shifts until the schedule was met. This often resulted in a flurry of activity toward the end of the month.

Other measures posed similar challenges, not the least of which was that for every measure there were multiple reporting systems. The printout of productivity report titles alone was over one-half inch thick. Nearly everything was counted and logged into the highly automated reporting systems. There were no simple ways for sorting through these productivity reports to determine which reports would be most useful for a meaningful evaluation of this intervention. There was no single expert who could advise on the utility of each report, many of which were produced solely to comply with reporting requirements and formats imposed from outside the Depot.

Due to the multiple reporting systems, it was not unusual to find a measure as non-complex as "sick leave useage" produce inconsistent or discrepant results. Often the data would be different because they were computed differently. For example, on one report, "sick leave" was the rate per employee per 100 work hours, while on another report it was the rate per pay period, or 80 work hours. Even the term "monthly" took on different meanings. On different report it might refer to the varying number of calendar days, the available number of workdays, or simply two pay periods without regard to the number of actual workdays involved. Seldom were these variations obvious from the report titles and painstaking investigations were required to resolve even the most obvious discrepancies. These individuals responsible for maintaining the various reports were often not aware of these discrepancies, since they maintained only one type of report. And although they were highly dedicated, those maintaining the printouts were not always the same people who knew the precise details of the derivation and meaning of the numbers being reported.

A second major measurement issue that emerged was the level of analysis. It has already been suggested that global measures, such as "number of aircraft produced," are often difficult to interpret and may mask real productivity gains made at more molecular levels. Figure 1 shows some of these levels and the types of subtasks that are required to process a helicopter airframe through its 18 overhaul stations. Over 60 relatively independent work centers (not all are shown) are involved in maintenance and overhaul of the helicopter airframe. Each work center has its function, from disassembly to electronics repair, to painting and flight test.

Reducing the specificity of measurement to this level might be intuitively attractive were it not for the tremendous volume of data that would somehow need to be collected, analyzed, and synthesized. This would represent not only be a collosal task for the researcher, but, more importantly, it would place considerable unexpected strain on the resources of the organization to duplicate and make all this data available.

Besides the volume of data that would be generated by a work center level of analysis, the researcher would now also have to face the challenge of comparing and aggragating data across work centers with very diverse technical processes and different outputs. Except for a few measures, such as sick leave, there were few productivity indicators that could be aggragated meaningfully across work centers. In addition, it raises the question of how a researcher can assess the productivity of one work center when the productivity of that work center depends significantly on the productivity of one or more other work centers? The paint shop, for example, cannot be expected to demonstrate high productivity, if it is not provided sufficient numbers of airframes to paint or if those airframes all arrive at the same time.

Other major issues are the reliability and validity of the measures themselves. Even some of the more promising measures proved on closer scrutiny to be highly variable from month to month and were likely to be of dubious quality for research or evaluation purposes. The large monthly variations could often be traced to unevenness in the submissions of data for the report, individual differences in the criteria used to report the data (e.g., What is a reportable defect?), computer system failures or software upgrading, undocumented changes in the way the data are recorded or calculated, re-establishment of the engineering standards or norms from which performance efficiency ratios are calculated, and so on.

The timing of the evaluation measurement is also critical. In an intervention as large and complex as that of the Army Depot, it was difficult to specify at what point in time a measure is

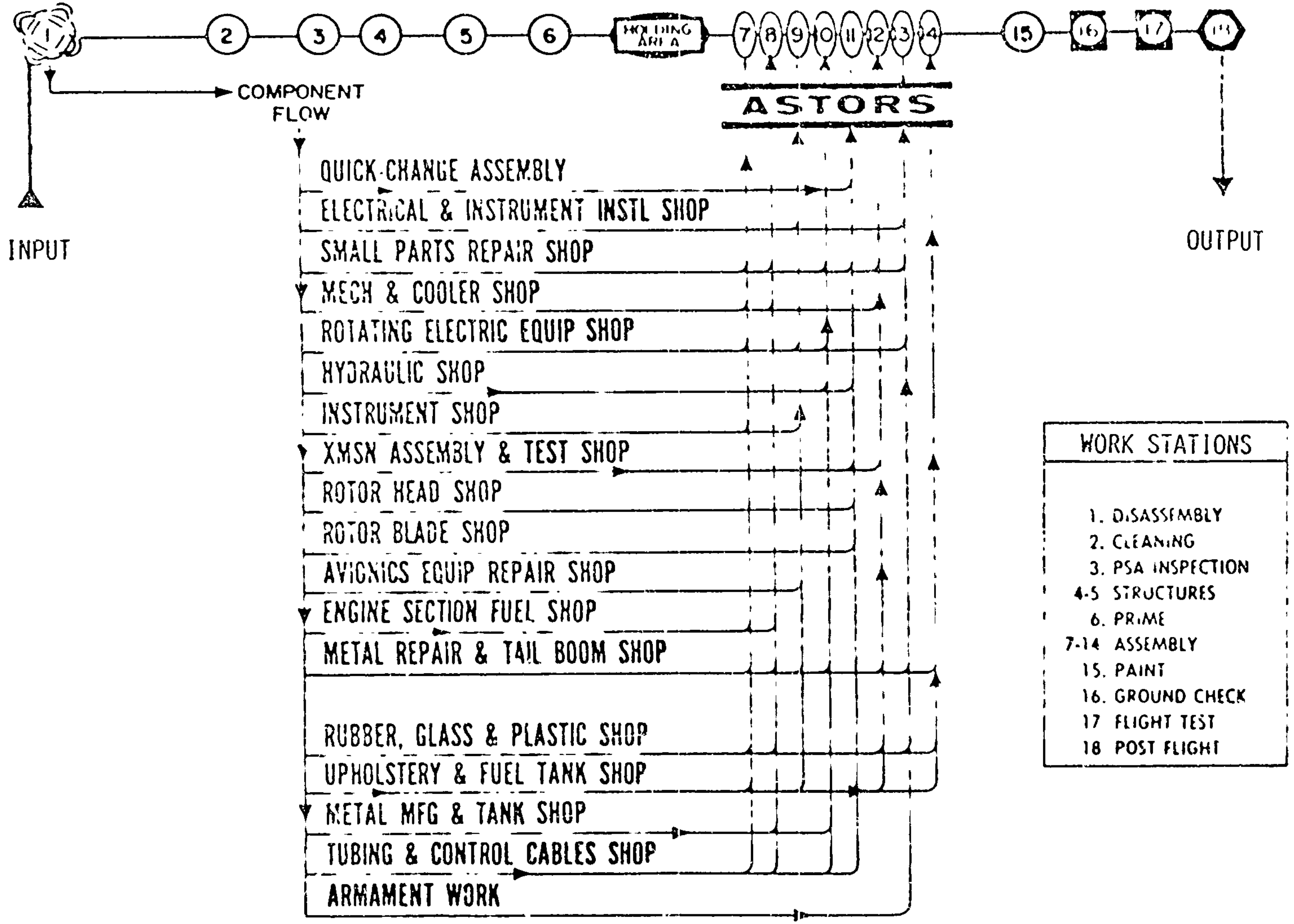


Figure 1. Schematic diagram of the workflow through the 18 work stations of the Airframe Division. "ASTORS" is the Automated Storage and Retrieval System.

before, during, or after implementation. This was due to the fact that different recommendations were implemented at different times and had different durations. Training, for example, can be expected to result in an early decrease in productivity with its benefits not becoming fully realized for years. Other recommendations, however, can have a more specific and immediate impact. Finally, it can be argued that implementation started the moment the analysis phase was begun -- long before the recommendations were implemented or even developed.

Another important measurement issue is the identification and control of confounding variables. It is highly unlikely that during an intervention of any magnitude that there are not also other factors that impact on the measures used in the evaluation. For example, during the intervention at the Army Depot, the Depot went into a massive hiring mode. Does this increase productivity? Or, does it decrease productivity as productive workers have to divert some of their attention to the training of the new and unskilled hires? Merely identifying these potential confounding factors is a difficult task, but this is not nearly as difficult as assessing the direction and magnitude of the effect, or of determining how the effect varies over time and from work center to work center.

Finally, the survey instrument, developed to study Depot organizational characteristics that might contribute to an effective intervention, proved to be insensitive to the changes that were ultimately implemented. This was because the survey focused largely on the structure and work processes of the Depot; while the intervention, as a whole, did not significantly impact on these areas. Given the nature of the organization, major organizational or process changes would have been highly unlikely. Consequently, it should come as no surprise that there were virtually no changes detected over the 14-months timeframe in which the survey was administered. The few questions that did demonstrate an attitude change (some over 20 percentage points) tended to be questions that related directly to the philosophy and principles of the STS approach to organizational change.

Lessons Learned

Based on the experiences with this particular intervention, a number of productivity measurement lessons were learned:

1. The measurement of an organization is likely to be much more complex than it initially appears.

2. Independent outside assessment of an organizational intervention is counterproductive. First, it does not focus the assessor on the most meaningful measures and it does not provide for a meaningful reduction of the multiple measures that are available.

3. Interpretation of the numbers alone, without full knowledge of the process from which they derive and the context in which they occur is likely to be highly misleading.

4. The assessor must work very closely with the organization -- from the very beginning of the intervention -- and assist the organization in identifying the measures that will be used to assess the success of the intervention. This requires defining the goals and expectations of the organization and the outcomes of the change recommendations in terms of measurable outputs.

5. To the extent possible, the assessor and the organization should be co-investigators and partners in the intervention, mutually learning about real organizational concerns and both being concerned that the data are meaningful and trustworthy and that the inferences derived from the data are logically sound.

6. To the extent possible, the different stakeholders in the intervention need to be identified early and their criteria for judging the pluses and minuses resulting from the intervention must be articulated.

7. Searching for unitary cause and effect relationships is inefffective. The assessor and organization must work together to clarify the major interdependencies and contexts in which measurement occurs.

8. The measurement aspects of an intervention need to be an integral part of every phase of the intervention. To the extent that positive outcomes can be identified early, the momentum of the intervention is likely to be sustained. It may be advisable to deliberately include in an intervention some recommendations that are likely to have an early demonstrable payoff. The more these payoffs can be expressed in terms of "hard" dollars saved or productivity gains, the more likely the intervention will continue to receive the top-management support it needs to remain resourced.

9. Finally, resources required to collaboratively assess an intervention are not trivial but are an essential and important investment, not just for the evaluation of the particular intervention, but also for the continued monitoring of an organization's productivity and quality of working life.

A MORALE AND MISSION RELATING STRATEGY

Raymond O. Waldkoetter
U.S. Army Soldier Support Institute
Fort Benjamin Harrison, Indiana 46216-5060

The Mission Area Analysis (MAA) assesses the long-range capability of a programed force to perform required combat tasks. The analysis is designed to discover task deficiencies and correct them with changes or solutions in doctrine, organization, training, and materiel. The MAA process also provides a basis for applying advanced technology to future Army operations with the inherent aim of increasing combat effectiveness. The MAA is performed by a study group at the Training and Doctrine Command (TRADOC, 1985) Centers or schools. There are 13 MAAs, projected to be conducted in groups of three once every four years, and are usually initiated prior to the research, development, and acquisition cycle. Throughout an MAA's analytic process, the soldier must be considered as an integral system. Deficiencies in combat effectiveness must be remedied either by improving soldier performance directly or by changing doctrine, organization, training, and/or materiel. In any case, the soldier is the key element in the combat equation.

The individual soldier must be represented as a "combat system" integrating TRADOC combat development activities, relating to improving individual performance with human technologies or equipment that will enhance individual capabilities during combat (Weisz, 1980). Soldier factor issues are identified during the MAA process to advise or assist proponent TRADOC combat developers in solving problems related to soldier motivation, capability, and performance. Direct assistance and advice are given to proponents so that the focus is on consideration of soldier factors as the means to enhance individual and unit performance. The Soldier Support Center (SSC) works with proponents in developing MAAs and study advisory follow-through to assure soldier factor consideration is also carried over into the materiel acquisition process (MAP).

The purpose of soldier factor consideration is to: insure that potentially critical soldier factors are considered in the MAA process; provide a guide for considering soldier factors in the MAA process; and promote efforts to assess the impact of soldier factors on MAA deficiency solutions. Study efforts in the MAA process should specifically include consideration of soldier factors. After-the-fact consideration of soldier factors is largely inefficient. The individual soldier should not be regarded as an add-on to a materiel system. Soldiers must have a defined role within the mission area. Serious consideration of soldier factors can reveal ways to avoid substantial degradation and enhance ability to perform the mission.

The views and opinions expressed in this paper are those of the author and should not be taken as an official policy of the Department of the Army.

METHOD

Soldier factor issues become critical and must be remedied to the extent that they detract from the soldier's job and combat performance. Soldier factors affecting morale and mission performance must be defined in some scheme to begin to understand how they may affect duty or combat. Historically, commanders have relied on the human/soldier dimension or "moral force" to decide conflicts when their forces were equivalent in other respects (Zais, 1985). Knowing relationships probably exist between the global notion of morale and mission success has not led to any readily reducible process to confirm these relationships. Since factors related to morale and mission success are so numerous, it seems best to devise a logical scheme to analyze "some," appearing to have noticeable affects.

Soldier factors, then, are defined in this paper to be such behavioral determinants as cohesion, stress, values or ethics, mission (sense of), and performance (sustained). Studied in the MAA context, these soldier factors are also seen as being affected by three primary determinants - systems/ weapons, individual/group characteristics, and leader/management actions - adding to the comprehensive soldier factor affects. Morale as a collective state-of-mind can be identified by the interaction of these soldier factors when producing a positive attitude expressing motivation and satisfaction (Motowidlo, Dowell, Hopp, Borman, Johnson & Dunnette, 1976). Thus, the fully adequate consideration of soldier factors also conveys the basis for understanding and assessing the level of unit soldier factors as morale.

Soldier performance is affected to a lesser or greater extent as soldier factors are used as corrective actions. Materiel has given limits, (i.e., tensile strength and payload), and consideration must be given to reliability, availability, and maintenance. Soldiers also have their given limits, (i.e., strength, endurance, and morale); and again, consideration must be given to "responsibility, availability, and maintenance" (fitness and readiness). A thorough consideration of soldier factors will likely guide the planning for and improvement of combat performance beyond the MAA process. The concept of soldier factors - cohesion, stress, values or ethics, sense of mission, and sustained performance - may be developed under the primary determinants with descriptive action-task (AT) statements suggested for each factor to improve mission analysis and corrective action design.

Each proposed corrective action or solution to an identified task deficiency is derived from concepts related to doctrine, training, organization, and/or materiel. A recommended solution following the soldier factor review will succeed only to the degree that it remains compatible with other projected changes and is supported within the given resource constraints. The perfect solution is an elusive goal and the best alternative may demand tradeoffs in deciding which soldier factors should have the most attention in obtaining a solution. As the MAA proponent task force or assisting soldier factor analyst examines the interface between solutions and soldier factors, the following procedures serve to implement consideration of soldier factors in the MAA process. Consideration of the soldier factors has to identify and determine for proposed solutions: action-task (AT) statements that will affect soldier combat effectiveness; task conditions and standards; the combination of soldier factor considerations that can most likely affect task and combat success; any expected changes in soldier factors related to proposed solutions; and, essential elements of analysis and measures of effectiveness for soldier factors. The probable improvement in critical tasks is determined in a proposed solution by considering the extent to which those negatively contributing factors can be reduced or countered, and positively contributing factors can be augmented and supported for corrective action. The total impact of soldier factors is evaluated and determined through the: impact of the proposed solutions on soldier factors; impact of the soldier factors on proposed solutions; soldier factor initiatives, constraints, and limitations in proposed solutions; and, acceptability and feasibility of implementing the proposed solutions.

RESULTS AND DISCUSSION

The analysis proceeds from evaluating the impact of proposed solutions on the most critical soldier factors to questioning how aspects of proposed solutions will affect one or more of the soldier factors and vice versa. For example, will a proposed solution "require leaders to increase training to cope with increased intensity of battle (stress)?" Action-task (AT) statements like this are used to judge the affordability and supportability of a solution. The AT statements can be used as operational task statements and prioritized in terms of importance for the proponent mission. Priority for mission accomplishment can be indicated by having a subject-matter panel judge whether they disagree or agree with the level of relevance of an action toward solving mission area needs. Then any deficiencies associated with the prioritized action can be identified and remedied by the most appropriate types of solution. A second way the proponent may utilize AT statements is by prioritizing actions most to least pertinent. Thirdly, if a deficiency would occur in any primary factor set, judges or raters can indicate to what degree it would increase the probability of inadequate performance. As corrective actions/solutions are reviewed in relation to selected soldier factor actions or issues, then the proposed solutions for doctrine, training, organization, and/or materiel will be defined and prescribed accordingly. If a solution is constrained by any soldier factors it may have to be altered.

Although numerous AT statements can be generated, in most cases 20 or less will suffice if an in-depth review is conducted. Some 105 AT statements were generated with 21 for each factor and seven for each factor under a given primary determinant. Any variation of AT statements is possible depending on the mission area concerns. Table 1 is presented as a working example using an analytical scheme based on occupational-task analysis procedures.

TABLE 1

SOLDIER FACTOR CONSIDERATIONS

Systems/Weapons (Parameters/capabilities of systems/weapons, how to employ, and what to expect from adequate use and effects.)	Individual/Group Characteristics (Physical health, skilled task performance, skilled interpersonal performance, integrity/mental health; group goals, skills, endurance, and efficiency.)	Leader/Management Actions (Refined mix of goals, behaviors, power, and style used to insure effective combat operations and productive effort.)
A. Cohesion (Sense of belonging, feeling a part of something, sharing of problems, or bonding together of soldiers and leaders to accomplish any purpose resolved.)		
1. Assure system/weapon training is related to suitable skill level.	1. Estimate whether personnel can function in more than one (1) job.	1. Communicate clear leader goals and objectives for survival.
2. Determine operational indicators to improve combat assignment.	2. Monitor impact of probable personnel actions on troop confidence.	2. Set proper examples in combat exercises and combat.
3. Analyze need to coach or add more skill for system/weapon team.	3. Identify special skills to cope with unusual problems or situations.	3. Initiate motivating rather than coercive L/M actions.
4 Identify expected operational measures to sustain efficiency.	4. Develop attitudes to support troop loyalty.	4. Review and apply cohesive practices and information.
5. Specify most helpful actions to support operator/teams.	5. Support job satisfaction interests in preparing personnel assignments.	5. Take part in a cross section of activities to help official direction.
6. Project critical time allocation for major system/weapon functions	6 Define duties to take advantage of available skills.	6. Supervise and participate in expected standards of excellence.
7 Assess strengths and constraints in achieving proficient operations.	7. Recognize examples of esprit and sacrifice in battle.	7. Demonstrate reliability and compassion in crises.
B. Stress (Reaction of the mind/body to extreme demands with sensations of tension and anxiety, and which must be managed to avoid degraded performance.)		
1 Confirm the adequacy of communications for system/weapon selected to be deployed	1. Analyze combat exposure levels and parity among troops.	1. Define procedures to manage stress continually in action.

Essentially, the impact of proposed solutions on soldier factors is found by indicating which factors are expected to be affected, with some idea of the magnitude of consequences anticipated. It may be adequate to merely identify the soldier factors affected and indicate the order of severity. A more detailed interrogative analysis may be desirable to define just what, how, and why, the factors are posing as unresolved and critical soldier factor issues. The impact of soldier factors on proposed solutions can be approached by identifying those factors which can alter constraints or limitations in solutions. These factors, if augmented or supported, can act as a "force multiplier" for continuous combat operations.

Conducting a productive consideration of soldier factors engages as much detailed analysis as a proponent can afford. An iterative consensus is recommended to relate the magnitude of soldier factor and AT statement judgments to primary determinants and mission success. Numerous questions are required in regard to mission areas to assess the importance of soldier factors and the interaction among such factors. Each proponent will have to decide on the scope and level of resolution needed to deal with soldier factor issues in the given mission area. Corrective actions and solutions for task deficiencies require detailed analysis to audit their development from task deficiency priority to type of solution (i.e., training). A speculative scenario will allow the analyst to relate the probability of mission success to morale by progressively assessing a primary determinant, soldier factors, and AT statements, as these indicate a need for corrective action.

The analyst or subject-matter panel can select one of the primary determinants for guiding soldier factor analysis (systems/weapons, individual/group characteristics, leader/management Actions). A value or rank of importance may be assigned in comparison with the other determinants, e.g., 1, 2, or 3. Next, the analysis process can estimate the magnitude or criticality for each of the five soldier factors toward assuring adequate combat effectiveness under the given determinant and threat scenario. If the five factors would not appear "critical to adequacy" to the analyst or panel, the scale magnitude would tend to be at the lower scale values. That is on a scale of 1 to 5, the relative priority of factors would tend to vary from 1 to 3. Now, the analyst or panel can select the AT statements considered most useful or needed to favorably modify the five soldier factors by augmenting or enhancing performance and related perception. Additional or new AT statements or items can be constructed as may be necessary. More detailed AT items might be prepared under each chosen item to bring about the consideration or desired improvement through one or more corrective actions or soldier factor solutions. Next, the magnitude or criticality would be estimated toward enabling the major AT item or subitem to be performed, i.e., by rating on the probability of inadequate performance on a scale of 1 to 5 (low to high).

