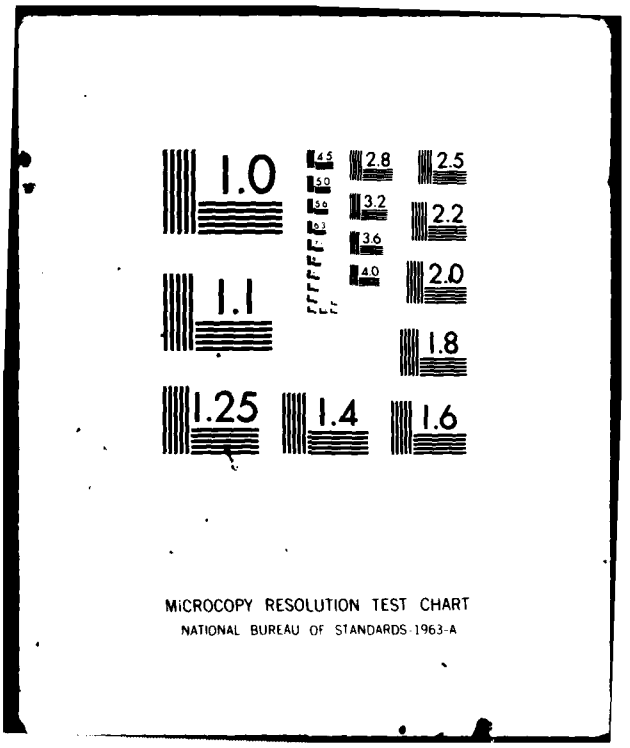
AD-A098 678 MILITARY TESTING ASSOCIATION F/G 9/10
PROCEEDINGS OF THE ANNUAL CONFERENCE OF THE MILITARY TESTING AS--ETC(U)
DEC 80
UNCLASSIFIED MTA-22-80-VOL-2 NL

1 of 6
AD98678




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD A098678

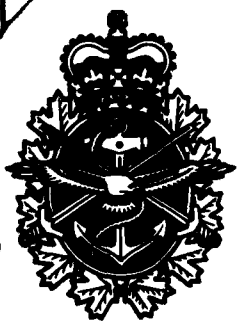
22nd ANNUAL CONFERENCE

PROCEEDINGS

MILITARY TESTING ASSOCIATION

LEVEL II

A098677





DTIC
MAY 21 1981

co-ordinated by

CANADIAN FORCES PERSONNEL APPLIED RESEARCH UNIT

Held in TORONTO, ONTARIO, CANADA

27-31 OCTOBER 1980

VOL. 2

DTIC FILE COPY

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

81 5 04 160

50272-101

REPORT DOCUMENTATION PAGE	1. REPORT NO. MTS-22-80	2. AD-A088678	3. Recipient's Accession No.

4. Title and Subtitle
Proceedings of the 22nd Annual Conference, Military Testing Association, Toronto, 27-31 October, 1980

5. Report Date
December, 1980

6. (date published)

7. Author(s)

8. Performing Organization Rept. No.
MTA-[illegible]-80

9. Performing Organization Name and Address
Military Testing Association
C/O Canadian Forces Personnel Applied Research Unit
4900 Yonge St. Suite 600, Willowdale, Ontario M2N 6B7

10. Project/Task/Work Unit No.

11. Contract(C) or Grant(G) No.
(C)
(G)

12. Sponsoring Organization Name and Address
CO, CFPARU
4900 Yonge St. Suite 600, Willowdale, Ontario M2N 6B7

13. Type of Report & Period Covered
Proceedings

14.

15. Supplementary Notes

16. Abstract (Limit: 200 words)

The Military Testing Association is open to members of the armed services of the United States, Britain, Canada and other allied nations, and to civilian employees of those armed services, who are employed in command, training, research and other activities involving the assessment of military personnel. Associate membership is available to civilians with parallel interests. The association meets annually to exchange information in the areas of behavioral measurement, occupational analysis, manpower analysis, simulation models, training programs, selection methodology, and survey techniques. The papers presented at the 22nd Annual Conference came from the military, government, educational and business communities of the United States, Canada, Britain, Australia, West Germany and Belgium.