The final phase of analysis is to derive a combined magnitude of the three levels - primary determinant, soldier factor, and AT item/subitem - to estimate the criticality expected for some composite probability of inadequate performance (Stark & Waldkoetter, 1985), specifying needed change or solution for mission success. The three-level estimate of criticality and multiplicative/additive combining of each level can yield a conditional/notional index of priority with analyst/panel consensus for each of the primary determinants. As the most simplistic example it could suffice to base a conditional analysis on a single primary determinant rather than all three. If Leader/Management Actions was ranked highest for a value of 3, the highest rated soldier factor with a value of 5 and one AT item alone rated for a value of 5 a product of 75 is obtained. Then, if the other remaining ranked factors

(1 to 4) have only one AT item each rated with values of 5, a sum of 50 would result to add with the 75 for a total of 125. The lowest possible overall total is 15 following such a forced solution. A scale mid-point of 70 and above may be arbitrarily given as the hypothetical area in which serious indications have morale implications and are noted for the probability of inadequate performance. Further, if Leader/Management Actions in Table 7 was considered of highest combat importance, the Cohesion factor was ranked highest for combat effectiveness (criticality), and only one AT item, 1. Communicate clear leader goals and objectives for survival, was given the highest probability for inadequate performance, then the analysis process would have yielded an indication of trouble in need of some corrective action. That action being to make certain the communication task was mandated in doctrine, training, organization, and/or materiel to build cohesive performance utilizing leader action to insure the expected resolution.

This pre-R&D overview offers a procedural basis for soldier factor (SF) consideration that will logically and systematically identify potential SF constraints and needs related to soldier impact on combat effectiveness. Analytical consensus and review should strongly influence this procedural model as a notional strategy. In this approach it seems logical to infer that if leader, cohesion, and communication actions are expected to be impaired by critically inadequate performance, then morale and mission success are inexorably related. Some principle of proportionality must apply in that any task action for either demands a complementary action for the other to achieve a dynamic balance.

REFERENCES

- Motowidlo, S.J., Dowell, B.E., Hopp, M.A.. Borman, W.C., Johnson, P.D. & Dunnette, M.D. (1976). Motivation, satisfaction, and morale in Army careers: A review of theory and measurement. Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

- Stark, R.W. & Waldkoetter, R.O. (1985). Tactical communications Mission Area Analysis: Human fatigue and stress study (draft). Fort Gordon, GA: U.S. Army Signal Center.

- U.S. Army Training and Doctrine Command (TRADOC). (1985). Mission Area Analysis Handbook (draft). Fort Monroe, VA: Headquarters, Deputy Chief of Staff for Combat Developments.

- Weisz, J.D. (1980). Human factors engineering in research, development, and acquisition. Aberdeen Proving Ground, MD: U.S. Army Human Engineering Laboratory.

- Zais, M.M. (1985). Ardant du Picq: Unsung giant of military theory. Army, 35, 56-64.

COGENT

A MICROCOMPUTER BASED CRITERION REFERENCED TESTING SYSTEM

DALE KERKMAN
TRAINING DEVELOPMENT UNIT
NAVAL TRAINING CENTER
GREAT LAKES, ILLINOIS

COGENT is a microcomputer based academic management tool that offers a workable, reliable, and efficient solution to the implementation and administration of criterion referenced testing. It provides a means of generating printed tests and evaluating student responses to not only each test but also to distinct content areas within each test. These content areas can be related to the objectives within the test. COGENT provides detailed test results which indicate mastery of the course objectives as well as student tracking and record keeping.

COGENT has a built in test authoring system that enables a person to easily enter test items into an item bank. An editing function allows the test author to make charges to an item, delete an item, insert an item, or move an item to another position within the item bank. Each test can be subdivided into up to 10 content areas. Each content area contains all of the test items for one discrete concept (terminal objective, enabling objective, lesson topic, etc). In addition, the content area parameters, which are used in the printing of the test and in determining if criterion was met for each content area, are easily entered into the system with the author program. Up to 100 test items may be entered into each distinct content area.

A test printing function enables the operator to easily print a hard copy test. Pre-defined parameters determine how many items are selected from each of the content area item banks. A maximum of 100 items can be selected from the content area item banks and printed. The test items can be randomly selected by the computer or manually selected by the operator. The test items can be printed in random or sequential order. The choices for each of the test items can also be printed in random or sequential order. If computer selected and random printing of test items and choices is chosen when printing new tests each test printed by COGENT will be unique from any other and will assist in minimizing the possibility of test compromise. Multiple choice and true/false items are accommodated in the test item banks.

COGENT will significantly decrease the number of manhours presently required to prepare, revise, type and print new examinations. New variations of a test can be generated as often as desired.

The tests are scored from optically scanned answer sheets. COGENT analyzes not only the entire test but also individual content areas to determine if each student mastered the subject matter contained in the content areas. Test results are printed after the last answer sheet has been scanned. The report lists all of the students who took the test along with their associated overall test scores. It also includes for each student a score for each of the content areas and a notation indicating any failed content areas. Individualized prescriptive assignments can also be printed for those students requiring remediation and retesting. Customized remedial tests for each of the students requiring retesting can also be printed. These tests will be made up of only those items from the failed content area.

COGENT allows students to be transferred from one class to another with all test data remaining intact. Students can also be put into a 'hold' status and then transferred to an active class. In addition, students can be disenrolled from school. Any student that is transferred, dropped or put into a hold status requires a corresponding code to be entered so that reports can be printed listing any and all students by any desired code.

COGENT provides the following reports:

*Class Roster - provides an alphabetical listing of the students including name, rate, SSN, and other pertinent student data.

*Student Cumulative Summary - provides a listing of all the students along with their overall score on each test. Included is a notation indicating the students' pass/remedial status for each test. This notation indicates the number of times each student was retested in order to achieve criterion on all of the content areas for each test. This report also includes each student's cumulative average and the class cumulative average.

*Content Area Performance Summary - prints a detailed analysis of each student's performance on each test. It includes the overall test score, the score for each content area the first time the test is given and any/all scores on subsequent remedial tests. It lists the class performance statistics such as: number of students taking the test; class test average; number of students passing all content areas on first attempt; number of student requiring remediation after first attempt; etc.

*Student Ranking - provides a printout of the class with the students ranked high to low based on each student's cumulative average. In addition, those students meeting honor graduate requirements can be recognized with an appropriate notation.

*Academic Review Board Worksheet - printed for those students who experience academic difficulties. It is used by the Academic Review Board in determining academic action.

*Student Memo - form printed for students who are assigned remedial study/testing. It lists the time and location where the remedial study/testing will take place and is used to assist in managing the remedial study and retesting program.

*List of Transfers/Drops - provides a printout of the students who were either transferred to another class, or disenrolled from school. It also includes the transfer/drop code and the date the action took place. This report can be printed for one student, all students, any specified transfer/drop code, or all transfer/drops codes.

*Item Analysis - provides a D&V item analysis and also a response count analysis which are used to determine the effectiveness of the test items. It also includes the mean, median, and standard deviation for the test.

Some of the benefits that can be realized by COGENT are:

1. Ability to identify specific problem areas within a test:

COGENT grades the answer sheets and evaluates each test by content area (objectives). This enables the proctor/instructor to identify problem areas and remediate to those specific areas. For example, with conventional testing a test covering subject matter such as Introduction to the Naval Chain of Command; Naval Terminology; AC Theory; DC Theory; and Safety Precautions would be evaluated on a students overall test score. It is very likely that a student could pass the test with an overall score of 70% but incorrectly answer the items on the Safety Precautions subject matter. With COGENT each of the individual subject matter areas (content areas) are evaluated and a report is generated listing all of the students and their scores in all content areas. The areas where the student failed to meet criteria are annotated so that individual remediation can be prescribed.

2. Reduction in remediation/retest time:

Without COGENT, if a student fails a test, he/she is remediated and retested on the subject matter that is covered by the entire test. Since specific problem areas within each test are readily identified by COGENT, the student is remediated and retested ONLY on the specific areas where he/she is experiencing difficulty. COGENT can automatically generate a content area specific remedial test that is tailored to the problem area/s of each student.

3. Reduction of errors in grading tests:

All tests are computer scored so that the possibility of an error in grading is greatly reduced.

4. Reduction in time to update/add/change items in the test item banks:

The authoring program allows the test manager to easily add or delete items in the test item bank. Also, changes can be made to existing items. For test audit purposes, each test item has the corresponding objective annotated along with the data each item was entered in to the item bank.

5. Reduction in time to validate test items:

The item analysis option allows the test manager to easily generate statistics which can be use to determine item difficulty and validity.

6. Reduction in time to print a new test:

The time to print a new test is reduced from hours to minutes because the computer automatically prints the tests from predetermined data.

7. Reduction of compromised tests:

All test items can be randomly drawn from a bank of items. The items can then be printed in random order with the distractors also randomized. This significantly reduces the possibility of test compromise.

8. Reduction of student/class management time:

COGENT maintains a record of all student performance/grades and provides various reports by individual or for an entire class. COGENT also provides a roster and a class ranking report with notation for prospective honor graduates. Printed forms for the Academic Review Board for those experiencing academic difficulties can also be printed. Students can also be transferred from one class to another with all grades remaining intact.

9. Ability to effectively administer a criterion referenced curriculum:

Without the automatic scoring and evaluation of tests, it would be impossible to attempt to administer an effective criterion referenced curriculum. COGENT automates all of the functions necessary to perform these tasks which results in a more highly skilled graduate.

COGENT is currently operating on an Apple computer with a 10 megabyte hard disk. In this configuration, it can accommodate up to 26 tests with each test consisting of up to 10 content areas. Each content area can have up to 100 items in the test bank. COGENT can manage the statistical data for up to 31 classes, each class consisting of up to 50 students.

COGENT is currently being converted to the Zenith model-120 with a 10 megabyte hard disk. This version will be able to manage multiple courses/curriculums with increased statistical reporting capabilities. Both versions utilize an Epson 80 column printer for the printing of reports and tests and a Scantron model 1200 optical mark reader for the grading of the students answer sheets. The selection of hardware was based solely on existing in-house equipment.

Mircocomputer-Based Field Testing for Human Performance Assessment

P. J. Merkle, Jr., R. S. Kennedy, M. G. Smith,
J. H. Johnson (Essex Corporation)

Abstract

The chief advantages of paper-and-pencil instruments for field research are economy and simplicity, but they can have low response rates, exhibit questionable subject anonymity and security, and typically they require a proctor. During on-site investigations of the side-effects from flight simulator operations, paper-and-pencil forms and an Automated Portable Test System (APTS) were administered. The APTS functioned reliably, data production increased, and a high incidence of simulator side-effect symptomatology was detected.

The APTS is comprised of a test battery and questionnaires embodied in a microcomputer. The battery includes tests of cognition, information processing, psychomotor skill, memory, reasoning, and others. The questionnaires are: a mood adjective checklist, motion sickness symptomatology (with automated scoring), and motion sickness history Originally conceived as a behavioral toxicology assessment tool, the APTS has applications for personnel selection and classification, education and training assessment, clinical diagnosis, and health care delivery systems.

Several military and university facilities have purchased, rented or borrowed an APTS and have begun collecting data to determine the sensitivity to various drug, environment or treatment conditions. In these preliminary studies one or more tests have been shown sensitive to morphine, chemo-radiotherapy, sleep loss, hypoxia, amphetamine, hyoscine, etc. Some of these results are reviewed.

Introduction

Prior to the advent of low cost, microcomputer systems (ca., 1972), computer technology was slow to find a place in human performance and laboratories. Several factors contributed to this lag. First and foremost, the cost of minicomputers was prohibitive for most facilities, while that of mainframe systems continues to be so. Additionally, the time-sharing operation mode of undedicated computers, required to make them cost-effective, can make them less reliable assessment instruments than traditional methods; the speed with which such computers operate is dependent upon the number of users accessing the system at any specific time. This factor results in inconsistent timing parameters from one testing session to the next. If the system is "down" for maintenance or development, all data collection comes to an abrupt halt. Although mainframes and minicomputers opened new frontiers for study, they did not prove to be cost-effective for most behavioral science laboratories.

Microcomputers expand the potential for the study of psychological phenomena in at least two ways. First, they permit more comprehensive measurement than traditional tests by providing latency and other information not ordinarily available with traditional approaches. Second, they are capable of controlling devices which produce speech, and they are suited for complex video displays, thereby increasing the number of sensory modalities that can be involved in a testing situation. Microcomputers are also well suited for memory tests, because unlike paper-and-pencil methods, they can present stimuli for short periods of time.

Potential advantages of microcomputers for psychological testing may include: (1) identification of true deficit in performance from developmental problems; (2) standardized presentation which may lead to improved comparability of tests; (3) higher test reliabilities due to more accurate control of stimulus material; (4) bypassing of infirmities (e.g., memory deficits, dyslexia, dystrophy) of certain groups, with performance testing in innovative modes (e.g., voice recognition, touch panels, large or back lighted keyboards, eye-tracked systems); (5) more comprehensive assessment of individuals; (6) potential for new assessment paradigms and perspectives for understanding of human performance; and (7) possible provision of more intrinsically motivating tests to subjects.

Automated Portable Test System Overview

The AUTOMATED PORTABLE TEST SYSTEM (APTS) is the first complete, compact (can be hand held) system of its kind. It is being produced expressly for human performance assessment whether in unusual environments, with toxic substances, or with other treatments. APTS is capable of controlling and administering complex psychological testing routines, while entering and collating responses and latencies with accuracy and precision. The battery includes tests of cognition, information processing, psychomotor skill, memory, reasoning, and others.

The APTS makes it possible to maintain a professional research workspace. Any data collected or documents written on the APTS can be printed out for examination, sent via modem to the home office, or loaded onto an external cartridge for easy transport to other locations. Also, information from other locations can be communicated through external cartridge or cassette tape (new tests to use, modified data formats, etc.).

The APTS is comprised of three subsystems: (1) hardware; (2) test programs; and (3) system control.

Hardware

The hardware subsystem has been developed around a notebook sized 8-bit personal computer: the NEC PC 8201A. Integral to the microcomputer is a 32K internal read only memory (ROM) containing, in addition to TELCOM and TEXT EDITOR, a version of Microsoft BASIC. The technical features of the microcomputer are more fully described in NEC User's Guide (1983). Within the small, lightweight package, the system has: substantial onboard random access memory (RAM) capacity expandable to 96K; an external battery option (8 A h) providing for more than 100 h of continuous operation; and a built in display. Refer to Figure 1 for a hardware overview.

Augmenting the notebook microcomputer are the wide variety of auxiliary components. Among these, the (32K) RAM cartridges have proved particularly useful in applications to date. For field applications, the APT System and Testing Programs are maintained in internal RAM, and, after data collection, data are transferred to a RAM cartridge for mailing or carrying from remote sites to a centralized data-base location. (See Figure 1) For laboratory applications, it is anticipated that researchers may find it useful to extend the capabilities of the microcomputer with an external display (CRT), floppy disks, and computer interfaces. Overall, the NEC PC 8201A has the expansion options required for a wide range of field and laboratory applications.

Test Programs

The APTS component programs are developed following an iterative three-stage process: identification, mechanization, and evaluation. The identification, until now, has been on the basis of sound metric properties. Future considerations will include operational relevance and prediction and construct validity. The mechanization is conducted in-house and is the work of Essex Orlando's Chief of Systems (M. G. Smith). The programming is in BASIC (Microsoft) and Assembly language. Evaluation has subjected microbased tests to a repeated measures analysis where they were compared to their paper-and-pencil analogues. Although there have been exceptions (and these were predictable), nearly all have exhibited strong commonalities.

System Control

The APTS has been developed to provide a human assessment capability suitable for use in remote operational settings. As presented in the system overview, the hardware, test program, and system control subsystems meet the requirements for such a system. The notebook-sized NEC PC 8201A provides the basis for an easily transportable and flexible assessment system with expansion options required for a wide range of field and laboratory applications. Additionally, the development of test programs is being conducted by a process to assure efficiency and construct validity. This process is based both on evaluation tools developed for computer tests and on lessons learned during the PETER Program (Bittner, Carter, Kennedy, Harbeson, & Krause, 1984; Smith, Krause, Kennedy, Bittner, & Harbeson, 1983). Lastly, the experimental control subsystem has been simplified for use by paraprofessionals with minimal training.

The APTS has substantial prospects for future growth and development. Attesting to this are recent and ongoing studies that have indicated that it has considerable promise for use in a broad range of unusual environments. For example, both an explosive decompression study and flight testing have found that the system is suitable for high altitude, chamber, or airborne applications. These applications, coupled with the evaluation of the design for NASA and NSF, have indicated that the system could easily be adapted for orbiting shuttle or space station use by applying spray coatings to the interior and to the exterior case. In addition, the NEC PC 8201A has demonstrated robust capabilities to operate in at least the range of 0° to 32° C, to survive drop tests, and to withstand multiple airport x ray exposures. The reliability of the system has been demonstrated during extensive field studies (>10^3 operational hours without failure).

Dr. Sam Schiflett of the USAF Aerospace Medical Research Laboratory at Brooks AFB, Texas, is using the short battery (Grammatical Reasoning, Pattern Recognition, Code Substitution, and Tapping) at two different altitudes in a parametric study. Thus far he has shown that performances are degraded in the following ways: greater deficits are seen at 25,000 ft. vs. 18,000 ft., and cognitive performances are more disrupted than motor.

Dr. Mary Williams at the University of New Orleans is studying whether learning curves of persons with identified learning disabilities reveal different slopes and different acquisition rates for different tasks, and/or whether there are relationships between initial score, rate of acquisition, and terminal score.

Dr. Darryl Mellard of the Institute for Research in Learning Disabilities at the University of Kansas, Lawrence, is working on a contract with the California Community College system to develop and establish eligibility criteria by identifying persons with learning disabilities. A sample of persons diagnosed as learning disabled are practicing 10 different tests following a similar paradigm to that being used by Dr. Mary Williams. One line of this research is the study of individual differences in the rate of acquisition of LD subjects versus normals.

Dr. Todd Jones of the US Coast Guard in Washington, D.C., is using several NEC PCs implemented with the short performance batteries to study fatigue at sea as well as the effects of ship motion on performance.

Dr. Lou Bandaret of the Army Natick Laboratories in Massachusetts is examining "Tower of Hanoi" and other complex games as tests. He is interested in performance effects of altitude and thermal stress. A study is now underway at that facility, under the direction of Dr. Charles Houston, on a protracted, simulated climb of Mt. Everest.

Dr. James May at the University of New Orleans is comparing the effects of performance of pre- and postexposure to optokinetic stimulation and pseudo-Coriolis stimulation to determine normals and persons with bilateral labyrinthine defects. He also has masking, meta contrast, and other temporally-based vision tests implemented on a NEC PC8201A, and is doing pilot work with learning disabilities.

Dr. Ann Streissguth at the University of Washington Medical Center, Seattle, is conducting a series of studies on the effects of Fetal Alcohol Syndrome on human performances. There are suggestions that recent memory loss is a key ingredient in the Fetal Alcohol Syndrome, and one test on the NEC PC8201A, the complex counting test which assesses a person's ability to keep track of several things at once with changing states, may be particularly useful.

Under contract to the Naval Training Equipment Center, Essex has tested 700-800 student pilots before and just after they were exposed to a ground-based flight trainer. In those studies > 10% reported motion sickness-like symptomatology. Performance data are being compared with subject reports.

Essex, through contracts with NASA and NSF, has evaluated the APTS and is issuing three reports. A final report of the Portable Human Assessment Battery (PHAB) has been submitted to NSF. The experimental work for these efforts has been conducted by R. L. Wilkes, at Casper College, Wyoming. For a complete reference list see Kennedy, Dunlap, Wilkes, and Lane MTA 1985.

Thus far, two studies have been completed for NASA(I and II). In the NASA I study, The proposed 6 minute (APTS) battery was administered four times along with analogous paper and pencil tests. Performance appeared stable and reliable. Two factors emerged. Reprints are available of this study and the preliminary work which preceded it. The NASA II study administered a longer battery and those data are presently being analyzed.

The study for NSF incorporated a long (10-minute) battery administered over 10 sessions and compared performance with measures of IQ (WAIS). Performance was stable for seven of the tests (vertical math and dynamic visual acuity [new tests] did not fare well). Four factors emerged. Good correlations with Performance Scale WAIS scores (multiple R = .89) were obtained; somewhat poorer with Verbal Scale scores (multiple R = .67). A preliminary draft is available.

Dr. Randall Kohl at NASA's Space Biomedical Research Institute, Houston, is undertaking a repeated measures performance testing paradigm using motion sickness drugs and provocative motion sickness tests while performance is assessed. Data collection is underway.

Dr. Pia Barth at the Fred Hutchinson Cancer Research Institute, Seattle, Washington, is administering a battery of tests from the PETER program on a NEC PC8201A to patients who have received bone marrow transplants as well as chemo radiotherapy, and to a related cohort control group over many months of replications. The project is in its 9th month and clear-cut differences in both learning and performance are evident in the treated group. Subsidiary studies are ongoing using morphine where performance decrements on different tests were shown.

Dr. Charles Wood at Louisiana State University in Shreveport, Louisiana, is studying the effects of dexedrine, hyoscine, and scopolamine on performance on the short (APTS) battery. Preliminary findings on eight subjects show performances in predicted directions with some statistically significant. Longer tests and more subjects are planned for future studies.

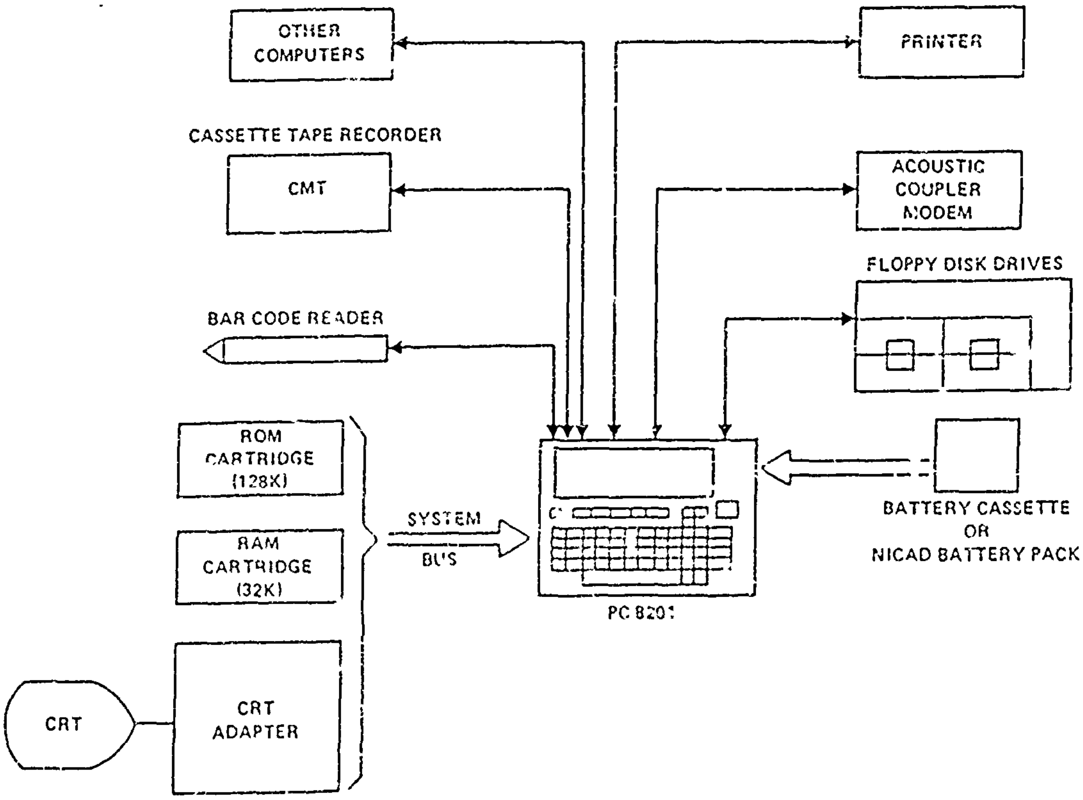
CAPT Wayne Coussens at the Tripler Hospital in Hawaii is in the process of evaluating the effects of work at altitude and mountain sickness on performance.

CDR Charles Hutchins and his students at the US Naval Postgraduate School, Monterey, California, has shown that up to 40 hours sleep loss reduces performance (p < .02) on APTS tests, particularly Code Substitution.

Dr. Michael McCauley of Monterey Technology, Inc., Monterey, California, under a US Coast Guard contract, has collected some at sea data to study workload, fatigue, and environmental stress.

The APTS has proven to be rugged and reliable, but it is the first generation of such a device. As more test are implemented on the current APTS and as new computer technologies develop an even more flexable system will emerge. A current limitation of the APTS is the Liquid Crystal Display (LCD) which is difficult to read under some lighting conditions. It is hoped that new flat panel displays will be integrated into the next generation of portable computers and hence a future generation of the APTS. Future plans have also been made to transport the APTS to the IBM PC compatable environment. This transportation will broaden the scope of the APTS applications.

Support for this project was provided by National Aeronautics and
Space Administration Contract NAS 9-16982 and the
National Science Foundation, Contract No. 00559



OVERVIEW OF PERIPHERAL INTERFACES

FIGURE 1

Applications of a Portable Training/Testing Computer

Joan Harman, Ph.D.

U.S. Army Research Institute for the Behavioral and Social Sciences

In recent years, the U.S. Army has developed and implemented increasingly complex and sophisticated weapons and communications systems. During this same period, entry-level soldiers have demonstrated declining reading, writing and computing skills. The discontinuity between increasingly demanding jobs and decreasingly skilled personnel has constituted a substantial training problem for the Army. One means of addressing this problem has been provided by the U.S. Army Research Institute (ARI) in the form of a hand-held, portable, computerized tutor that teaches technical, job-related subject matter.

Most computer-based instructional systems consist of desk top devices that are very costly and are confined to a site to which users are also confined in order to benefit from the instruction provided. The four-pound, battery-operated device called the Hand-Held Tutor, that was developed in accordance with ARI's specifications, is intended for out-of-classroom environments (mess halls, motorpools, barracks, etc.) to convert waiting periods into training opportunities. Each soldier, therefore can work with a tutor independently in a variety of settings rather than sharing a microcomputer terminal at a fixed location. The device was also required to incorporate the following features:

1- diagnostic pretests

2- self-paced instruction

3- gaming

4- instruction compatible with varying initial knowledge levels
initial motivation levels
rates of learning

5- frequent corrective feedback

The exterior features of the tutor include a 9" by 11" plastic case with an indentation molded on its top surface to hold a 5" by 5" booklet that provides instructional information and directions for interacting with the computer. Above the booklet is a multifunction liquid crystal diode display screen that includes a two digit counter and twenty-nine character space for questions, instructions, definitions and feedback. Below the booklet is a keyboard equipped with domed conductors to provide tactile feedback to the user. The keyboard displays numerals 0 through 9, letters A through E, and the words SAY, ERASE, and GO. Beside the display screen on the upper front surface of the tutor is a built-in speaker and in the rear, a jack for alternative earphones. Also on the back of the tutor casing is a jack for a battery recharger, a switch/volume control, and a receptacle for plug-in modules that encase a computer chip programmed for the Military Occupational Specialty (MOS) instruction provided in the accompanying courseware booklet. The plug-in nature of the module offers the potential to permit the essential hardware features to accommodate a great variety of MOS instruction. Selections of all features of the tutor were based on cost, availability and human factors considerations. For example, the display screen was chosen to optimize brightness, contrast ratio, size, character font and legibility within size and cost constraints. The printed courseware booklet represents an economical alternative to systems that store text and graphics in computer memory for display on a CRT.

The major considerations in courseware development included multiple teaching techniques (gaming, drill and practice, etc.) to maximize a match with individual learning styles, initial knowledge levels and rates of learning. Users can make selections from a menu of teaching/testing options that include gaming. The booklet includes many pictures and other graphic presentations and the computer provides both immediate and delayed visual and oral feedback to responses to multiple choice questions.

The courseware is divided into units that are sequenced from less to more difficult to promote an early experience of success by the user. Each unit consists of a Pretest, Explanation, Picture Battle and Word War. Users can choose any unit to work with and any component within the unit selected.

The Pretests are short tests that are intended to establish whether the user is knowledgable about the subject matter being presented. If all but one or every question is answered correctly, the final score is presented vocally and the user is permitted to move to any other component or any other unit or, if desired, to review the Pretest. If more than one answer is wrong, the user is directed to return to the first Pretest item, reviews the test with accompanying corrective feedback, and then is directed to the Explanation component in which the subject matter is taught. This component includes test questions as a check on the progress of the instruction.

The Picture Battle component requires matching pictures or graphic presentations with visual/oral stimuli. This component displays projectiles at each end of the display screen representing friendly and enemy targets. Correct responses result in movement of the friendly projectile toward the enemy target and incorrect responses result in the same kind of movement of the enemy projectile. The objective is to destroy the enemy target befo~e it reaches the friendly one. The impact wi th the enemy target is accompanied by a sound resembling an artillery shell exploding. The impact with the friendly target only results in both projectiles returning to staring positions to re-start the game.

Word War is a component that is independent of the booklet. Both questions and multiple choice answers are presented by the computer in the form of electronic flashcards on the display screen. The instructional method calls for drill and practice in an increasing ratio review format. That is, incorrect responses result in the question being presented again after one succeeding question, and once again after three additional items have been presented. Multiple choice answers to questions answered incorrectly are randomly selected from other choices stored in the tutor's chip. Also, the position of the correct answer choice is randomly varied. The success of increasing ratio review has been demonstrated to shift learned information from short to long term memory.

The tutor, therefore, incorporates varying teaching techniques, presentation modes and kinds of feedback in order to enhance acquisition and retention of the selected subject matter. The courseware is heavily weighted with frequent, short tests to permit the user to monitor progress in acquiring the needed information and to focus attention on the most relevant materials.

In 1981, ARI awarded a contract to Franklin Research Center (FRC) and its subcontractor, Educational Testing Service (ETS) for initial development of the tutor. To test the feasibility of the device, it was decided to take advantage of the existing research foundation in computer-based methods for training vocabulary.

Accordingly, a vocabulary tutor was developed to teach technical terminology to Cannon Crewmen. This tutor was evaluated at Fort Polk, Louisiana, and Fort Drum, New York. Results demonstrated substantial increases in scores on a vocabulary test, that soldiers enjoyed using the tutor and that they found it trouble-free and easy to use. Most soldiers tried all of the components in the units and most completed all of the units. There was a considerable difference between the fastest and slowest time required to complete all units, which suggests that self-pacing is an appropriate feature of the program.

In 1983 FRC and ETS were awarded a contract that called for the adaptation of the tutor to teach mathematics. The courseware was developed to address the needs of MOS 12B, Combat Engineeers. In addition, the contractor constructed a RS-232-C serial interface for the tutor. This effort provided a hardware/software data link through which the tutor can communicate with other computers. The demonstration model permits the desk top microcomputer to download course materials to the tutor, which can then be disconnected and transported to another site for study. A diagnostic feature allows for the microcomputer to upload responses to test questions, assess the needs of the user, then download appropriate homework on which the user can practice before returning for retesting.

This development greatly increases the flexibility of the tutor and provides the potential for storing instructional materials for a variety of MOS, each set of which can be transferred to the portable device as needed.

In 1984, ETS and its subcontractor, Advanced Technology Laboratories, were awarded a contract that required development of plug-in modules to accompany the mathematics courseware. This application of the tutor is under evaluation now at the Naval Ordnance Station in Indian Head, Maryland. This evaluation site is appropriate because the mathematics needs of the MOS 12B soldiers proved to be of so general a nature that a broad range of service members are expected to benefit from the instruction. In addition, the contractor is in the process of adapting the tutor to provide M1 Tank Commanders with instruction in fire commands and degraded mode gunnery. This application will be evaluated in January of 1986.

The next application of the tutor planned by ARI is for instruction in English-as-a-second language. We intend that this version incorporate a miniaturized tape recorder to simultaneously teach reading and understanding spoken English as well as pronunciation.

DESIGN OF AN OCCUPATIONAL DATA ANALYSIS SYSTEM

MAJOR C.P. WHEELER ROYAL ARMY EDUCATIONAL CORPS

ARMY SCHOOL OF TRAINING SUPPORT

INTRODUCTION

1.The British Army has for some time based its occupational data analysis programme on a version of CODAP implemented in the IBM format and currently running on an IBM 3083 mainframe computer. This version has been superseded by CODAP 80 which was considered to be the natural progression in CODAP evolution. Its proposed implementation however highlighted difficulties that would constrain the exploitation of the power of CODAP 80. It was decided therefore, that in view of the progress made by the United States Air Force Human Resources Laboratory in the development of Advanced CODAP, the time was appropriate to review the Army`s requirements and investigate the possibility of creating a dedicated occupational data analysis system.

AIM

2.The aim of this paper is to describe the analysis of the total requirement of the Army and the subsequent design and specification of a suitable system that best meets those requirements.

SYSTEM ANALYSIS

3.The decision to use a formalised system analysis technique was taken at an early stage in the investigation. The technique chosen was Learmonth and Burchett`s Structured System Analysis and Design Method, (SSADM) which is a Ministry of Defence approved system. SSADM is a data driven method which takes as input an initial statement of requirement and produces the following outputs:

a. Program specifications
b. User clerical procedures

c. Operating instructions
d. File design or data base schema
e. Plan for testing and quality assurance

4.System analysis and design is tackled as six phases with each phase broken into steps and activities. There are clearly defined interfaces between steps in the form of working documents and criteria for review and project reporting. These six phases fall nicely into two sections:

a. The What - Systems Analysis
(i). Analysis of Current System
(ii). Specification of the Required System
(iii). Selection of an Option for Implementation

b. The How - Systems Design
(i). Detailed Data Flow Design
(ii). Detailed Procedure/Processing Design
(iii). Optimisation of the Physical Design

Current System Analysis

5. The system analysis in this instance was a fairly simple process and did

not require a great depth of investigation of data flow. The current system, though incorporating an automatic data processing procedure, is fundamentally a manual one. There is no standardised questionnaire format and the design is very much a matter of study team style. The transcription of raw data from the questionnaire to a computer readable format is a data preparation bureau activity which produces a magnetic tape containing the data in ICL format. This tape is mailed to the computer centre together with a magnetic tape transcription of the job control cards and run within the CODAP shell. Results printouts are sent to the initiating agency for subsequent analysis, and, if necessary, re-runs of the data with other options from the CODAP suite of programs are completed. Finally a report is produced with appropriate recommendations. A modified data flow diagram is given at figure 1 which, for the purposes of this paper, adequately describes the current system.

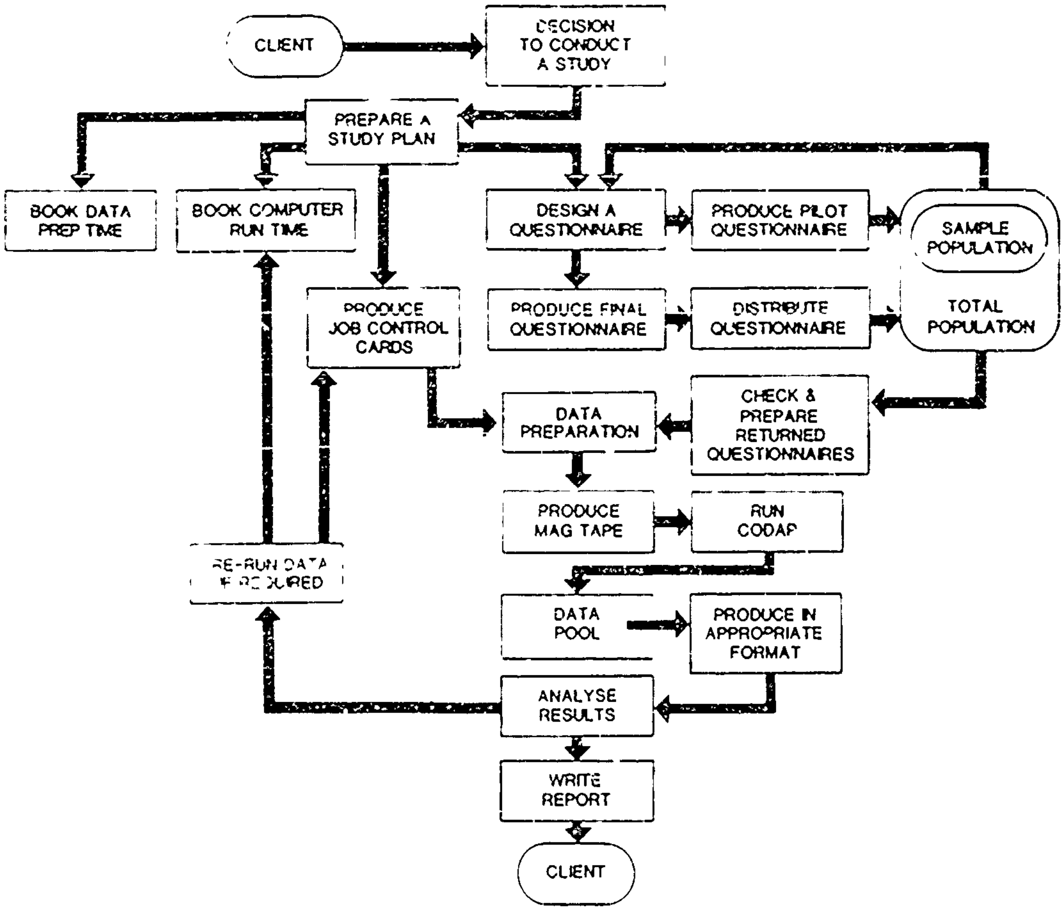


figure 1 - Current System Data Flow Diagram

The production of this document achieved two aims, firstly it ensured the analyst had a complete understanding of the system and secondly, it enabled him to identify those processes that were responsible for system operational shortfalls.

Problem Definition

6. There were two areas where problems occured, those that were inherent in the CODAP system and those that were a result of the particular implementation

at the Royal Army Pay Corps Computer Centre,(RAPC CC).They are addressed separately below.

a. Inherent Problems

(i). The version of CODAP in use is a translation of the original USAF SPERRY 1100 CODAP and reflects the state of development in 1977, (the year of acquisition). On-line data base interrogation is not possible. Confirmation of a point or the establishment of a trend requires a re-run of the raw data with fresh job control cards. This process is lengthy as fresh computer run-time slots need to be booked.

(ii). The processed data is output in the form of tables which require some expertise and patience to translate. When identifying trends this is arduous and encourages misinterpretation of the facts.

(iii). There is a limit of 999 task history and/or time data items available to each study. Though for some studies this is more than adequate in many of those proposed it may well be a constraining factor. In a recent study of Army majors' education the questionnaire had to be "sliced" to accommodate some 2200 data items from 600 respondents. This considerably added to the analysis task.

b. Implementation Problems -

(i). CODAP was designed to utilise Optical Mark Reader, (OMR) data input peripherals. The host Computer Centre did not permit the use of this facility in the original implementation for security reasons. It was feared ASTS control of the data,(and by implication the computer), would compromise the integrity of the system and its other applications programs. The result is a resource intense and time consuming data preparation process. Currently based on a bureau, the data magnetic tape is prepared in ICL format which has to be transcribed into IBM format prior to run-time. Slippages do occur in this procedure giving worst case times in the order of 25 weeks for data preparation. The best case was 3 weeks with an average of 10 weeks.

(ii). CODAP is a low priority task at the Computer Centre consequently their resource and time allocation to it is limited. This further aggravates the problem indicated above.

Specification of the Required System

7.Traditional job analyses for training have concentrated on the collection, amongst others, of information relating to task difficulty, importance and frequency, (DIF analysis). Though apparently simplistic, all of these factors, if taken in isolation are liable to produce data that will ultimately result in a wrong interpretation by the analyst and subsequent incorrect training decisions being made.

8.To seek estimates of task difficulty alone can result in invalid training decisions. Estimates can be made by the job incumbent based on differing factors, size, weight, environment, availability of spares as well as the more obvious and expected responses concerning lack of training, lack of practise or lack of ability. It is now accepted that a less ambiguous measure of difficulty is the time taken to learn to perform the task to some required level. This factor has some value in skill retention analysis.

9.Similarly estimates of importance of a task are of little value without further amplification of the frame of reference. The estimates of "importance to whom" and "importance for what" can become quite subjective and difficult to quantify. If it is importance that we wish to measure then the job incumbent's supervising officer is better placed to make that assessment. Within defence requirements, importance can be considered to be a composite of:

a. Immediacy of skill requirement after training

b. Consequence of inadequate training

10.Precise measures of frequency of performance cannot be made by job incumbents, often due to the cyclic or seasonal nature of jobs. They can, however, indicate that they spend more time on one task than another in relative terms. It is accepted that such a relative time scale can produce a more reliable measure of time spent on tasks and that subsequent calculation of percentage involvement is more accurate.

11.These factors were accepted by the Army at the time CODAP was first implemented, in reality it was a major reason for its aquisition. The situation has not changed, CODAP is still considered to be the most appropriate analytical tool available to military occupational analysts and training designers. It was decided that any future occupational data analysis system should be based on the most recent development of CODAP. Therefore system design should concentrate on improving those areas of operational shortfall identified in the analysis.

12. The most important requirement for a modern occupational data analysis system is an automated data capture facility. Data can be rapidly transcribed from standardised questionnaires and input directly to computer memory. This is an OMR operation. The ability to interrogate the data base in real-time is an advantage. The analyst should be able to manipulate data in an interactive manner using an on-line query language. Care also must be taken to ensure the system file and record lengths are of sufficient magnitude to handle the largest studies. Failing this ´slicing´ of the data should not seriously detriment the study.

13.The output of the system should be readable with an appropriate use of graphics to display data in a format such that trends are highlighted. A more thorough examination of data relevant to the problem is then possible. The total system must be able to respond rapidly, (within one week), to raw data input from respondents. It must therefore be co-located with the system managers, the system hardware being dedicated to occupational data analysis.

Possible Solutions

14. The alternatives available to ensure the continuation of an occupational data analysis service are as follows:

Option 1 - Continue with the existing system.
Option 2 - Implement CODAP 80 on the existing computer.
Option 3 - Implement a commercial package, (eg. SPSS), at the RAEC Centre.
Option 4 - Implement USAF CODAP at the RAEC Centre.

15.Option 1 - Though this is a solution option, the problems outlined previously preclude it from consideration in all but the most extreme of circumstances. The service available to the Army's Training Organisation would be subject to continual degradation leading eventually to inefficient training with all that that implies.

16.Option 2 - CODAP 80 was designed to overcome the deficiencies of CODAP. It is written in FORTRAN and well documented. There have however, been some shortfalls in the expected program run-time target efficiency. Though this may well be acceptable on a dedicated system, this is not the case with our host computer. Run-times measured in hours or even days would consign CODAP studies to those periods least used by the operational system. The outcome of this would be an even longer turn around time for processed data. It is possible that the RAPC CC would consider the task too resource demanding and not continue to host CODAP. A second and perhaps more important point is the provision of interactive facilities to ASTS. Users of CODAP 80 are able to interrogate the data base on-line. The analyst is able therefore, to manipulate the data and run the requisite program that best meets his needs from a remote terminal. To achieve this a communications link would need to be established between ASTS and RAPC CC, (a distance of some 70 miles). The link would be a British Telecom leased line, (telephone line), and would require IBM compatible terminals at each end to ensure correct communications protocol. At this time however, the RAPC CC insist on the terminal being able only to access data and not to manipulate it. The reason for this is the security of the RAPC CC computer system may be compromised and unauthorised access to other programs possible. CODAP 80 is essentially an interactive medium and without the ability to sort data and run programs from a local terminal, much of its power would be lost. Similarly CODAP 80 can accept data from OMR data input devices, in the past RAPC CC have not permitted their use.