17. Document Analysis a. Descriptors

testing, personnel, assessment, training, evaluation, performance, appraisal, achievement, measurement, ability, prediction, selection.

b. Identifiers/Open-Fnded Terms

c. COSATI Field/Group

18. Availability Statement
Release unlimited

19. Security Class (This Report) UNCLAS

20. Security Class (This Page) UNCLAS

21. No. of Pages 1095

22. Price

DO NOT PRINT THESE INSTRUCTIONS AS A PAGE IN A REPORT

INSTRUCTIONS

Optional Form 272, Report Documentation Page is based on Guidelines for Format and Production of Scientific and Technical Reports, ANSI Z39.18–1974 available from American National Standards Institute, 1430 Broadway, New York, New York 10018. Each separately bound report—for example, each volume in a multivolume set—shall have its unique Report Documentation Page.

1. Report Number. Each individually bound report shall carry a unique alphanumeric designation assigned by the performing organization or provided by the sponsoring organization in accordance with American National Standard ANSI Z39.23–1974, Technical Report Number (STRN). For registration of report code, contact NTIS Report Number Clearinghouse, Springfield, VA 22161. Use uppercase letters, Arabic numerals, slashes, and hyphens only, as in the following examples: FASEB/NS–75/87 and FAA/RD–75/09.
2. Leave blank.
3. Recipient's Accession Number. Reserved for use by each report recipient.
4. Title and Subtitle. Title should indicate clearly and briefly the subject coverage of the report, subordinate subtitle to the main title. When a report is prepared in more than one volume, repeat the primary title, add volume number and include subtitle for the specific volume.
5. Report Date. Each report shall carry a date indicating at least month and year. Indicate the basis on which it was selected (e.g., date of issue, date of approval, date of preparation, date published).
6. Sponsoring Agency Code. Leave blank.
7. Author(s). Give name(s) in conventional order (e.g., John R. Doe, or J. Robert Doe). List author's affiliation if it differs from the performing organization.
8. Performing Organization Report Number. Insert if performing organization wishes to assign this number.
9. Performing Organization Name and Mailing Address. Give name, street, city, state, and ZIP code. List no more than two levels of an organizational hierarchy. Display the name of the organization exactly as it should appear in Government indexes such as Government Reports Announcements & Index (GRA & I).
10. Project/Task/Work Unit Number. Use the project, task and work unit numbers under which the report was prepared.
11. Contract/Grant Number. Insert contract or grant number under which report was prepared.
12. Sponsoring Agency Name and Mailing Address. Include ZIP code. Cite main sponsors.
13. Type of Report and Period Covered. State interim, final, etc., and, if applicable, inclusive dates.
14. Performing Organization Code. Leave blank.
15. Supplementary Notes. Enter information not included elsewhere but useful, such as: Prepared in cooperation with . . . Translation of . . . Presented at conference of . . . To be published in . . . When a report is revised, include a statement whether the new report supersedes or supplements the older report.
16. Abstract. Include a brief (200 words or less) factual summary of the most significant information contained in the report. If the report contains a significant bibliography or literature survey, mention it here.
17. Document Analysis. (a). Descriptors. Select from the Thesaurus of Engineering and Scientific Terms the proper authorized terms that identify the major concept of the research and are sufficiently specific and precise to be used as index entries for cataloging.

    (b). Identifiers and Open-Ended Terms. Use identifiers for project names, code names, equipment designators, etc. Use open-ended terms written in descriptor form for those subjects for which no descriptor exists.

    (c). COSATI Field/Group. Field and Group assignments are to be taken from the 1964 COSATI Subject Category List. Since the majority of documents are multidisciplinary in nature, the primary Field/Group assignment(s) will be the specific discipline, area of human endeavor, or type of physical object. The application(s) will be cross-referenced with secondary Field/Group assignments that will follow the primary posting(s).
18. Distribution Statement. Denote public releasability, for example "Release unlimited", or limitation for reasons other than security. Cite any availability to the public, with address, order number and price, if known.

19. & 20. Security Classification. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED).

21. Number of pages. Insert the total number of pages, including introductory pages, but excluding distribution list, if any.

22. Price. Enter price in paper copy (PC) and/or microfiche (MF) if known.

(14) MTA-22-80-Vol-2

(6)

PROCEEDINGS of the

22ND ANNUAL CONFERENCE of the MILITARY TESTING ASSOCIATION (22nd)

held in Toronto, Ontario, Canada, 27-31 October 1980. Volume 2.