17.Option 3 - Several commercial packages were investigated and a general comment on their suitability was that none were specifically designed for occupational data analysis in a military environment. Also as far as can ascertained from the literature they are not able to produce a job description. Commercial packages are quite expensive , typical costs over a five year period are £47750, (costs for SPSS).

18. Option 4 - This option is based on the most recent implementation of the USAF CODAP running on a SPERRY UNIVAC 1100/81 mainframe computer. A preliminary investigation into this system is encouraging, it would appear that those shortfalls of the current CODAP system have been overcome. A representative of ASTS will shortly visit the USAF HRL at Brooks AFB to confirm the suitability of it to meet the requirements of the British Army. An implementation of this version of CODAP would be hosted on a SPERRY Series 11 computer running under the SPERRY 1100 operating System and located at ASTS.

Proposed System

19.The proposed system should be based on Option 4 and should consist of USAF CODAP, (if found to be suitable), running on a SPERRY System 11 computer with automated data input peripherals, OMR, and appropria output devices, VDUs and printers. This system to be located at ASTS and managed by the Systems Group of the RAEC and be dedicated to occupational data analysis. The departure from the systems analysis routine in specifying hardware without a design phase is justified by the fact that CODAP was written for this series of computers and exploits much of its hard wired power. Any re-configuration to utilise other computers would result in an expensive and lengthy process with no guarantee of operational efficiency at the end, thus narrowing the field of choice. The OMR question has been thoroughly investigated and a suitable machine identified. This is a KAISER OMR 80 which has the ability to read both sides of a document on each pass and identify those documents that contain errors and segregate them. The OMR 80 can read up to 10,000 documents per hour.

20.The disadvantage of this proposal is its high initial cost, in the region of £250,000. The advantages however would more than compensate for this in specific cost savings of training time and in the establishment of an effective and efficient training system. This system would be able to take advantage of future development by the USAF HRL with the improvements to the service that that entails. Also the Series 11 computer would possess some spare capacity for other tasks, eg office management systems, library and course administration which will add to the general efficiency of this establishment.

Justification

21.In control systems theory, it is a fact that a system without a clear purpose and suitable controlling feed-back will eventually degenerate to chaos. The Army's Training System is based on a scientific methodology, SAT, that requires accurate training objectives derived from comprehensive job descriptions. Feed-back in the form of internal and external validation provides the control and ensures the training objectives are valid such that training courses fit the soldier for his task. The current implementation of CODAP has short comings in the capture and presentation of data. There is some doubt as to the willingness of RAPC CC to continue to host CODAP. The time is right to make good the implementation problems with a modern version based on a dedicated computer under the direct control of ASTS.

22.It is the recommendation of this paper therefore, that:

a.With the proviso above, the USAF version of CODAP be adopted for use by the British Army.

b.A suitable purchase of hardware be made that is able to run CODAP, eg a SPERRY System 11 computer, peripherals and a KAISER OMR 80.

c.A CODAP consultancy and management cell be established at ASTS to control, advise and manage the occupational data analysis requirements of the Army with a possible extension to the Royal Navy and the Royal Air Force.

IMPLEMENTATION OF AIR FORCE ASCII CODAP: NEW VERSUS OLD SYSTEM

William J. Phalen
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

Michael R. Staley
Johnny J. Weissmuller
Texas MAXIMA Corporation
San Antonio, Texas 78209

I. NEED FOR REDESIGN OF THE CODAP SYSTEM

A. BACKGROUND

1. Introduction

The principal occupational analysis technology in the Air Force is the Comprehensive Occupational Data Analysis Programs (CODAP) software system, which has supported a major occupational research program within the Air Force Human Resources Laboratory (AFHRL) and a major operational occupational analysis program within the Air Force Occupational Measurement Center (USAFOMC) since 1967. From a software standpoint, CODAP can be defined as a package of computer programs used to input, process, organize, and report occupational data from job inventories. From a measurement standpoint, CODAP can be defined as a set of procedures which focus on the analysis and comparison of individual and group job descriptions and their associated biographical data, as well as on individual tasks and groups of tasks (task modules) and their associated characteristics. From an applications standpoint, CODAP can be defined in terms of its significant contributions to the mission accomplishment of the Air Force manpower, personnel, and training management (MPT) functions. These contributions include: providing a data-based approach to evaluating and updating Air Force officer and enlisted classification structures, providing an empirical means of restructuring and redesigning jobs, providing data that has been instrumental in eliminating unnecessary training and in pinpointing specific training requirements, and providing a scientifically scund basis for realigning entry-level aptitude requi ments across career fields.

2. Problem

The CODAP system began 20 years ago as a software package of approximately 15 general-purpose programs. However, in order to keep pace with the rapidly expanding needs of occupational researchers and analysts since that time, the system was forced to expand unsystematically into a somewhat confusing aggregate of more than 60 general-purpose programs. Over time, the system became increasingly difficult to maintain, modify, or augment without extensive programmer training and experience. Many of the programs had been hastily developed to meet short suspenses, with little time available for producing proper program documentation; and many of these programs were partially redundant with already existing programs. Files maintained in the system had been put in a variety of formats and stored on a variety of media by a succession of programmers using a variety of programming standards and styles. New and advanced developments that had been needed in the system for some time required that many of the existing programs run more efficiently and interface more easily.

The CODAP system had not only grown in size and complexity since its inception, but the evolutionary process had become increasingly more dynamic as a direct result of significant hardware improvements to the AFHRL Sperry system, major enhancements to the basic structure of the CODAP software system, a growing sophistication in the methods and procedures required by Air Force occupational researchers and analysts, and a wider range of users and applications. Consequently, it became increasingly urgent that a major system redesign effort be launched to consolidate and use to better advantage the many additions and modifications that had been incorporated into CODAP over the years and to create new technology and software to meet the current and anticipated methodological and applications requirements of CODAP users.

B. OBJECTIVES

The objectives of the CODAP redesign project were to improve Air Force occupational analysis methodology, translate methodological improvements into R&D technological capabilities, standardize occupational analysis procedures, and generate efficiencies in the occupational analysis process that would significantly improve quality of product with substantially less expenditure of scientist/analyst man-hours and computer processing time. Increased data processing efficiency would also make feasible the development of complex analysis programs, especially in the area of automated job analysis, that would otherwise have been too costly to run.

C. MEANS

Early in 1983, AFHRL negotiated a task ordering contract with an 8A (small business, minority-owned) contractor, the MAXIMA Corporation. The first task submitted to MAXIMA was a 26-month project (later extended to 29 months) to rewrite/convert the CODAP system, as needed, to bring it in line with the most recent standards for software development, some of which represented requirements established by the AFHRL Technical Services Division (TS). The initial meeting with the contractor identified the following as the most urgent needs for revision of the CODAP system:

1. Converting the general-purpose CODAP programs from FIELDATA FORTRAN, which was no longer being supported by TS, to the FORTRAN 77 Standard (ASCII FORTRAN). "ASCII" is an acronym for the American Standard Code for Information Interchange.

2. Converting the utility programs from the old, TS-developed PILOT language, which was no longer being supported by Sperry, to the new, TS-developed PRISM language.

3. Converting the system files from a variety of formats to standardized mass storage files, wherever feasible.

4. Developing simplified processing procedures (runstream generators) for often-used program strings.

5. Improving formats of printed reports to enhance readability and interpretability.

6. Conducting exploratory development of new analytic capabilities, especially in the areas of profile analysis, nonhierarchical clustering,

two-way clustering (cases x tasks), module technology, and automated job typing.

At this meeting, also, major new Air Force manpower, personnel, and training (MPT) programs which would require the extensive use of occupational survey data were discussed in terms of how these projects might impact the planning of the CODAP redesign. Examples of these emerging applications included such research and development (R&D) projects as: the Training Decisions System (TDS), the Basic Cognitive Skills Project, the Advanced On-the-Job Training System (AOTS), th Task Qualification Assessment (TQA) Project, the Task Identification and Evaluation System (TIES), and various Performance Measurement projects. In addition to major programs, the new CODAP system would be required to support new operational applications for Air Force managers and decision makers. An example of such an application would be the use of CODAP-based occupational survey data to help establish test outline "testing importance" specifications for the development of enlisted promotion tests.

As a consequence of this meeting, three major project objectives were identified: increased operational efficiency, improved system maintainability, and expanded analytic capability. Each major project objective was, in turn, further broken down into technical goals, which will be presented and discussed in the next section, no longer as goals, but as redesign accomplishments. These accomplishments show how the redesigned CODAP system (ASCII CODAP) differs from and is an improvement upon the old CODAP system (FIELDATA CODAP).

II. IMPLEMENTATION OF CODAP REDESIGN: NEW VERSUS OLD SYSTEM

As compared to the old FIELDATA CODAP system, the new ASCII CODAP system has achieved the three major project objectives listed in the previous section, as follows:

A. <u>In ASCII CODAP, Operational Efficiency Has Been Increased.</u>

1. <u>Run setups have been simplified.</u>
   - Functional program names used
   - Mnemonics employed to name control card options
   - Control cards standardized
   - Program execution by processor control cards
     - increased flexibility in filename specification
     - elimination of separate filename cards

2. <u>Audit trails have been improved.</u> The first page of every report contains information such as the report ID, when the report was created, the IDs of all input files, options selected, and applicable selection or cutoff parameters for cases and/or tasks.

3. <u>Resource utilization has been improved.</u>
   - Reduced core requirements for many programs
   - Reduced running times for many programs (e.g.. job description program runs 10 times faster)
   - Elimination of three-reel files

4. Turnaround has been speeded up.
   - Reduced core requirements and running times of many frequently run, multiple-execution programs, as well as the elimination of three-reel files, has permitted daytime running of these programs.
   - Conversion of files to mass storage format has speeded up processing and allows some programs which were previously run only in batch mode to be run in demand mode.

5. Fewer runs are required to accomplish a standard analysis.
   - Elimination of redundant programs
   - Combining of programs which perform related functions
   - Development of functionally pure programs (e.g., sample selection extracted from multiple programs and consolidated into a single program)

6. Training requirements for computer technicians have been simplified.
   - Development of computer-based training package (in progress)
   - Better documentation
   - Reduced number of programs
   - Functionally pure programs
   - Mnemonic control card options
   - Visible cue report format specifications
   - Standardized control cards
   - Automatic process generation
   - Separ te report file process

B. In ASCII CODAP, system maintainability has been improved.

1. Structured programming used.
2. Conversion to FORTRAN 77 Standard (ASCII FORTRAN) and PRISM.
3. Fewer programs (reduced from 120 to 83).
4. Functionally discrete programs.
5. Reduced amount of source code.
   - Assembly language code reduced
   - Total lines of code reduced (ASCII: 27,186 vs. FIELDATA: 48,477)
     - elimination of redundant code
     - interfacing with non-CODAP software rather than retaining similar CODAP programs

6. Increased internal documentation (43.4% more lines of comments).

7. Increased supporting documentation, which is all in automated form.
   - Users manual
   - Standards documentation
   - Programmer reference guide
   - Subroutine documentation
   - Subprogram documentation
   - File format documentation

8. System has been standardized, wherever feasible.
   - Program code
   - Program names
   - Program documentation
   - File formats
   - Subroutin library

9. Formalized test and acceptance procedures to ensure reliability.

C. In ASCII CODAP, analytic capability has been expanded.

1. Increased system limits.
   - From 1,700 to 3,000 tasks
   - From 1,000 to 9,999 task modules
   - From 1,726 to 3,000 tasks per module
   - From 999 to 2,000 history variables
   - From 500 to 9,999 computed variables

   Unchanged system limits.
   - 20.000 cases total
   - 7,000 cases for clustering
   - 26 duties
   - 66 characters for variable description

2. New features and capabilities.
   - Clustering technology expanded
     - additional overlap indices incorporated (four new indices)
     - task clustering capability added
     - nonhierarchical clustering capability added
     - profile analysis capability added
   - Module technology expanded
     - tasks can be summarized into modules
     - modules can be used like tasks
     - module values can be clustered
     - module data collected from job incumbents/expert raters can be processed directly
   - Job typing technology expanded
     - pairwise group comparison program (AUTOJT) capabilities expanded
     - core task analysis program (CORSET) capabilities expanded
     - program developed which makes initial selection of job types automatically (JOBSET)
   - Interrater reliability technology expanded

3. Improved final products.
   - More readable and interpretable report formats
   - Greater availability and more flexible use of free text
   - Use of upper- and lower-case letters and overstriking
   - Greater standardization of formatting for reporting similar kinds of data

4. External system interfaces to non-CODAP software.
   - Data distribution routines
   - Correlation and regression software
   - Factor analysis software
   - Other statistical packages, such as SPSS, BMDP and IMSL

III. ISSUES RELATED TO CODAP TECHNOLOGY TRANSFER

A. Transfer of Air Force ASCII CODAP

Air Force Regulation (AFR) 300-6 governs the release of Air Force-owned or developed computer software package. It states in para 11-71(2): "Computer programs and related technical data are not considered 'records' within the congressional intent of 5 U.S.C. 552; these items are considered property." It further states in para 11-7b(1): "Software packages will not be released to the private sector except when in the best interest of the government." In the event that Air Force software, such as ASCII CODAP, is released to a private sector firm, the requester must certify that the package will not be published for profit or in any manner offered for sale to the government and will not be sold or given to any other activity or firm without the prior written approval of the Air Force. On the other hand, AFR 300-6 permits the release of ASCII CODAP to the public sector, such as government agencies and universities, provided that the recipient signs a statement of terms and conditions which frees the Air Force of any liability and or responsibility for the software and requires the recipient to state the intended use of the software.

b. Cost of Compatible Hardware Versus Cost of Software Conversion

Sperry representatives were asked to speculate as to what would be the necessary configuration for a minimal Sperry system to meet the requirement that there would be 5 to 10 users on a system dedicated to running CODAP. It was agreed that a minimally adequate system would consist of a Sperry System 11 or Sperry 1100/60 or Sperry 1100/61 with a Mega word of main memory, an 8470 disk system (and controller) containing 50K tracks (89 million words), three 1600 BPI tape drives with tape controller, console, upper-lower case printer, and a communications subsystem. Such a system would cost approximately $250,000. ASCII CODAP could be transferred to such a system at little cost and be up and running almost immediately. On the other hand, the contractors who produced ASCII CODAP estimated that it may cost as much as $250,000 and require one year for two senior systems analysts and two programmers conversant with CODAP and familiar with both the Sperry system and the non-Sperry host system to accomplish the conversion of ASCII CODAP for running on the non-Sperry system. Given these two alternatives, the choice of which way to go should be an easy one. Nevertheless, justifying the purchase of another computer system when your organization already has one is usually more difficult than getting authorization and funds to do a software conversion.

IV. CONCLUSION

The new Air Force ASCII CODAP system is now undergoing final test and acceptance by the AFHRL Technical Services Division (AFHRL/TS). The process is moving along smoothly and the new system should be ready for release to the USAF Occupational Measurement Center (USAFOMC) and other Sperry/CODAP users by 1 January 1986.

REFERENCES

AFR 300-6. (1980, 11 July). Automatic Data Processing Resource (ADPR) Management.

Phalen, W.J., Weissmuller, J.J., & Staley, M.R. (1985, May). Advanced CODAP: New analysis capabilities. Fifth International Occupational Analysts Workshop, San Antonio, TX.

Staley, M.R., Weissmuller, J.J., & Phalen, W.J. (1985, May). ASCII CODAP: The impact of system design for emerging applications. Fifth International Occupational Analysts Workshop, San Antonio, TX.

Weissmuller, J.J., Staley, M.R., & Phalen, W.J. (1985, May). ASCII CODAP: Quarterly status report. Fifth International Occupational Analysts Workshop, San Antonio, TX.

Occupational Research Data Bank
(ORDB)

Martin E. Ellingsworth

Air Force Human Resources Laboratory
Brooks AFB, Texas 78235-5601

Louis F. Olivier
Glenda J. Pfeiffer

OAO Corporation
Brooks AFB, Texas 78235-5601

INTRODUCTION

The Occupational Research Data Bank (ORDB) is a computer-based occupational information system. It provides researchers and managers a means of rapid on-line retrieval of a variety of current and historical occupational information on Air Force enlisted specialties and the people performing duty in them through a set of user-friendly, tutorial programs. Use of ORDB streamlines background research, permits quick in-depth orientation to specialties, and provides a database for a rapid response capability for research and management concerns. The ORDB is of great value as a tool to increase productivity and as an instrument for longitudinal and cross-specialty analyses.

Development of the ORDB began in 1978 with the investigation of available data that would contribute to Air Force occupational analysis and management. A number of sources and types of information were identified and have been obtained for inclusion in the ORDB (Carpenter, Archer, & Camp, 1979; Stephenson, 1979; Camp, 1982). The primary contractor for this effort is the General Services Administration data processing services contractor, currently OAO Corporation. OAO personnel assigned to this project are collocated with the monitoring activity (AFHRL/MOMM) at Brooks AFB, Texas.

The ORDB is composed of both digital and hard copy data consisting of technical reports and studies, Air Force Regulation 39-1 information by career area, statistical variables summarized for occupations from individual Air Force members and technical training course data, and Comprehensive Occupational Data Analysis Programs (CODAP) (Christal, 1974) studies performed at the Air Force Occupational Measurement Center (OMC). These types of information have been obtained and are incorporated in the ORDB. The subsystems which provide for storage and on-line retrieval of the information are described in the following section. Before going into the ORDB subsystems, here are two notes of interest:

1. When referencing an Air Force Specialty Code (AFSC), there are three levels of detail that can be shown using the five numbers of the AFSC. Here are the types, abbreviations, and examples of these three levels.

Skill Level	(S)	e.g., 27230, 44550G (skill levels 1, 3, 5, 7, 9, & 0)
Ladder	(L)	e.g., 973X2, 113X0C
Career	(C)	e.g., 61XXX, 42XXX

2. References to the figures at the end of this paper are actual slides from the elaberative case study scenario of the 426X2 (Jet Engine Mechanic) ladder AFSC, which was the major part of the briefing illustrating how an occupational researcher could benefit from using the ORDB. Figures 1 and 2 are actual output from the ORDB.

SYSTEM OVERVIEW

The ORDB operates on the AFHRL Sperry 1100/81. Seven subsystems are tailored to the types of data and kinds of retrieval needed by the user. These subsystems are linked together by a front-end program to simplify the use of the ORDB. The programs are designed to interact with the user, assisting in the choice of the appropriate subsystem, and in selecting the desired information. Each subsystem is described below.

1. Computer Assisted Reference Locator (CARL). The CARL subsystem is used to reference hard copy occupational data items, such as recurring reports, occupational survey reports, job inventories, films, and microfiche, which are stored at the Air Force Human Resources Laboratory, Brooks AFB, Texas. References are based on user selected keywords. CARL was obtained from the Navy Personnel Research and Development Center (NPRDC) and modified to operate on the 1100/81 (Sands, 1978; Sands & Hartman, 1979). Additional modifications were made to accept AFSCs as keywords and to clarify user selection of output options.

Each reference stored in the CARL subsystem includes such information as author, name or title of the reference, type of reference, a brief narrative description, and an associated list of keywords for each reference (see Figure 1).

To speed the referencing process, two search techniques have been added to CARL--Quick and Smart. Quick is a binary search on a given list of keywords available upon request, while Smart is a character string search across the list of available keywords. Both respond with the number of references located and ask if and how the user would like to see output. In addition, the user can expand or reduce the number of references by using additional keywords or character strings.

2. Aptitude Requirements Component (ARC). The ARC subsystem contains AFSC descriptions (for ladder and career field), progression ladders, and prerequisite data for the years 1978 to the present (1985). The ARC has an AFSC number change history file which tracks all changes from March 1965 through the present. In addition, aptitude requirements information for each AFSC is stored and accessible (Figure 2). Also, the ARC subsystem contains KINTON (a contractor) and OMC study information including Average Task Difficulty Per Unit Time Spent (ATDPUTS), validity/reliability information, statistics, and minimum/maximum task information. It should be noted that the KINTON reports (approximately 200) were produced in the Aptitude Requirements benchmarking research, while OMC studies are completed with each occupational survey of an AFSC (Kinton uses a 25-point scale while OMC uses a 9-point scale for ATDPUTS).

Since the ORDB is a menu-oriented, tutorial system, access and use of this and the following subsystems are quite easy and efficient. You only really need to have an idea about the AFSC structures in the Air Force. This structuring is readily apparent in the use of the statistics subsystem.

3. Statistical Variable. The statistics subsystem contains demographic, aptitude, education, training, turnover, and duty related information on Air Force enlisted personnel. Information is sorted by AFSC, population group, and year or a total of 125 different variables for the most current 5 years of data. Population group is based on enlistment status by Total Active Federal Military Service (TAFMS) (0-4 years, 5-8 years, 8+ years, total sample, or current year's accessions). See Figure 3 for a partial list of variables and their level of detail (L, C, or S). An AFSC must be valid for the type of detail available and it must have existed at the end of the year for which data are being requested (validity of an AFSC can be checked in the ARC subsystem, and the validity of detail can be checked on the menu available in the statistics subsystem).

While 5 years of data are stored on-line, earlier years of statistical data will be accessible via batch run. The sources are the Uniform Airman Record (UAR), Pipeline Management System (PMS), Airmen Gain and Loss (AGL), and Processing and Classification of Enlistees (PACE) files stored at AFHRL.

This subsystem uses System 2000 (S2K) Data Base Management System with COBOL extension (Intel Corporation, 1982).

4. CODAP Report Display. This subsystem was developed to provide the task scientist and manager with the ability to rapidly retrieve OMC and AFHRL CODAP reports and review them on the terminal screen. To accommodate the standard CODAP report format, Datagraphix 132 character remote terminals are in use at principal user sites. Studies can be selected by either AFSC or study number (from a menu of available studies). Studies from 1978 to the present have been loaded, and any report retrieved on the screen can also be printed at the user's option. This subsystem is programmed in the Programming Instructions for String Manipulation (PRISM) language (AFHRL/TS, 1982).

From the initial listing of CODAP reports contained in the OMC studies, a determination is made as to which CODAP reports will be loaded into the ORDB (Figure 4). The determination is based on the requirement of a report that concerns a population group and an AFSC in a similar manner to the statistics subsystem organization.

When the user selects this subsystem from the ORDB introductory screen, he or she is provided a choice of the five different CODAP retrieval features.

a. CODAP report display--The user can view the text of individual reports and obtain a hard copy of desired reports.

b. CODAP edit feature--The user can use editing commands to select lines from one or more reports from one or more studies and sort them into a customized hard copy output.

c. Task-level cross-study--The system will retrieve tasks containing key words from multiple studies and print hardcopy, if the user desires.

d. Background cross-study analysis--The system will retrieve 15 background variables from multiple studies and print hard copy if the user desires.

e. Title deck Specialty Training Standard (STS) items--The system will create a file containing STS items for an AFSC in the format of a title deck.

5. Cross-Study Analysis. This subsystem was developed in response to the need to compare CODAP reports across specialties. Since CODAP variable numbers and titles are not necessarily standard, identifying corresponding data in different studies presents a difficult task. To solve this problem, studies are indexed as they are loaded to the ORDB for a set of 15 variables and 8 groups. The variables include: Number of Tasks, ATDPUTS, Job Difficulty Index, Grade, Major Command, Time in Career Field (TICF), TAFMS, Eligible to Reenlist, Eligible for Retirement, Job Interest, Talent Utilization, Training Utilization, Sense of Accomplishment, Plan to Reenlist, and How Assigned to Present Career Field

Groups that can be analyzed include: Total Sample, Skill Levels 3, 5, 7, 9, 1-48 months TAFMS or TICF, 49-96 Months TAFMS or TICF, and 97+ Months TAFMS or TICF. On-line retrieval of corresponding data from multiple studies on one or a number of job groups can be performed using this system (Figure 5). For example, job difficulty of airmen with 1-48 months TAFMS can be retrieved for comparison across any number of AFSCs. This subsystem is written in PRISM and uses the same data files as the CODAP Report Display subsystem.

6. Statistical Package for the Social Sciences (SPSS)/Statistical Interface. A total of four SPSS procedures are interfaced with the Statistical Variable Subsystem of the ORDB: ANOVA, BREAKDOWN, T-TEST, and CROSSTABS (Figure 6). ORDB allows the user to produce statistical analyses of ORDB variables using SPSS without requiring him to be familiar with formatting SPSS run cards. The interface program provides easy to follow instructions for user inputs.

The user may initiate a batch run which will automatically retrieve the ORDB statistics and create a runstream of SPSS control cards. This will result in an SPSS run with output. The user has the choice, however, of running the SPSS automatically or having the file containing the runstream retained for additional modification.

7. Comments. The comments subsystem provides an opportunity for users and developers to record information related to the ORDB while using a remote terminal. Comments can include anything relevant to data contained in the system, or to the system operation itself. It has been especially useful as a means of obtaining user feedback and for announcing the implementation of enhancements or changes.

CONCLUSION

ORDB is currently used by many of the major projects at AFHRL (e.g., Training Decision Systems, Basic Skills, Aptitude Requirements, Performance Measurement, and RIVET Work Force), and it is planned for use with Advanced On-the-Job Training). At OMC, ORDB is used to provide quick in-depth orientations to AFSCs as well as to rapidly respond to high level management queries.

ORDB relates many dispersed sets of data into a consolidated data bank which can be rapidly accessed. Instead of the normal laborious and time-consuming task of finding background information by formal requests to computer databases, searching Air Force regulations, and/or digging through a library of technical reports and previous studies, the ORDB allows the user to streamline data retrieval while saving computer resources. ORDB is valuable for aiding research design, for conducting historical and cross-specialty analyses, and for guarding against duplication of effort and inconsistencies between databases. Clearly, ORDB enhances researcher and management productivity.

References

AFHRL [illegible]. (198[illegible]. Programming instructions for string manipulation (PRISM). Developed and maintained by Air Force Human Resources Laboratory, Technical Services Division, Brooks AFB, TX.

Camp, R.L. (1982, October). Implementation of an Air Force occupational research data bank. Proceedings of the 24th Annual Conference of the Military Testing Association. San Antonio, TX.

Carpenter, J.B., Archer, W.B., & Camp, R.L. (1979, October). Establishing an Air Force occupational research data bank. Proceedings of the 21st Annual Conference of the Military Testing Association San Diego, CA.

Christal, R.E. The United States Air Force occupational research project. (1974, January). AFHRL-TR-73-75, AD-774-574. Occupational Research Division, Air Force Human Resources Laboratory, Lackland AFB, TX.

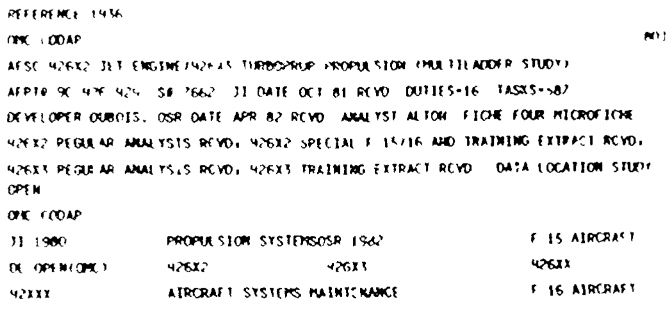
Intel Corporation. (1982). System 2000 (S2K) data base management system. Intel Corporation, P.O. Box 9968, Austin, TX.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., & Bent, D.H (1975). Statistical package for the Social Sciences (SPSS) (2nd ed.). McGraw-Hill, St. Louis.

Sands, W.A. (1978, October). Computer assisted reference locator (CARL) system: An overview. Proceedings of the 20th Annual Conference of the Military Testing Association, Oklahoma City, OK.

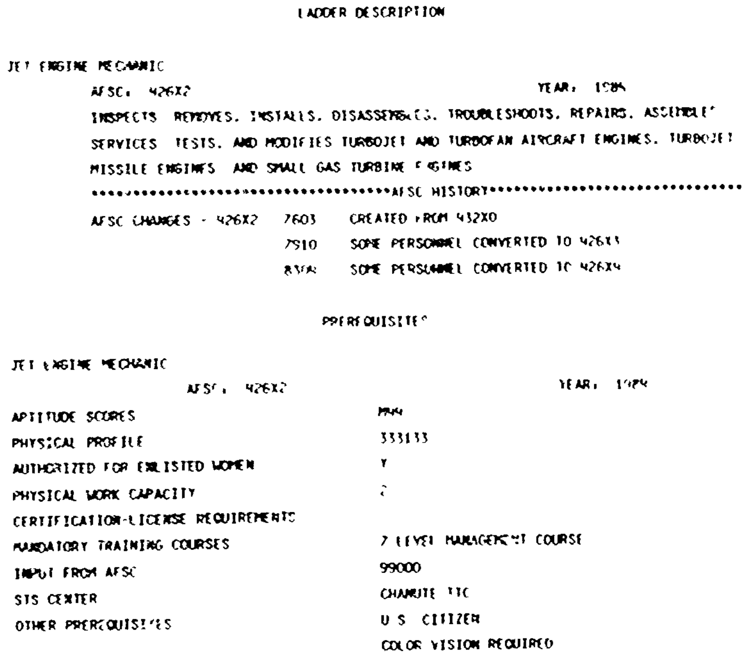
Sands, W.A., & Hartman, L. (1979, October). Research management applications of the computer assisted reference locator (CARL) system. Proceedings of the 21st Annual Conference of the Military Testing Association, San Diego, CA.

Stephenson, R.W. (1979, October). Development plans for an Air Force occupational research data bank. Proceedings of the 21st Annual Conference of the Military Testing Association, San Diego, CA.



```
REFERENCE 1936
OMC CODAP                                                          NO)
AFSC 426X2 JET ENGINE/426X3 TURBOPROP PROPULSION (MULTILADDER STUDY)
AFPT0 9C 42E 429  S# 7662  JI DATE OCT 81 RCVD  DUTIES=16  TASKS=587
DEVELOPER DUBOIS. OSR DATE APR 82 RCVD  ANALYST ALTON  FICHE FOUR MICROFICHE
426X2 REGULAR ANALYSIS RCVD, 426X2 SPECIAL F 15/16 AND TRAINING EXTRACT RCVD,
426X3 REGULAR ANALYSIS RCVD, 426X3 TRAINING EXTRACT RCVD  DATA LOCATION STUDY
OPEN
OMC CODAP
JI 1980            PROPULSION SYSTEMSOSR 1982                  F 15 AIRCRAFT
DL OPEN(OMC)       426X2              426X3                    426XX
42XXX              AIRCRAFT SYSTEMS MAINTENANCE                F 16 AIRCRAFT
```

FIGURE 1



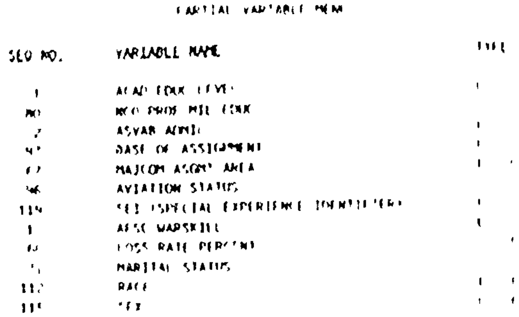
LADDER DESCRIPTION

```
JET ENGINE MECHANIC
          AFSC, 426X2                                    YEAR, 1985
          INSPECTS  REMOVES, INSTALLS, DISASSEMBLES, TROUBLESHOOTS, REPAIRS, ASSEMBLES
          SERVICES  TESTS, AND MODIFIES TURBOJET AND TURBOFAN AIRCRAFT ENGINES, TURBOJET
          MISSILE ENGINES  AND SMALL GAS TURBINE ENGINES
          ******************************AFSC HISTORY******************************
          AFSC CHANGES - 426X2   7603   CREATED FROM 432X0
                                 7910   SOME PERSONNEL CONVERTED TO 426X3
                                 8306   SOME PERSONNEL CONVERTED TO 426X4
```

PREREQUISITES

```
JET ENGINE MECHANIC
                AFSC, 426X2                                YEAR, 1985
APTITUDE SCORES                         M44
PHYSICAL PROFILE                        333133
AUTHORIZED FOR ENLISTED WOMEN           Y
PHYSICAL WORK CAPACITY                  2
CERTIFICATION-LICENSE REQUIREMENTS
MANDATORY TRAINING COURSES              7 LEVEL MANAGEMENT COURSE
INPUT FROM AFSC                         99000
STS CENTER                              CHANUTE TTC
OTHER PREREQUISITES                     U S  CITIZEN
                                        COLOR VISION REQUIRED
```

FIGURE 2

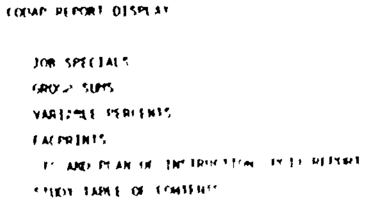


PARTIAL VARIABLE MENU

```
SEQ NO.     VARIABLE NAME                              TYPE

   1        ACAD EDUC LEVEL                             I
  NO        NCO PROF MIL EDUC
   2        ASVAB ADMIN                                 I
  47        BASE OF ASSIGNMENT                          I
  62        MAJCOM ASGMT AREA                           I    I
  96        AVIATION STATUS
 119        SEI (SPECIAL EXPERIENCE IDENTIFIER)         I
   1        AFSC MAPSKILL                               I
  [illegible]  LOSS RATE PERCENT                             I
  [illegible]  MARITAL STATUS
 112        RACE                                        I    I
 115        SEX                                         I    I
```

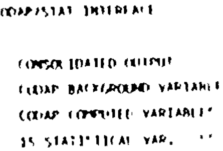
FIGURE 3



CODAP REPORT DISPLAY

```
   JOB SPECIALS
   GROUP SUMS
   VARIABLE PERCENTS
   FACPRINTS
    [illegible] AND PLAN OF INSTRUCTION  [illegible] REPORT
   STUDY TABLE OF CONTENTS
```

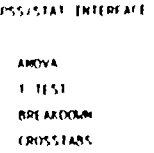
FIGURE 4



CODAP/STAT INTERFACE

```
   CONSOLIDATED OUTPUT
   CODAP BACKGROUND VARIABLES
   CODAP COMPUTED VARIABLES
   15 STATISTICAL VAR.
```

FIGURE 5



SPSS/STAT INTERFACE

```
   ANOVA
   T TEST
   BREAKDOWN
   CROSSTABS
```

FIGURE 6

Effects of Absenteeism on the Performance of Air National Guard Personnel

Robert P. Steel and Guy S. Shane
Air Force Institute of Technology

Recent reviews of the absence literature give substance to claims that considerable research on absenteeism has been performed (Mowday, Porter, & Steers, 1982; Muchinsky, 1977). This considerable body of research has often yielded mixed findings and uncertain conclusions (e.g., Muchinsky, 1977), although some real progress has been made in evaluating traditional common-sense absence frameworks (Scott & Taylor, 1985). For example, Scott and Taylor's meta-analysis has brought some reliable support to the long-standing faith of researchers in job satisfaction - absence relationships.

Traditionally, researchers have regarded absenteeism as a dysfunctional response emitted by an employee attempting to withdraw from an unpleasant work experience. However, this view may be too simplistic; Mowday et al. (1982) note that both **functional** and **dysfunctional** outcomes may be associated with employee absenteeism. However, cases of absence abuse are still considered by most theorists to be primarily dysfunctional behavioral patterns, and their effective control is assumed to pay significant dividends to the organization.

Absence control policies have long been justified on the grounds that absence abuse contributes to labor costs, and it has also been linked conceptually to lower organizational and individual productivity. While evidence of the labor costs associated with absenteeism continues to mount (cf. Mowday et al., 1982), little systematic thought has been given to the purported absence - performance linkage (Mowday et al., 1982). This relationship appears to have been taken for granted.

Common-sense beliefs about absence correlates have increasingly been questioned by researchers and theorists (e.g., Scott & Taylor, 1985). Likewise, the assumed relationship between absenteeism and declining productivity should be subjected to empirical review. Although occasional investigations (Keller, 1984; Sheridan, 1985) have reported incidental correlations between absenteeism and job performance measures, this relationship has not been systematically studied. The present study reports correlations between absenteeism and job performance ratings.

Very little published research has been performed examining sick leave usage rates in the federal sector. This is perhaps unfortunate given the rather liberal sick leave policies currently in existence in the federal government. Leave administration for Department of the Air Force civilian personnel is defined in Air Force Regulation 40-631 (USAF, 1971), and these policies are uniform throughout the entire federal service. The policy authorizes 13 days of sick leave per year regardless of seniority. There is no absolute accumulation limit on the amount of sick leave employees may accrue, and they may "bank" accrued sick leave for other purposes. Accumulated sick leave may be used to retire early without reducing service time annuity payment calculations or to increase the service time used in calculating retirement and survivor annuity payments. In effect, this policy eliminates the "use or loose" incentive contained in many sick leave plans.

Several studies have established links between absence use patterns and organizational policy (Dalton & Perry, 1981, Larson & Fukami, 1985). Policy in force in the federal sector will undoubtedly impact the way sick leave is used, and furthermore, it may impact the way in which absence relates to other organizational outcomes, notably job performance. A second major focus of the study was on federal sector absence - performance relationships.

Method

Sample

Data for the study were provided by 134 full-time employees of an Air National Guard base in the western United States. The typical respondent was male (93%) and between 31 and 40 years old.

Sick Leave

Absenteeism in the present study was operationalized in terms of authorized sick leave usage. The organization provided sick leave data for all employees covering the period January 1, 1983 to December 31, 1983. Absence records for each employee were organized into 26 biweekly pay periods.

Considerable debate has occurred on the selection of an absence metric. Most authors regard frequency indices as psychometrically superior to time lost measures (Chadwick-Jones, Brown, Nicholson, & Sheppard, 1971; Hammer & Landau, 1981; Muchinsky, 1977). A frequency index expresses absence in terms of the number of incidents of absence, regardless of their individual durations. In contrast time lost indices express absence in terms of the amount of time lost (i.e., hours or days lost). Given the impact that choice of index may have on study outcomes, the present investigation employed both a time lost index and a frequency index.

As the present data were collapsed into 26 units of information, this organization imposed some limitation on development of a frequency index. Therefore, frequency represented the number of pay periods in which an absence event occurred, regardless of duration (possible range 0 - 26). Time lost was calculated as the total number of hours lost during the entire year.

Performance Appraisals

Two sources of employee performance appraisals were utilized. Self-appraisals were embedded within a larger survey questionnaire. The self-appraisal method was based on a technique of self-appraisal developed by Steel and Ovalle (1984) called Feedback Based Self-Appraisal. This method is comparable to traditional self-appraisals in most respects except that employees are instructed to base their ratings on feedback they have received from their immediate supervisors. Steel and Ovalle (1984) found Feedback Based Self-Appraisals to be more highly related to supervisory evaluations than were conventional self-ratings. Evaluations were made on five performance dimensions: quantity, quality, efficiency, problem-solving capacity, and dependability. Seven point agree - disagree rating scales were used.

Supervisory appraisals were solicited from each employee's immediate supervisor These ratings were made on the same dimensions as the self-ratings. In addition, an overall rating of employee effectiveness was obtained. The supervisory ratings were distributed on 7-point scales ranging from (1) "this employee is *far worse* than the typical employee" to (7) "this employee is *far better* than the typical employee."

Procedure

Self- and supervisory appraisals were collected during the summer of 1984. The data were collected in conjunction with a larger research project, and participants were told that their evaluations would be used exclusively for research purposes only. The questionnaire containing the self-evaluations was administered on-site in small group meetings. Respondents were notified that their responses were confidential and anonymous, and they were informed about possible uses to which the data might be put. Social security numbers were requested from participants, and they were used to link appraisal and absence data on a case-by-case basis.

Results

Pearson rs were calculated between the performance appraisal measures and the two absenteeism metrics, time lost and frequency. These correlations are displayed in Table 1. The correlations tended to be small and nonsignificant. Most of them were negative as we anticipated.

No significant correlations were obtained between the self-appraisals and either absence index. These correlations were uniformly quite small.

Correlations between the supervisory ratings of job performance and the two absence metrics produced inconsistent results. The

Table 1

Absence - Performance Correlations

Performance Dimension	Time Lost	Frequency
	Self-Appraisal	
Quantity	-.07	-.06
Quality	-.03	-.04
Efficiency	-.05	-.03
Problem-solving	-.01	.00
Adaptability	-.04	-.03
	Supervisory Appraisal	
Quantity	-.12	-.06
Quality	.00	-.06
Efficiency	-.15*	-.10
Problem-solving	-.10	-.09
Adaptability	-.19*	-.12
Overall rating	-.10	-.08

Note. Self-appraisal n = 71; supervisory appraisal n = 153.

* $p < .05$

performance - absence frequency correlations were all nonsignificant. However, correlations between the time lost absence measure and two performance dimensions, efficiency and adaptability were statistically significant ($p < .05$). Given the general unreliability and insensitivity usually attributed to performance appraisal measures (e.g., Schmidt, Gast-Rosenberg, & Hunter, 1980) and the unreliability of time lost absence measures (Muchinsky, 1977), the magnitude of these significant correlations probably underestimates the true relationship between absence and performance. These data may reasonably be construed as indicating that excessive sick leave usage is likely to have detrimental effects on the work performance of federal employees.

Discussion

Although at first blush the significant correlations appear to be small and potentially trivial, we believe they probably underestimate true performance - absence relationships. We believe the cards were stacked against finding significant correlations. The purported reliability of both variables is low, and as Hunter, Schmidt, and Jackson (1982) have stridently pointed out, such artifacts as error of measurement contribute substantially to the apparent magnitude of bivariate relationships. Furthermore, we believe the generous sick leave accumulation policy in force in the federal sector removes the "use or loose" incentives associated with most sick leave plans. This policy should have the effect of reducing absence criterion variance because less total sick leave will be taken. Hence, we believe the results give genuine support to the long-standing belief among personnel managers that absence prone employees are also poor performers. However, the correlational nature of the data prohibit any causal inferences. It is impossible to determine whether the performance of sick leave abusers is rated low because they are absent so often (i.e., performance judgments are influenced by general negative impressions) or because their performance genuinely suffers from so much lost work time.

The study yielded a couple of surprises. Many researchers have observed that frequency absence metrics are more reliable than time lost absence scales. The degree of reliability of an absence measure should have a profound effect on the ability of that metric to correlate with other variables. This fact was clearly manifest in Scott and Taylor's (1985) research. However, the present study failed to obtain significant correlations between performance and absence frequency, although significant rs were obtained for the normally inferior time lost index. We suspect that collapsing the data into 26 pay periods reduced the total amount of variance obtainable with the frequency metric, and perhaps this constraint on the potential variance produced the poor results with this absence index.

References

Chadwick-Jones, J. K., Brown, C. A., Nicholson, N., & Sheppard, C. (1971). Absence measures: Their reliability and stability in an industrial setting. *Personnel Psychology*, *24*, 463-470.

Dalton, D. R., & Perry, J. L. (1981). Absenteeism and the collective bargaining agreement: An empirical test. *Academy of Management Journal*, *24*, 425-431.

Hammer, T. H., & Landau, J. (1981). Methodological issues in the use of absence data. *Journal of Applied Psychology*, *66*, 574-581.

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

Keller, R. T. (1984). The role of performance and absenteeism in the prediction of turnover. *Academy of Management Journal*, *27*, 176-183.

Larson, E. W., & Fukami, C. V. (1985). Employee absenteeism: The role of ease of movement. *Academy of Management Journal*, *28*, 464-471.

Mowday, R. T., Porter, L. W., & Steers, R. M. (1982). *Employee-organization linkages: The psychology of commitment, absenteeism, and turnover.* New York: Academic Press.

Muchinsky, P. A. (1977). Employee absenteeism: A review of the literature. *Journal of Vocational Behavior*, *10*, 316-340.

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology*, *65*, 643-661.

Scott, K. D., & Taylor, G. S. (1985). An examination of conflicting findings on the relationship between job satisfaction and absenteeism: A meta-analysis. *Academy of Management Journal*, *28*, 599-612.

Sheridan, J. E. (1985). A catastrophe model of employee withdrawal leading to low job performance, high absenteeism, and job turnover during the first year of employment. *Academy of Management Journal*, *28*, 88-109.

Steel, R. P., & Ovalle, N. K. 2d. (1984). Self-appraisal based upon supervisory feedback. *Personnel Psychology*, *37*, 667-685.

United States Air Force. (1971). Civilian personnel: Leave administration. AFR 40-630. Washington: HQ USAF.

IN SEARCH OF HONESTY

Allen N. Shub, Ph.D.
Department of Management
Northeastern Illinois University, Chicago, Illinois

In the 4th Century B.C., legend has it that a philosopher named Diogenes wandered about the Athenian countryside in broad daylight with a lighted lantern clutched in his outstretched hand. He was searching for an honest man.

Now, more than 2,300 years later, the search continues as governmental agencies and private industry look for honest employees. In most employment settings today the demand for help is great and often continuous, especially where lower wages have effected maximum turnover and lower reenlistment, where responsibilities are great and temptations are even greater, and where an indifferent society has spawned an atmosphere more prone to dishonesty and violence than its work-ethic-oriented predecessors.

Is it that hard to find an honest and stable employee?

Consider these facts. According to the U.S. Department of Commerce, at least 30% of all business failures each year are due to employee theft and other forms of employee crimes. A study conducted in 1983 concluded that approximately one-third of all employees steal from their employers.

What can be done to combat these losses? To answer this question, it is helpful to understand theft in terms of the profile of a dishonest person and the theft triangle.

Profile of the Dishonest Individual

Research by Dr. William Terris and his associates at London House, Inc., a Chicago-based psychological testing firm, has shown that employee thieves perceive themselves as average persons in a basically dishonest world. In contrast, honest and stable persons perceive themselves as above-average persons in a basically honest world. The employee thief accepts rationalizations of theft behaviors and is tolerant of and less punitive toward theft. If caught, the employee thief responds:

> "I did it because...
> my supervisor is doing it.
> I needed it more than they did.
> the company pays me so little for what they get from me.
> the company took advantage of me.
> it's covered by insurance. No one loses."

In addition, Terris's research has shown that most employee thieves interviewed do not acknowledge their acts of theft as crimes; a criminal is someone who steals more than they do.

The Theft Triangle

Researchers have noted that three elements must be present for an employee to steal: opportunity, need, and attitude.

Opportunity. Most employees have daily opportunities to steal without fear of being caught or punished. In many situations neither apprehension and punishment nor security devices completely deter employee theft. In fact, in some situations, security systems provide the motivated thief with the challenge to be even more creative.

Need. Everyone has needs and wants. Faced with increasing consumer prices (rent, unexpectedly high medical bills, or other expenses), honest employees will either work harder at their jobs in hopes of getting a raise or promotion or they will take on an extra job. The dishonest employee, however, will turn to theft, wherever it is easiest. Often, it is easiest where he or she works.

Attitude. Attitudes determine how needs are satisfied. Honest employees will resist the temptation of theft because of strong moral attitudes.

It would appear that if employers could somehow assess attitudes toward theft, they would be able to screen out individuals prone to theft. In the past, however, employers have used the Diogenes technique--the lighted lantern in broad daylight--and have gotten about the same results. Let's examine some of these traditional hiring techniques in more detail.

Selection Techniques

Good business practice dictates that whatever selection procedure is used in the hiring situation, it should be validated; that is, it should be job-relevant and predictive of future job performance.

Selection Interviews. The interview is widely used in the selection process. It is hard to imagine hiring or recruiting an individual without first having a face-to-face discussion with the applicant.

The typical interviewer relies on personal, subjective values, opinions, and intuitions when making hiring or recruiting decisions. Obviously, there are many pitfalls to this approach. Winning the confidence of an interviewer (and possibly winning the position) can depend upon whether a person i. [illegible] however, does not reliably indicate that a person will have the necessary skills and abilities to successfully perform a job. And it certainly does not reliably indicate whether a person is likely to steal on the job. In fact, data on the validity of the interview as a selection device has been poor. Yet, the interview is often the single most used method of selection.

Application Blanks. Nearly all employers require the completion of application blanks. While responses to questions on an application may be somewhat distorted or exaggerated, falsification is minimized to the extent that an applicant believes that the responses are verifiable. Rarely will an individual admit to leaving a past position because of theft (although one creative individual, later discovered to be a thief, gave as a reason for leaving his last job: "They couldn't afford to keep me").

Resumes. The purpose of a good resume is to sell the applicant to the recruiter. To that end, it is neither in the applicant's best interest, nor reasonable within the format of the typical resume, to include an admission of theft. The trend today is for "formula" resumes, especially with the proliferation of employee outplacement, counseling, and marketing organizations. Certainly, resumes will continue to be used by recruiters, but the resume cannot be expected to be an adequate indicator of future job success or of future employee theft.

Reference and Background Checking. Reference checking generally consists of verification of previous employment, education, personal and business references, and any other information supplied by the applicant that is considered to be job-relevant and deemed important enough by the recruiter to investigate. When the situation warrants, credit and criminal background checks are also made. The name of a reference is provided by an applicant with the expectation that the reference will be favorable. Even when names of references are not specifically provided by an applicant, as in some background investigations, most references tend to be positive. Previous employers are fearful of liability in revealing employee poor performance or dishonesty and are, therefore, very reluctant to supply negative information. Moreover, since some 80% of employee theft goes undetected, previous employers may not even be aware of a former employee's crimes. Reference checks are necessary, but they simply are not enough.

Another pitfall of reference checks and background investigations is that they can be worthless when done superficially by untrained investigators. And many personnel assistants or governmental clerks tend to be just that.

Polygraph Examinations. Diogenes was not the only person in ancient times searching for the honest individual. Both the ancient Chinese and the Arabian Bedouins, pioneers in the development of lie detector tests, believed that the dishonest person's mouth became dry while engaging in deception. To test for dry mouth, the Chinese required that the person chew and spit out rice powder and the Bedouins required that the person lick a hot iron. Dry expectorated powder and burned tongues were considered as evidence of lying. While Diogenes [illegible], presumably some of the Chinese and Bedouins passed this crucial test.

A distinction must be made between two types of dishonesty--lying and stealing. The polygraph was designed to assess untruthfulness. It does so by monitoring the candidate's physiological reactions (pulse rate, blood pressure, respiration, and galvanic skin response) to direct questions. The polygraph is especially useful in facilitating the process of obtaining admissions, often even before the actual polygraph examination has begun.

Besides being costly (often $25 to $50 for preemployment screening), polygraph testing has been barred or restricted as a condition of employment in at least 19 states and has been the subject of continued legislative opposition in many levels. Mechanical truth verification remains highly controversial in legal circles and has generated a great deal of negative publicity. Use of this approach for preemployment screening carries with it the stigma of negative employee relations In addition, polygraph examinations require skilled examiners, but standards and training vary greatly across the country, thus rendering the results of this screening method highly variable. While possibly effective at screening out those candidates with previous criminal activity, there is no evidence that the polygraph can screen out those who, given the right circumstances and opportunity, are prone to stealing.

Paper-and-Pencil Tests. Until comparatively recently, paper-ard-pencil tests for assessing proneness to theft have been virtually nonexistent. Originally developed as a substitute or alternative to the polygraph, paper-and-pencil honesty tests have become entities in their own right.

To validly predict proneness to theft, a good honesty test should assess beliefs about the extent of theft in society, positive attitudes toward theft, ruminations about theft, perceived ease of theft, interthief loyalty, likelihood of detection, knowledge of employee theft. punitiveness, rationalizations about theft, assessments of own honesty, and theft admissions.

In addition, the test should contain a distortion scale, to detect attempts to "fake good" on the test.

All major honesty tests available today attempt, with varying degrees of success, to meet the above requirements in measuring attitudes toward theft. Only two tests, however, appear to be backed up by major, published research: the London House Personnel Selection Inventory and the Reid Report. There are several other tests for which very little technical data are available, including the Stanton Survey, the TA Survey, and the Wilkerson Pre-Employment Audit.

London House Personnel Selection Inventory. The Personnel Selection Inventory (PSI), developed by psychologist William Terris, was first published in 1975. There are several forms of the PSI available to measure attitudes toward and opinions about dishonesty and which include checklists for dollar values of merchandise, property, and money stolen (the admissions section). The only test of its type to do so, the PSI has an independent distortion scale. Other PSI forms available include additional scales for drug abuse, violence, emotional instability, and safety locus of control (accident prevention)

Reid Report. The Reid Report was first published in 1950 by John E. Reid, a polygrapher. The Reid Report consists of two parts: yes/no questions designed to assess the candidate's attitudes about honesty and theft and a set of application blank-type items, followed by items designed to obtain admissions of theft and other crimes.

Comparison of the Two Major Honesty Tests

There are many ways to compare tests that purport to measure the same traits. Key issues for comparison include:

Ease of Administration. Both the PSI and the Reid Report are rather easily administered. The directions are clear and easy to follow. For both tests, it is necessary that the test be given on company premises under controlled testing situations. The tests should never be given or mailed to applicants for them to take on their own.

Ease of Scoring. The scoring of both tests is controlled closely by the respective publishers. The primary reason for this control is test security; the publishers thus ensure protection of the scoring keys and the norms.

Both publishers provide a mail-in service whereby the client mails in test booklets and receives a report, generally mailed within 24 hours of receipt. Using the mails has obvious disadvantages, of course, for clients needing more immediate results; hence, both publishers offer a telephone method of scoring. In this approach, information from the test is clerically compiled by the test administrator and then phoned in to the publisher's test center; the test data are entered on-line, scored, and reported back to the client within seconds. For telephone scoring, the PSI has the advantage of a simpler preliminary coding than the Reid Report, whose preliminary work is more complicated and requires more manipulation.

On site computer scoring is another scoring option. The PSI has a software scoring program that can be run on certain personal and mainframe computers, thus providing for immediate results.

Ease of Interpretation. Both tests provide a measure of dishonesty, with assistance from the publishers in determining recommended/not recommended standards for the hiring of candidates. The PSI, in addition, provides an independent distortion scale that indicates the extent to which the candidate is being truthful about the answers given

Response Format. The Reid Report has mostly yes/no questions, whereas the PSI allows for five to seven or more choices for each test question, thus encouraging more admissions of dishonesty, while reducing applicant stress.

Legality. Both publishers show evidence that their tests are nondiscriminatory and meet various guidelines on employee selection procedures. In addition, the publishers of the PSI and the Reid Report stress in their literature their willingness to assist their clients in any litigation that may be enacted against them

Research Base. Both the PSI and the Reid Report have published research regarding reliability, validity, fairness, and other technical information. Reid supplies eight publications; London House provides 50 published studies on the PSI.

Review by Buros. Both tests have been reviewed in Buros's Mental Measurements Yearbook. Evaluation of The Reid Report appears in the eighth edition (1978); the review of the PSI is in press for the ninth edition and is currently available in retrieval format from the publishers of Buros. The reader is invited to consult the Buros reviews for additional information.

Concluding Remarks

Paper-and-pencil honesty testing has been somewhat of a controversial area. Psychologists, especially those in the testing profession, have shown a healthy skepticism about whether a paper-and-pencil test can really predict proneness toward employee theft. The major publishers of honesty tests, especially London House and Reid, have countered with a number of well documented studies that show correlations with polygraph results. For those who are uncomfortable with polygraph examinations as the criterion, London House has also conducted a series of PSI studies with other criteria and methodology--including time series studies (where retail store shrinkage was reduced following the introduction of the PSI for employee screening), predictive studies (where PSI scores were related to future detected thefts), and contrasted groups (where it was found that charity collectors who were more honest as measured by the PSI turned in more money per day then those who scored as less honest). For those who survey the research literature, it is hard not to be convinced of the validity and viability of a paper-and-pencil approach to screening for honesty.

Recently, the field of honesty testing was given even greater legitimacy by a 1984 review by Sackett & Harris in the prestigious Personnel Psychology. Employers have always known that paper-and-pencil ability tests can be highly predictive of future job success. Now the personnel and psychological communities have recognized that paper-and-pencil honesty tests can indeed predict future employee theft.

And, oh yes, remember Diogenes? too bad that honesty tests weren't available in his day. If so, he would not have had the opportunity to immortalize himself as the seeker of honest men. You see, Diogenes and his father were earlier exiled from their native Sinope, reportedly for tampering with the country's currency!

A MODEL OF PSYCHOLOGICAL STRESS AND CONTROL

M.Kastner and W.Neef
University of the Armed Forces,
Faculty for Economics and Organizational Science
Munich, Federal Republic of Germany

Summary

A model for psychological stress which integrates the approaches of McGRATH (1976) and LAZARUS and LAUNIER (1981) is presented. In the context of the above-mentioned model, a further facet of control (emotional/rational) is added to those which have, up to now, been generally accepted (unstable/stable; external/internal, and global/specific); some empirical proof has been found.
Consequences for diagnosis (esp. the construction of questionnaires) and therapeutic measures by means of which optimal stresslevels can be reached are drawn from the linking of stress and rational/emotional control.

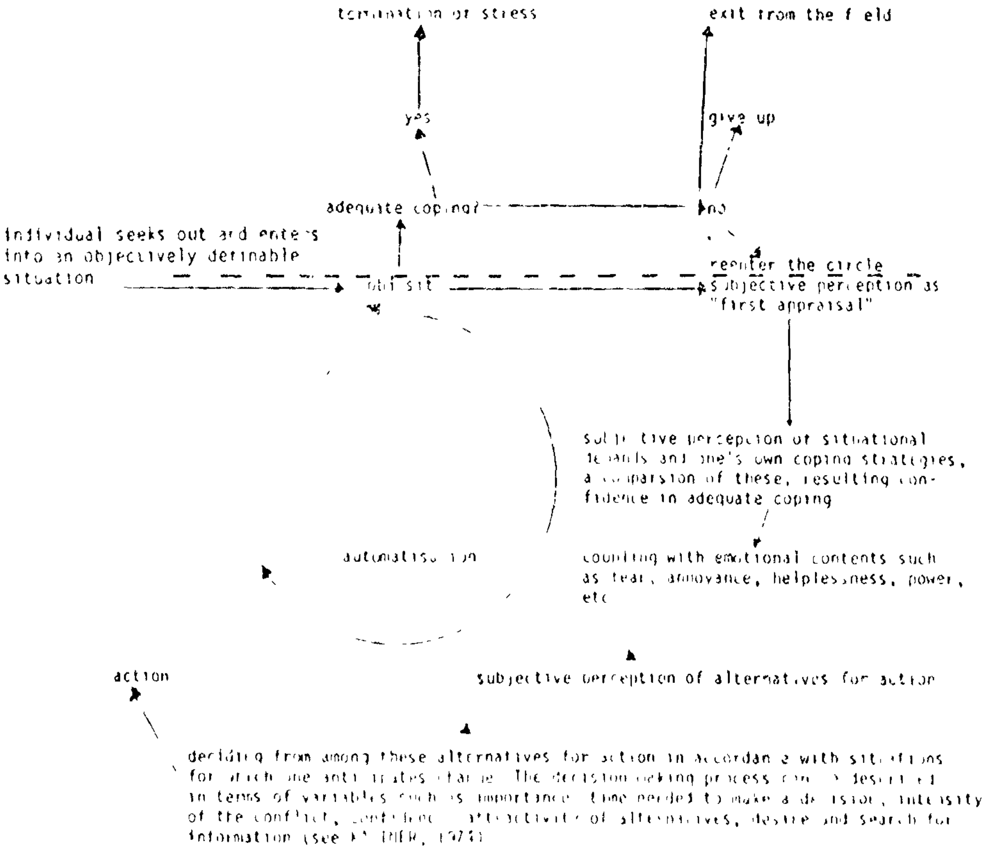
1. A simplified version of the model

The model consists of four components: one based on a "Handlungs"-theory (i.e. a theory of activity), a transactional one, a part based on facet theory, and the aspect of simultaneous registration of physical and psychological parameters. According to the "Handlungs"-theoretical aspect, the unit of observation must be an action. Such an action can only be described as the dynamic reciprocal action between human and situational components. This introduces the second, the transactional aspect. Reciprocal actions should not be described ideographically in their dazzling uniqueness, but these too should be classified, in order to make diagnosis and measures (including the training of actions) possible. Facet theory is brought into play as a method of classifying situations on the one hand and persons on the other. Objective, physically definable parameters and variables of their psychological subjective representation are needed for these situations. For persons we need objective psychological and physiological (test-) variables as well as variables which subjectively represent a person's own psychological and physiological state. We have herewith arrived at the fourth aspect: how can biological-biochemical and psychological parameters be linked with each other? The goal of diagnosis and of therapy is always to attain an optimal level of stress, i. e., the oscillation of actions involving stress within an intraindividually varying range of stress.

Later on, the contribution of control to the description of and to predictions concerning the above-mentioned process shall be discussed.

First the model of an individual action:
If the discrepency between subjective situational demands and an individual's own coping potential becomes too great, the individual no longer tries to meet the demand, but instead exits from the field, rationalises, etc. Reflex actions eliminate the right side of the evaluation process. Felxible actions, in the sense of an optimal level of stress, consist in an optimal combination of automatic "small circles" and "large circles" on higher levels of action within the checking and planning process.
Too little stress leads to insufficient experience with these combinations. Too much stress leads to ineffective, uncoordinated combinations, in which normally automated actions rise in hierarchy and therby use valuable cognitive capacity.

Figure 1: Circular process of an action performed under stress (under the interrupted line) and possible ways of terminating this activity (above the interrupted line)


termination of stress
exit from the f eld
yes
give up
adequate coping?
no
individual seeks out and enters into an objectively definable situation
obj sit
reenter the circle
subjective perception as "first appraisal"
subjective perception of situational demands and one's own coping strategies, a comparison of these, resulting confidence in adequate coping
counting with emotional contents such as fear, annoyance, helplessness, power, etc
automatisation
subjective perception of alternatives for action
action
deciding from among these alternatives for action in accordance with situations for which one anticipates change. The decision making process can be described in terms of variables such as importance, time needed to make a decision, intensity of the conflict, confidence, attractivity of alternatives, desire and search for information (see [illegible], 1974)

2. Facets of control

How can such an action performed under stress (see figure 1) be influenced by processes of control? Let us first, in the tradition handed down to us by ROTTER (since 1966), take a look at external and internal control. An externally controlled person might possibly focus more on the demands of the situation than on his own potential for coping with it. (The opposite can be said for an internally controlled person.) The comparison between situational demands and a person's own potential for coping with the situation will be strongly dependent on external factors. This may have consequences for the choice of alternatives for action. An externally controlled person will tend to create an alternative for action which he subjectively believes exists. He will most likely "construct" his environment to suit his own taste. The circular model of stress can be similarly played through for the various facets of control which have, up to now, generally been accepted.
Especially when the stress-model was applied to depressive behavior and experience (KASTNER, 1984), it became clear that the "generally accepted" facets of control were not adequate. Questioning test persons again and again made it clear, that depressive persons obviously suffered from a "control split", i. e., the rational elements of control did not correspond to the emotional elements. This led to the splitting up of control into rational and emotional control.
Rationaling control can here be defined as a calculated comparison, which is "void of emotion", between the environment and oneself, i. e. the comparison between the stress inherent to a situation and a person's own potential for coping with the situation as well as the appropriate corresponding causal attributions. Emotional control can be defined as the feeling of being in control, of pulling the strings oneself, or of being controlled and having the strings pulled by someone else, i. e., of helplessness.

If we consider only high and low degrees of both facets, then the following systematic combination results:

		Emotional Control	
		+	-
Rational Control	+	A	B
	-	C	D

Thus four groups of persons are formed, those who are both rationally and emotionally controlled (A), those who are rationally controlled but emotionally uncontrolled (B), those who are rationally uncontrolled but emotionally controlled (C) and finally, persons who are both rationally and emotionally uncontrolled (D).

3. Empirical proof to data

Studies on stress among motorists and persons in a state of depression have shown that

- both facets of control meaningfully describe definable qualities and the existence of the four types refered to above can be proved;
- stress and performance differ for each of the four different "types of control", in accordance with the type and difficulty of the given tasks;
- the best performance for simple speed tasks was given by those who were rationally controlled but emotionally uncontrolled (B);
- the best performance for complex (intelligence) tests was given by those who were both rationally and emotionally controlled.
- The persons in group C accomplished somewhat more in both types of tasks than the persons in group D.
- Persons with a high degree of rational and emotional control experience their actions in stressful traffic situations as being more strongly automated than do uncontrolled persons.
- In addition, rationally and emotionally controlled persons perceive stress situations as being less optically complex. Thus, simplifying the situation and meaningfully integrating it into the context of ones actions seems to be an important function of control.
- With respect to physiological variables, persons who are emotionally controlled but rationally uncontrolled are more excitable than test persons who are both rationally and emotionally uncontrolled (see KASTNER & GUILLOT, 1983).

4. Consequences for future diagnoses and measures

In all of our empirical studies, control as a construct applicable to persons and controllability as an attribute of a situation, proved to be the most predictable variable. The psychological variables we used, taken as a whole, described actions under stress better than all the physiological variables (see KASTNER, 1980; KASTNER & GUILLOT, 1981, 1983). In attempts to link different variables in various manners, it was shown:

- that, first of all, multiplicative linkage of the variables intensity, duration and controllability of the action performed under stress proved more useful than additive linkage,
- that, secondly, control was always the determining variable when it was negatively attributed.
- When control was positively attributed, on the other hand, intensity and, above all, duration, became more important.

Thus, when describing and predicting stress factors, it will be necessary to try to systematically include all facets of control. First of all, controllability must be regarded as peculiar to situations. This contritutes to the classification of the situation. Secondly, but of no less importance, control must be regarded as peculiar to persons. In this case, the process in which stress is dealt with will differ, according to whether a person is externally or internally directed, instable or stable, globally or specifically oriented, emotionally or rationally controlled, or according to how these various possibilities are combined.

The diagnosis can be made using a questionnaire which is constructed on the basis of facet theory, and which systematically varies and combines these facets of control. The therapeutic consequences, on the other hand, should take this differentiation of control into consideration. A person who is rationally controlled but emotionally uncontrolled must be differently treated, taught, trained etc., in consideration of his emotional processes, than a person who is emotionally controlled but rationally uncontrolled.

References

Kastner, M. (1980): Eine Validierung von Beanspruchungsindikatoren im Labor. Cologne: BASt Research Report.

Kastner, M. & Guillot, G. (1981): Beanspruchung von Kraftfahrern im kontrollierten Feld. Cologne: BASt Research Report

Kastner, M. & Guillot, G. (1983): Beanspruchung des Kraftfahrers in Abhangigkeit von Persönlichkeitsmerkmalen. Cologne: BASt Research Report

Kastner, M. (1984): Psychische Beanspruchung und padagogische Massnahmen im Hinblick auf eine Humanisierung des Arbeitsprozesses. Zeitschrift fur erziehungswissenschaftliche Forschung, 18, 131-144

Lazarus, R.S. & Launier, R. (1979): Stress-related Transactions between Person and Environment. New York: McGraw Hill

McGrath, J.E. (1976): Stress and Behavior in Organizations. In: Dunette, M.H. (ed.): Handbook of Organizational Psychology. New York: Rand McNally

Methodological Considerations in CBI Research

Theodore M. Shlechter and John A. Boldovici
US Army Research Institute, Fort Knox, KY

The purpose of this paper is to examine some methodological considerations in CBI (computer-based instruction) research as applied to military training. We hope that this discussion will help military decision makers in assessing the usefulness and limitations of CBI. We hope also that the issues discussed here will help behavioral scientists and others who provide technical advisory service to the military in assigning priorities to, and evaluating the outcomes of CBI research.

The Armed Services are planning to spend millions in the next few years to expand the use of CBI. The US Army Armor School, for example, is planning to spend over three million dollars in the next three years to develop 350 hours of courseware for various aspects of armor training. The plans include training various levels of people to perform tasks ranging from using volt meters to land navigation and tactical decision-making.

A major source of support for using CBI in military training is the results of research studies which suggest advantages of CBI over conventional instruction. Several reviews of literature on CBI, Kemner-Richardson, Lamos, and West (1984), and Orlansky (1985), for example, have indicated that the use of CBI led to reduced training time and concomitant reductions in price, with little or no decrease in effectiveness. Those reviews also suggest that students favor CBI over conventional instruction. The cited reports have not, however, analyzed possible methodological problems in the reviewed CBI research. Without such analyses one is ill-equipped to make recommendations and decisions about the use of CBI for military training. Our discussion will address a few of the methodological considerations which seem relevant in assessing the relative merits of CBI and conventional instruction as instructional delivery systems.

Methodological Considerations

The methodological considerations in CBI research affect conclusions about costs, attitudes, and effectiveness.

Costs

Orlansky (1985) indicated that the promise of CBI lay in the potential for saving time and money: millions of dollars could be saved by reducing military training time. Orlansky provided the following formula for computing time savings attributable to CBI: Percent Saving = (CBI Time/Conventional Time) x 100. However, for this formula, CBI was being compared to conventional classroom instruction, which was not an equivalent medium to CBI. Avner, Moore, and Smith (1980) have argued that CBI should only be compared to other self-paced individualized instructional media, e.g., programmed texts. Their argument is especially relevant for training time data, because CBI and other forms of individualized instruction allow the students to proceed at their own pace, while conventional instruction does not. Time savings are thus ascribable not to CBI, but to self-pacing, which characterizes nearly all modern instructional innovations. Time savings of approximately 50 percent for both CBI and programmed texts as compared to conventional instruction have in fact been reported when all three media have been simultaneously tested (Orlansky, 1985). And programmed texts are usually a fraction of CBI's cost for initial implementation.

The costs of alternative media can be examined, not only by comparing training time, but also by estimating life-cycle costs. Kemner-Richardson et al., (1984) noted, for example, that the PLATO system as compared to conventional instruction would eventually lead to $180,000 a year savings. Such long-term cost analyses of CBI systems versus other systems traditionally have involved estimating administrative costs and increases or decreases in the number of faculty which would accompany implementation of new programs. Such analyses, however, have provided only a partial picture of long-term costs. Personnel costs, for instance, also should include expenses associated with training teachers to be "computer-experts." As Shavelson, Winkler, Stasz, Feibel, Robyn, and Shea (1984) observed, using a CBI system in a typical school necessitated a staff development program which trained teachers to be thoroughly knowledgeable in computers. Clearly, such hidden-costs as teacher training programs would reduce the potential savings for CBI.

Cost estimates also should include determining the reliabilities of the compared systems. System reliability is important to measure, because repairing malfunctions will cost money in repairing the problem area and in training time losses. One formula (Frances, Welling, & Levy, 1983) is (Failures Per Hour x Terminals Affected)/(Working Days by Terminals Affected) x 100. The formula, however, only provides a partial index to the educational and financial losses associated with faulty delivery systems. The index does not take into account the potential problems of an implemented system. For the most part, evaluations are conducted for a prototype system which may not relect a system's reliability when implemented. Data should then be collected after the system is fielded and fully implemented.

Evaluators should also collect data on administrative costs after the system has been fully implemented. Orlansky (1985) surprisingly demonstrated that assumptions about administrative costs associated with CBI had rarely been systematically analyzed. For example, very few CBI evaluations have examined the actual expenses (and problems) associated with a CBI system for updating the instructional materials and for hiring the additional personnel needed to operate and maintain the system's equipment. Information obtained when a system is fully implemented is then needed to compare actual costs to estimated costs and correspondingly to determine the validity of certain assumptions about CBI's value for administrative purposes.

Attitudes

Students' attitudes toward CBI may be influenced by instructor views. Clark and Leonard (1985) have suggested that teachers' believe that CBI would help the educational program. One would expect that students would have similar positive views on CBI, because students' attitudes toward an instructional system are influenced by their teachers' attitudes (King, 1975). Determining teachers' attitudes toward CBI is needed to provide insights into students' attitudes. Teachers rather than the system itself may dictate the students' views.

Instructor obtrusiveness is relevant in evaluating military CBI programs, because instructors usually are involved in the evaluation. Draxl and Aggen (1981) described the need to give briefings to enlist military instructors' interest in, and support of, new instructional systems. Such briefings may have engendered positive attitudes in instructors, while militating against obtaining unbiased reports from students.

Discrepancies between students' responses to questionnaires on the one hand, and more objective data on the other, were found by Shlechter (1985): Soldiers reported that using a light pen was easy, when in fact many of their errors related to problems with using this responding mechanism, such as holding the light pen incorrectly. Additional efforts to compare self-reports with objective performance measures seem warranted.

Shlechter also found that students showed fatigue while completing CBI lessons, but did not report such fatigue on the subjective evaluation questionnaire. Objective measures of fatigue are an important issue in CBI research because the human-factors literature (see McVey, Clauer, & Taylor, 1984) indicates that VDTs (visual display terminals) may be "visual discomfort terminals" with students' not being able to use certain types of terminals for extended periods of time. Evaluators must then measure through objective indexes human factors variables which may affect users' comfort.

Consideration also should be given to examining other factors which may confound attitude reports. Orlansky and String (1981), for example, reported that half the courses which they reviewed lasted one week or less. Such limited duration might not be adequate for ascertaining students' attitudes toward a CBI system. King (1975), and Clark (1985) have noted that most students initially enjoyed working with a CBI program; however, questions remain about a system's ability to sustain students' motivation and enthusiasm once the novelty has worn off.

Effectiveness

Objective comparisons of CBI and other media require that the compared groups be treated identically in every respect except those under investigation. Instructional content, for example, should be the same for the compared groups, as it was in Morrison and Witmer's (1983) comparison of computer-based and print-based job aids. Other media comparisons did not, as noted by Clark (1985), deal with the issue of matching content. For instance, ETS's often cited evaluations of PLATO and TICCIT (Alderman, Appel, & Murphy, 1978) did not include a specific comparison of course materials presented in the compared classroom. Any differences in ETS's evaluations might have been due to presenting different content and not to any features inherent in the CBI system. Evaluators cannot make any claims about the relative effectiveness of CBI as as instructional delivery system unless unconfounded comparisons are made with other media.

Unconfounded comparisons between CBI and other media also require that both educational programs be created with the same degree of effort. Stone (1985) described the extensive effort involved in developing an in-house set of CBI lessons. He showed that developing this courseware involved a team of

Army Captains, civilian educational specialists, and outside consultants working for a year. What typical Army Instructors go to such an effort to develop their daily lesson plans? CBI evaluators must then take into consideration such differences in instructional efforts when discussing their evaluation results. If differences between CBI and conventional instructions only revolve around this issue of instructional effort, then the Army should consider putting forth similar efforts in creating instructional plans to be used by classroom teachers.

CBI evaluators should also determine a system's instructional efficiency by its ability to help the student accomplish different levels of transfer. Clark and Voogel (in press) suggested that CBI was usually geared for immediate transfer and frequently ignored the skills and strategies necessary for long-term retention. They also suggested that CBI courseware was not geared for cognitively oriented tasks, such as problem solving. These trends are disturbing to educators, who place a premium on long-term mastery of information and the ability to help students develop problem-solving skills. Unfortunately, very few CBI evaluations have examined the long-term and cognitive impact associated with this medium.

The apparent effectiveness of CBI may also be an artifact of unwarranted instructional prompting. Clark and Leonard, after reviewing 42 randomly selected civilian CBI programs, found that teachers usually provided CBI groups with more instructions, e.g. prompts, to complete the tasks than they provided for control students. The extent to which prompting is a problem depends, of course, on the observed amount of prompting relative to the amount intended by the system's designers. Unwarranted prompting should then be another variable measured in CBI research.

Summary and Conclusions

In summary, the following methodological considerations have been discussed regarding CBI research: 1) make unconfounded comparisons between the CBI system and other appropriate educational media; 2) measure hidden life-cycle costs associated with the delivery system; 3) determine the system's reliability; 4) measure actual life-cycle and reliability costs for an implemented system; 5) determine teachers' attitudes toward the CBI system; 6) control for possible "instructor obtrusiveness" effects; 7) substantiate subjective evaluational data with more objective measures; 8) measure human factors variables with objective indexes; 9) control for possible confoundings due to insufficient testing duration; 10) make sure that the compared media have the same content; 11) control for differences in instructional efforts; 12) examine students' long-term and cognitive mastery of the information; and 13) measure unwarranted prompting. Other methodological issues which cannot be described in this paper involve the possible interactions which may exist between student characteristics and the experimental treatment. As argued throughout this paper, clearer answers about CBI's inherent value as a delivery system can be obtained if these considerations are incorporated into the evaluation process. One cannot conclude that CBI is the superior educational medium when confounded comparisons are made with inappropriate media.

These methodological problems reflect the complexities in CBI research. To the extent that CBI research continues to compare CBI to other instructional media, then some time-consuming and expensive research procedures must be employed. For one thing, both cross-sectional and longitudinal data should be collected. A cross-sectional design is needed for initial assessments of students' CBI performance, while longitudinal data are necessary to ascertain the long-term learning associated with the CBI system. Secondly, systematic programs of research are needed in which priorities are assigned to independent variables and variables are systematically manipulated and measured in successive experiments. Such research programs are imperative to analyze systematically as many of the previously cited methodological considerations as possible. Systematic analyses would provide, for example, information about the relationship between estimated and actual CBI life-cycle costs. Programmatic researches needed to provide further insights into whether CBI problems are due to courseware or hardware limitations.

Evaluators should perhaps shift focus from questions of inherent superiority to the identification of the conditions under which CBI and alternative media produce and do not produce desired results. Various media have various strengths which must be first enumerated and then matched with intended instructional settings objectives, and resources. Zemke's (1984) conclusion that CBI may best be used as a supplement to classroom instruction is a case in point.

Military educators should only begin widespread implementation of CBI after clearer answers are provided about this medium's instructional and financial value. History has shown that educational innovations which were implemented without sufficient research and planning were always abandoned for later technological innovations (Montague & Wulfeck, 1984). There is currently some evidence that this abandonment process is beginning to occur for some CBI programs (D Reed, Personal Communications, 6 November 1985). With sizeable financial and personnel investments associated with large-scale CBI implementation, the military can ill afford to continue this historical process. This abandonment process would also be unfortunate because computers--if used properly--could be a valuable instructional tool.

REFERENCES

Alderman, D. L., Appel, L. R., & Murphy, R. T. (1978, April). PLATO and TICCIT: An evaluation of CAI in the community college. *Educational Technology*, *18*, 40-46.

Avner, A., Moore, C., & Smith, S. (1980, May). Active external control: A basis for the superiority of CBI. *Journal of Computer-Based Instruction*, *6*(4), 115-118.

Clark, R. E. (1985). Confounding in educational computing research. *Journal of Educational Computing Research*, *1*(2), 137-147.

Clark, R. E., & Leonard, S. (1985, March). Computer Research Confounding. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Clark, R. E., & Voogel, A. (in press). Transfer of training principles for instructional design. *Educational Communication and Technology Journal*.

Draxl, M., & Aggen, W. (1981). *Computer-based instruction project: A final report*. College Park European Division: University of Maryland.

Frances, L. D., & Welling, L. G., & Levy, G. W. (1983). An evaluation of the CAI (Computer-Assisted Instruction) segments of the TOW field-test site and the HAWK CW radar repair course. (TDI-TR-83-3). Fort Monroe, VA: US Army Training and Doctrine Command.

Kenner-Richardson, S., Lamos, J. P., & West, A. S. (1984). The CAI decision handbook. Lowry AFB, CO: US Air Force Human Resources Laboratory.

King, A. T. (1975). Impact of computer-based instructions on attitudes of students and instructors: A review. (AFHRL-TR-75-4). Brooks AFB, TX: US Air Force Systems Command.

McVey, B. W., Clauer, C. K., & Taylor, S. E. A comparison of antiglare filters for positive and negative image displays. (HFL-50). San Jose, CA: IBM Human Factors Center.

Montague, W. E., & Wulfeck II, W. H. (1984, Winter). Computer-based instruction: Will it improve instructional quality? Training Technology Journal, 1(2), 4-19.

Morrison, J. E., & Witmer, B. G. (1983). Comparative evaluation of computer-based and print-based job performance aids. Journal of Computer-Based Instruction, 10(3 & 4), 73-75.

Orlansky, J. (1985, October). The cost-effectiveness of military training. A paper presented at a NATO symposium on Military Training and Cost-Effectiveness of Training. Brussels, Belguim.

Orlansky, J., and String, J. (1981, second quarter). Computer-based instructions for military training. Defense Management Journal, 46-54.

Shavelson, R. J., Winkler, J. P., Stasz, C., Feibel, W., Robyn, A. E., & Shea, S. (1984, March). "Successful teachers" patterns of micro-computer based mathematics and science instructions. Santa Monica, CA: Rand Corporation.

Shlechter, T. M. (1985, November). CBI and students' attitudes. Paper to be presented at the Annual meeting of the Mid-South Educational Research Association, Biloxi, MS.

Stone, C. S. (1985, March). Instruction at the US Army Engineer School: Evolution and management of an in-house coding team. Paper presented at the International Meeting of the Association for the Development of Computer-Based Instructional Systems, Philadelphia.

Zemke, R. (1984, May). Evaluating computer-assisted instruction: The good, bad, and the why. Training, 21, 22-47.

The views expressed in this paper are those of the authors and do not represent the official opinions or policies of the US Army Research Institute or the Department of the Army.

THE EFFECTIVENESS OF COMPUTER GRAPHICS IN VISUAL RECOGNITION TRAINING

BARBARA MCDONALD
AND
BETTY WHITEHILL

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
SAN DIEGO, CALIFORNIA 92152-6800

Abstract

The effectiveness of computer-based graphics in training a visual recognition skill (radar jamming) was investigated. In recognizing jamming, Naval personnel must differentiate between several types and their visual characteristics. Computer graphics were developed to enhance the features of each type of jamming. The graphics were accompanied by captions which gave the title and a one-sentence description. An experiment was conducted to systematically test the effects of graphics. Subjects were students at the Fleet Combat Training Center, assigned to one of six conditions in a 2x3 design. Subjects received training on animated graphics, with or without captions, still graphics, with or without captions, or captions only, or were assigned to a control group. Subjects were then tested using videotapes of actual jamming. Test results are discussed in terms of the relative merits of animated versus still graphics, and the value of captioning.

Introduction

The use of graphics to enhance the learning process has been of interest to the educational community for years. Research on the use of graphics and text illustrations has shown that they can improve learning by (1) helping students understand what they have read, (2) having pictures substitute for words, and (3) enhancing learner enjoyment. They can also increase retention of information by having more than one way of encoding information (i.e., pictures plus words) and through repetition (Levie & Lentz, 1982).

With the advent of computer based training, graphics have been an even more interesting aspect of instruction since they can be made to move, flash, and present in color. However, the evidence on how critical they are to the success of the instruction is mixed. Since they increase development cost significantly, whether they are necessary and the way in which they are used is of importance (Moore, Nawrocki & Simutis, 1979). It is likely that the importance of graphics depends largely on the task to be learned.

In the present study, graphics were a logical extension of the instruction, the purpose of which was to train visual recognition skills. The graphics were designed to offer the students a series of simplified visual patterns to help organize the learning of important visual characteristics. In the present experiment, two aspects of graphics were explored in addition to the effectiveness of the graphics in general. First, the value of animation was addressed and second, the value of captioning was explored.

Background

The present experiment was designed to investigate the influence of computer-based graphics on recognition of visual patterns. The training and testing of visual recognition skills in Naval personnel has been of longstanding importance. A critical skill in the Navy is the recognition of electronic countermeasures (ECM) also known as "jamming", on radar systems. In recognizing jamming several jamming types and their visual characteristics must be differentiated. NPRDC developed a training program to teach the recognition of 11 different types of jamming. This program used computer-based instruction combined with videotapes of jamming. The computer-based training presented the instruction and also presented graphics of the jamming types to enhance the visual characteristics. This training program was extremely successful (see McDonald & Crawford 1983, 1985).

The computer-based instruction for recognition of jamming was designed to focus on categorical features of the different types of jamming. Each type of jamming has critical visual features which are unique and which are important in defining and identifying the type. In real life however, jamming can be very difficult to identify because the patterns are complex and can be ambiguous with respect to the critical features. In training, these features were difficult to display in real-life videotape presentations. As a result, the videotapes by themselves were not sufficient to train students to identify each type of jamming. Because of this, graphics were used to highlight the critical features. These graphics were "cartoon-like" renditions, making the features of each jamming type very clear and exaggerated. The literature on concept learning provides evidence that in order to teach students to recognize the prototypical concept, lots of examples must be given and prototypical features must be emphasized (Gagne & Briggs, 1979). It was the purpose of the present experiment to test the use of the graphics as "feature enhancers" since they were considered extremely helpful in teaching the students by helping the students attend to the features critical to identification of each type. Three questions were of interest in the present experiment (1) were the graphics effective in teaching the recognition of different types of jamming (2) were animated graphics more effective than still graphics, and (3) would captions (short verbal descriptions of the jamming types) add to the value of the graphics?

In order to test the effects of graphics without the confounding influence of the accompanying instruction the experiment was conducted prior to the start of the instruction. For experimental purposes the pretest to the instruction (which consisted of a videotape with examples of actual jamming), served as the dependent variable. In the experiment, there were 5 graphics conditions used and the students previewed the graphics before they took the pretest and went on to the instruction. The five conditions used were animated graphics, with or without captions, still graphics with or without captions, and captions alone. These five conditions were compared to a control group that did not receive any pre-instruction. The use of these particular conditions permitted evaluation of whether or not graphics were helpful and also were designed to tease out exactly which aspect of graphics use was the most helpful.

Method

Subjects

Subjects were 114 male enlisted Naval personnel, attending ECCM classes and Advanced Warfare classes at FCTCPAC. Their Navy rates included Operations Specialists (OS), Fire Control Technicians (FT) Electronic Technicians (ET) and Electronic Warfare Technicians (EW). These personnel ranged in rank from E3's to E7's.

A control group was formed from previously collected data in which students attending the same course were given the training as a whole package and did not receive the graphics as a preview condition. 93 students comprised the control group.

Experimental Materials

Two types of graphics were used within the computer-based instruction and were therefore tested in the experiment. In the narrated instruction the graphics used were still graphics with captions. The captions had a title of the jamming and a short description of what the jamming looked like. In the other part of the training the graphics were animated with captions. The captions used were the same as used in the still graphics. In order to tease out the contributing factors to the graphics effectiveness there were five graphics conditions in the experiment. They are as follows

a) STILL GRAPHICS WITH CAPTIONS two computer frames of each type of jamming (11 types) with a visual representation (graphic) of the jamming type. Underneath the graphic was the title of the jamming and verbal description of approximately 1 sentence.

b) ANIMATED GRAPHICS WITH CAPTIONS. Two frames of each type of jamming in which the graphic of the jamming type was animated (i.e. the picture presented a moving jamming type on the simulated radar scope). Underneath the graphic was a Title and a verbal description.

c) STILL GRAPHICS WITH TITLE AND NO CAPTIONS. The graphics in this condition were the same used in the STILL GRAPHICS WITH CAPTIONS described above except that underneath the graphic was the title only and no verbal description.

d) ANIMATED GRAPHICS WITH TITLE AND NO CAPTIONS. The graphics in this condition were the same used in the ANIMATED GRAPHICS WITH CAPTIONS except that underneath the graphic was the title only and no verbal description.

e) JAMMING TITLES WITH CAPTIONS AND NO GRAPHICS. In this condition which served as a "control" to the graphics conditions there were no graphics presented, only the Jamming titles and verbal descriptions of each type of jamming were presented.

Experimental Procedure

The data for the graphics experiment was collected from May, 1983 to June 1984. The experiment was designed so that the preview condition would "fit" onto the front end of regular course materials and computer- based instruction. ECM and Advanced Warfare classes were used for this experiment. These classes were each scheduled once a month or every two months. Each class was comprised of 8-10 students. Subjects from each class were randomly assigned to one of the five preview conditions. The students took the preview condition on one day. After a one-day delay students in all conditions took the pretest. After the pretest they proceeded to complete the computer-based training (which is not of concern in this paper).

Dependent Measures

The dependent measures used in the experiment were the scores obtained on the pretest used in the computer-based instruction. The pretest consisted of videotaped examples of actual jamming. There were 2 videotaped examples of each type of jamming (22 items). In addition, the response time for recognition of each example of jamming was measured in seconds. Finally, for experimental groups total time spent in the preview condition was measured (in minutes).

Equipment

The microprocessor used for this training was the TERAK, an LSI-11 based dual floppy disk drive system with a 32K work memory capability. It included a keyboard and CRT display for the presentation of black white graphics and text. The computer was used for presentation of instruction and graphics, for presentation of tests and test results and for data collection. The video equipment used for videotape presentations was a Betamax videotape player and TV monitor.

Results

Table 1 presents means and standard deviations for pretest scores for experimental groups and the control group. An analysis of variance was performed and revealed statistically significant differences between the experimental groups (1-4) and the control groups (5 and 6), F(1,205) = 11.55, p < .001. Further analysis revealed that the animated conditions (groups 2 and 4) had statistically higher scores on the dependent measure than the still graphics conditions (groups 1 and 3) F(1,92) = 6.17, p < .01. Finally, according to subsequent tests for differences, the animated with captions group accounted for the significant differences. An analysis of variance between the captions groups versus the no captions groups showed no significant differences, F(1,92) = 2.82, p = .09

PREVIEW TEST SCORE MEANS FOR ALL GROUPS

Condition	N	MEAN	SD
1 Still Captions	24	48.41	21.69
2 Animated Captions	20	62.50	14.46
3 Still No Captions	25	46.24	15.59
4 Animated No Captions	25	51.16	15.15
5 Captions only	20	40.95	19.58
6 Control	93	43.47	17.89

Table 1 Means and Standard Deviations for Preview Test Scores for all Groups

Response times for recognition of jamming was monitored as well as performance scores. The time in seconds was measured for all instances of recognition from the moment the jamming segment was presented until the computer recorded the student response. Table 2 presents the means and standard deviations for response times for experimental and control groups. Statistical comparisons revealed that groups did not differ significantly in response times, F(5,201) = 1.00, p < 0.419. It is interesting to note that while there were no significant differences, performance was slower for the text only group (5), and the control group. These groups did not have the advantage of previewing the graphic displays of each type of jamming.

PREVIEW TEST RESPONSE SECONDS

Condition	N	MEAN	SD
1 Still Captions	24	10.83	14.87
2 Animated Captions	20	13.90	15.03
3 Still No Captions	25	10.68	15.23
4 Animated No Captions	25	11.32	12.74
5 Captions only	20	15.30	10.40
6 Control	93	15.88	15.28

Table 2 Means and Standard Deviations for Response Seconds for all Groups

Table 3 presents the total time (in minutes) spent actually previewing the graphics for the experimental groups. An analysis of variance revealed significant differences between experimental groups, F(4,109) = 18.09, p < .001. A subsequent test for differences revealed that the animated graphics group with captions (group 2) spent more time viewing the graphics prior to taking the pretest.

PREVIEW TOTAL MEAN MINUTES			
Condition	N	MEAN	SD
1 Still Captions	24	10.66	1.46
2 Animated Captions	20	13.10	[illegible]
3 Still No Captions	25	9.72	1.20
4 Animated No Captions	25	11.92	1.28
5 Captions only	20	16.45	1.95

Table 3. Total Mean Minutes Viewing Graphics For All Groups

Conclusions and Discussion

Three questions were of interest in the present experiment. First, the overall effectiveness of graphics in teaching the skill of jamming recognition was explored. Second, the value of animation was tested and finally, the use of captions to accompany the graphics was addressed. The results of the experiment revealed that graphics were helpful and specific information about the use of animation and captions was provided as well.

The graphics groups outperformed the control groups on the pretest, so even though they had not seen actual jamming before, the graphic provided them with enough information to recognize actual instances of the jamming. A more detailed look at the results showed that it was the animated with captions group which accounted for the significant differences. This finding suggests the importance of both animation and captions. Interpretation of this finding must be made with caution, however, because of the fact that there was some similarity between the animated condition and the test condition (they were both moving pictures). Captions appeared to add to the value of the animated condition although there were no main effects for captions. The captions may have provided the animated group with a more efficient way to process and retain the preview information.

Analysis of the time spent previewing the graphics provides another clue to the superior performance of the animated with captions group. This group spent more time in previewing the graphics than did the other graphics groups. Here again, the combination of animation with captions may have given this group more information to process thereby accounting for more time spent. In terms of actual response time in the test itself, there were no significant differences between groups.

The results of this experiment provide some interesting guidelines for the use of graphics. First, graphics appear to be a useful way of giving students advance information about visual characteristics of complex patterns. The findings of this study do suggest that it is important to consider the characteristics of the visual pattern to be learned in developing the graphics. In our case, simplifying the visual pattern was important However including the movement through animation turned out to be critical even though it did add some complexity Finally, the use of captions cannot be trumpeted based on this experiment but the findings do suggest that they added to the value of the animated graphics

References

Gagne R.M. & Briggs, L.J. Principles of Instructional Design
New York Holt Rinehart & Winston, 1979

Levie W.H. & Lenz, R. Effects of Text Illustration A Review of the
Literature, Educational Communication and Technology Journal, 1982,
Vol 30, 4 195-232

McDonald B.A. & Crawford, A.M. Remote Site Training Using Microprocessors,
Journal of Computer-Based Instruction, 1983, Vol 10, 83-87

McDonald, B.A. & Crawford, A.M. Microprocessor-Based Training School and
Shipboard Evaluations, Navy Personnel Research and Development Center
Technical Report 51-85 06, In Press

Moore M.V., Nawrocki, L.H. & Simutis, Z.M. The Instructional Effectiveness of
Three Levels of Graphics Displays for Computer-Assisted Instruction, U.S.
Army Research Institute for the Behavioral and Social Sciences Technical Paper
359 April 1979

Recent Developments in Job Evaluation Research

Stanley D. Stephenson
School of Business
Southwest Texas State University
San Marcos, Texas 78666

Job evaluation is widely used to establish a specific salary for a specific position. The point system is the most widely used job evaluation procedure. In this system, one or more factors are used to rate each job. Research has typically yielded four factors: Skill, Effort, Responsibility, and Working Conditions. Each of these factors can have one or more subfactors, or scales. For example, under the Skill factor Fraser, Cronshaw, and Alexander (1984) used four scales: Education, Experience, Accuracy, and Complexity. Within each factor (or scale) there are several levels of worth; each level is assigned a certain number of points. Judges (typically experienced employees) then rate each factor, usually based on information contained in a job description, and assign the established number of points. The total number of points across factors equals the job's worth to the organization. All jobs are then ranked according to their point total, and wages are set.

In spite of the wide spread use of job evaluation to determine the salary paid to an employee, recent research has been limited. However, in the late 1970s job evaluation in general came under somewhat of an attack, and consequently some research has been conducted in the last several years.

This research was generated by critical comments concerning the use of job evaluation for determining wages. For instance, one major report (Treiman & Hartmann, 1981) concluded that the reliability of job evaluation methods for determining wage levels has not been established. Treiman (1979) noted that there have been no studies of the validity of job descriptions which in theory form the basis for any job evaluation system. The real impact of these criticisms is that there may now be a "perceived" lack of credibility concerning job evaluation.

More specifically, job evaluation methods came under closer scrutiny due to the comparable worth movement which is rapidly becoming a dominant personnel issue of the 1980s. Comparable worth refers to receiving equal pay for equal job value. Proponents of this movement argue that the historical male-female wage difference is due primarily to a bias inherent in the job evaluation process. Consequently, comparable worth advocates are calling for the use of job evaluation methods that can objectively assess differences in basic job factors, a measurement quality that may not exist in current job evaluation techniques.

The chief criticism of job evaluation is that it is inherently judgmental and therefore possibly biased. Bias can enter into the process at two points: in the writing of the job description and in the evaluation of the job description with respect to the factors/scales selected. In other words, the rating of jobs and the writing of the job description are potentially subjective events.

The sum of the recent questioning of job evaluation techniques has been the publication of several research studies designed to test some of the underlying assumptions, and recent criticisms of bias in job evaluation. This paper attempts to summarize the findings of these articles.

Rater Reliability

Treiman (1979) has stated that "the reliability of ratings is not particularly encouraging"(p. 40). Doverspike, Carlisi, Barrett, and Alexander (1983) conducted their own review of the literature and found, contrary to Treiman, rather high interrater reliability across studies. They then had 10 raters rate 20 job descriptions, for jobs ranging from clerk to accountant to sales correspondent to supervisor, on a point system job evaluation instrument. The instrument had four factors (skill, effort, responsibility, and working conditions) and a total of 11 scales.

They used a relatively new analytical technique, generalizability analysis. This procedure uses a random effects ANOVA design and permits the analysis of each potential source of error that may affect job ratings. In this study, three facets were analyzed: jobs, scales, and raters. The authors found that the rater factor and its interactions with both scales and jobs produced little variance. With adequate training, sufficient job information, and a properly designed point job evaluation system, the job scores produced by the 10 raters yielded adequate levels of reliability. Moreover, "reliability dropped only slightly when the number of raters, assumed to be from the universe of trained raters, was reduced from 10 to 1"(p. 481). These results certainly do not agree with Treiman's (1979) conclusion; they also question the need for the usual recommendation to have a minimum of 10 raters. Others (e. g., Fraser et al, 1984; Doverspike & Barrett, 1984; Madigan, 1985; and Stephenson, 1985) produced similar results. Also, on a related dimension, neither sex of the job incumbent (Schwab & Grams, 1985) nor sex of the job evaluator (Doverspike et al, 1983; Doverspike & Barrett, 1984) appear to influence final job scores.

In sum, it appears that job evaluation raters contribute very little variance to the job evaluation procedure. If properly trained and adequately informed with respect to the jobs being evaluated, raters can and do reach reliable consensus about the worth of jobs. Moreover, this consensus seems to be obtainable with as little as three raters. Consequently, on the surface raters do not seem to have a biasing impact on job evaluation.

The term, on the surface, was used to preface a finding reported by Madigan (1985). As noted, he found rater reliabilities of at least .85. However, he also reported that the lowest standard error of measurement was ± 40 points at r = .90. "Hense the 95% confidence interval range of 160 points encompassed four possible classification assignments" (p.145). Doverspike et al (1983) also reported large confidence intervals even with high rater correlations. Error variances this large might be unacceptable for comparable worth evaluation.

Madigan (1985) also reported that in the best of three job evaluation methods he found classification level agreement in only 61 of 120 interrater comparisons. Moreover, he reported classification differences of two or more pay grades in 11% of the cases. Gomez-Mejia, Page, and Tornow (1982) suggested a need to examine evaluation 'hit rates' (the percent of cases for whom the estimated and actual grades were the same) as well as the traditional interrater correlational results. Operationally, these authors considered a position a 'hit' if it was classified by the job evaluation system within ± 1 grade of its assigned grade. They noted that the relationship between correlations and hit rates is uncertain; a correlation of .82 between predicted and actual grade may have a lower hit rate than a correlation of .62. In their best job evaluation method, ± 0 hit rates occurred in 40 percent or fewer of the cases.

Consequently, even though rater reliabilities may be acceptable, this measure may not be capturing the entire picture. Even when correlations are in the .90 range, the percent of actual hits may be unacceptable for actual classification purposes. Madigan (1985) expressed the belief that assessment of potential error variance in evaluation measures must go beyond traditional measures to include an analysis of impact on pay or classification decisions. He also expressed the need to establish acceptable error intervals for determining wages.

Madigan (1985) summarized, "The psychometric adequacy of job worth measures generated by point, guide chart, and PAQ [Position Analysis Questionnaire] evaluation methods is open to serious question. Results of this investigation indicated that previous studies of job evaluation understated potential inconsistency in classification decision making attributable to measurement error." "Consequently none of the three evaluation methods evaluated here exhibited the psychometric qualities desired of a procedure that will serve as the governing criterion in pay classification decisions"(p. 146).

Job Evaluation Techniques

Gomez-Mejia et al (1982) rated management positions using seven job evaluations systems. Similar ratings were reported across methods. "It appears that, given a common job-analysis data base and data-collection tool, various methods used to transform data into a grade prediction can yield essentially comparable results."(p. 806). Madigan (1985) reported that the same job rating decisions were reached across three job evaluation methods. Consequently, both Gomez-Mejia et al (1982) and Madigan (1985) found that job evaluation method apparently does not contribute to job evaluation bias.

Factors/Scales

As noted earlier, Doverspike et al (1983) found little variance due to raters. Scales and jobs and their interactions produced most of the variance reported. Doverspike et al also calculated confidence intervals for scales and jobs and found them to be relatively large. To the authors, this result suggests that a replication of the study or the use of different jobs or scales could produce different results. However, the literature suggests that four factors (skill, effort, responsibility, and working conditions) do seem to be reliable across studies.

These results seem to indicate that the popular job evaluation methods have the capability to distinguish between jobs and that factors/scales do measure different dimensions. Certainly, one would want a job evaluation method both to differentiate between jobs and to have independent scales.

Doverspike and Barrett (1984) conducted a more detailed analysis of the impact of scales as they relate to possible sex bias in job evaluation results. While they did find high rater reliability, they also reported that reliabilities across scales were not the same. Moreover, (1) the correlations between scale scores and total scores varied across sexes; (2) a factor analysis produced a different set of factors for males versus females; and (3) partial correlations between scale totals and sex group scores produced some scales that favored males and some that favored females. Their results are particularly noteworthy in that, while raters provided reliable results, some scales proved to be biased toward one sex or the other.

Worthy of separate mention is the interpretation provided by Doverspike and Barrett (1984) for scale-sex interaction. They suggested that for male sex-typed jobs complexity of interactions with things was associated with higher skill demands while for female sex-typed jobs greater interaction with people was associated with lower skill demands. "Thus, the worth or meaning of interactions with people and things differed for male and female sex-typed jobs"(p.657).

Literature Summary

Obviously, the job evaluation methods that have seemingly stood the test of time are now coming under closer scrutiny. Rater reliability may be acceptable, but classification hit rates may be unacceptable. Moreover, traditional methods of measuring reliability may overstate rater reliability. Also, the concept of what is an acceptable error rate in classification has not been addressed.

The factors that can be used in job evaluation has reached some consensus; skill, effort, responsibility, and working conditions. The factors themselves do not appear to be biased. However, subfactors (i. e., scales) may not produce similar results for all jobs. Scales that emphasize interacting with things may favor the job evaluation of male jobs while scales that emphasize interacting with people may produce lower scores for female jobs. Obviously, if such scales are not equally represented in a job evaluation method, bias would result.

Implications of Job Evaluation Research for Job Analysis As Conducted in the Armed Services

The natural question arises as to why research on civilian job evaluation methods would be of interest to the Armed Services. Perhaps the most compelling reason is that DoD has always stated that it supports Affirmative Action and Equal Employment Opportunity programs in spirit regardless of whether or not DoD is actually included in Civil Rights legislation. Obviously, comparable worth is an ever expanding part of the civil rights movement. Given that job evaluation and its impact on comparable worth is coming under closer scrutiny in the civilian sector, then the Armed Services should also be interested in these issues. The fact that women are playing an ever expanding role in the Armed Services only adds to this argument.

Another reason is that the Armed Services make extensive use of job analysis data. An interesting trend of the job evaluation research is that it is leading more and more to the issues of the validity and reliability of job descriptions and the underlying data, often job analysis data, that generate these descriptions.

Or, as stated by Gomez-Mejia et al, "... the instrument that is used in gathering job analysis data is a critical element in building a valid and practical job evaluation system"(p. 806).

There are other reasons why the Armed Services should be interested in the results of job evaluation research. First, within the Air Force for example, promotion rates for enlisted personnel are heavily influenced by the results on skill knowledge tests, tests whose questions are directly linked to the task inventories generated in the job analysis process. Second, job descriptions are an important part of the military personnel structure. Since the on-going research in job evaluation is heading towards a more definitive study of job descriptions, the military should also be interested in any validity and reliability research done on job descriptions. Third, we are beginning to witness the inclusion of corresponding civilian job incumbents in job surveys of military specialties. Job evaluation research would be of interest to GS employees. Fourth, from an academic view, the military personnel R & D community should be interested in what the current developments are in the civilian sector both to stay abreast of current developments and also to have the opportunity to add to the investigation of the issues. Related to this fourth reason is the uniqueness of the military environment. Since the military does not have the usual `wage contamination' found in civilian jobs, it may be able to conduct purer research than is possible in the civilian community.

Suggestions as to the direction military research should take start with the job analysis system (CODAP) because the task inventories used in the CODAP system are the basis for much of the military classification system. The first suggestion is to re-analyze existing CODAP data using recently developed measurement techniques. For example, generalizability analysis could be conducted on job specialties which contain large numbers of female job incumbents. This analysis could take the form of treating individual tasks as job evaluation factor scales and determining the proportion of variance attributable to tasks, raters, factors, or sex group. Partial correlation could also be used to link task inventory responses to group membership. If the task inventories prove to be non-biased in terms of sex preference, then the next step would be to measure the validity of existing job descriptions by having judges independently evaluate jobs based on the current job description and based on the results of the underlying task inventory.

Given that job evaluation research has reported that there is a job by scale variance factor, task factor ratings, such as Job Difficulty and Training Emphasis, should be investigated as well as the number of tasks per job factor. These investigations should study the impact of job factors on job structuring in general as well as the impact, for example, of the job difficulty ratings received by male and female job incumbents.

The basic research goal should be to develop methodologies that can be used both to validate a job analysis task inventory and also to produce and validate subsequent job descriptions. In addition, the Armed Services personnel R & D community should also be able to provide more empirical data on job evaluation classification level confidence intervals and standard errors of measurement, two stated needs from recent research.

The question is not whether or not current evaluation techniques are working but whether or not they are biased. Current research suggests that there is bias and that this bias begins in either the job description or the data (often job analysis data) that generates the job description. Given the Armed Service's interest in job analysis, it should join in this research. The fact that there has not been much job analysis research done of late by the military is also not an issue; there had not been much job evaluation research done prior to 1979 either. The issue is simply that, just as we need measures of validity of any assessment technique, we need to develop measures of the validity of both (1) the task inventories used in job analysis and (2) the job evaluation methods themselves. The Armed Services can play a major role in this research.

References

Doverspike, D. & Barrett, G. V. (1984). An internal bias analysis of a job evaluation instrument. Journal of Applied Psychology, 69, 648-662.

Doverspike, D., Carlisi, A. M., Barrett, G. V., & Alexander, R. A. (1983). Generalizability analysis of a point-method job evaluation instrument. Journal of Applied Psychology, 68, 476-483.

Fraser, S. L., Cronshaw, S. F., & Alexander, R. A. (1984). Generalizability analysis of a point method job evaluation instrument: A field study. Journal of Applied Psychology, 69, 643-647.

Gomez-Mejia, L. R., Page, R. C., & Tornow, W. W. (1982). A Comparison of the practical utility of traditional, statistical, and hybrid job evaluation approaches. Academy of Management Journal, 25, 790-809.

Madigan, R. M. (1985). Comparable worth judgements: A measurement properties analysis. Journal of Applied Psychology, 70, 137-147.

Schwab D. P., & Grams, R. (1985). Sex-related errors in job evaluation: A "real world" test. Journal of Applied Psychology, 70, 533-539.

Stephenson, S. D. (1985). Job evaluation for the small business. A paper presented at the 1985 Southwestern Small Business Institute Association Meeting.

Treiman, D. J. (1979). Job evaluation: An analytic review. Washington, DC: National Academy Press.

Treiman, D. J. & Hartman, H. I. (Eds.) (1981). Women, work, and wages: Equal pay for jobs of equal value. Washington, DC: National Academy Press.