(11) Nov 80

(12) 523

co-ordinated by

CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT (CFPARU)
TORONTO, CANADA

DTIC ELECTE MAY 11 1981 S C

co-hosted by

DIRECTORATE OF MILITARY
OCCUPATIONAL STRUCTURES (DMOS)
NATIONAL DEFENCE HEADQUARTERS (NDHQ)
OTTAWA, CANADA

TORONTO DOWNTOWN HOLIDAY INN

27-31 October, 1980

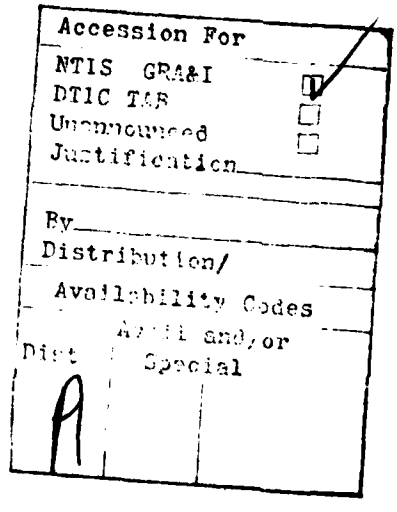
DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

404408

KANTOR, Jeffrey E., and IDEEN, Capt Dana R., Air Force Human Resources Laboratory, Brooks, AFB, Texas.

IMPACT OF STRESS IN AIR COMBAT: MODELS FOR PREDICTING PERFORMANCE (Thu A.M.)

Operations within an air combat environment are typically associated with subjective feelings of strain, pressure, or tension. These feelings are referred to as stress and can impact on aircrew performance. A Combat Stress Questionnaire was administered to 560 pilots with extensive SEA combat experience and from these data, the frequency of different fighter mission events, the stressfulness of those events, and individual estimates of maximum sortie rate were identified. Four models for predicting maximum sortie rate were developed using frequency and stressfulness of combat events and these models were evaluated using linear regression analyses. Results of these analyses and the implication for predicting combat performance are discussed.



THE IMPACT OF STRESS IN THE AIR COMBAT ENVIRONMENT: Models for Predicting Performance

Jeffrey E. Kantor and Captain Dana R. Ideen

United States Air Force Human Resources Laboratory

The combat mission of a fighter pilot is a stressful, complex, and highly demanding task. While there are many different theories regarding the various factors which affect performance in combat, one aspect which is widely cited (e.g., Youngling, Levine, Mochamuk, and Weston, 1977) is the pilot's ability to manage the psychological stress inherent in the combat arena. Psychological stress is being defined here as the pilot's subjective feelings of strain, pressure, or tension experienced in conjunction with the events of a combat mission. Research and anecdotal reports indicate that these feelings are common to almost all personnel in combat and are typically not moderated but rather increase with the individual's length of exposure to the combat setting (Shaffer, 1951).

Although flying combat has long been recognized as stressful, only a few scientific studies have investigated the relationship between stress and performance in air combat. Roff (1950) surveyed a group of aviators at the conclusion of their combat tour and reported that good pilots were described as having a greater tolerance for stress than poorer pilots. Strawbridge and Kahn (1955) found in one group of pilots a significant correlation between aviation combat performance scores and an aiming stress test. Finally, Austin (1969), using a urine analysis test for chemical indicators of stress, reported an increase in stress associated with combat flying and some indication that stress levels varied proportionately with levels of responsibility within the combat flights. In general, these studies appear to confirm that stress can be attributed to combat flying and that there appears to be some relationship between stress and performance.

Though not yet confirmed empirically, it is likely that the relationship between stress and performance in air combat follows the more general stress-performance relationship described by the Yerkes-Dodson Law, i.e. that the more complex the task, the more performance is disrupted by stress. Since piloting a high performance aircraft to a target, avoiding enemy threats, operating a sophisticated weapons delivery system, and returning a possibly battle damaged aircraft to base is an extremely complex task, then pilot performance in the combat arena should be particularly sensitive to stress induced decrements. To study various aspects of the relationship between stress and performance in air combat, a data base was developed from the most recent combat experiences of American fighter pilots. The objective of this study, one of several using that data base, is to develop a mathematical model which describes and predicts the impact of stress on air combat performance.

METHOD

Subjects

The subjects in this study were 563 members of the Red River Valley Fighter Pilot's Association. Membership in this organization is limited to aircrew who have flown combat missions into North Vietnam past the Red River,

an area which included Hanoi and other heavily defended targets. Members of this organization have accrued the most recent American operational theatre air combat experience.

Development of the Combat Stress Data Base

A Combat Stress Questionnaire developed by Kantor, Klinestiver, and McFarlane (1978), was used to build the data base to assess the impact of stress on combat performance. The main body of this questionnaire is a listing of specific events which a fighter pilot might encounter during any one combat mission flown against a heavily defended target. The list of events was developed through consultation with combat experienced fighter pilots and included such events as mission briefing, cockpit checkout, engine start, weapon armament, different types of take offs, aborts, refueling, flying different types of formations and different slots within the flight, threats from ground and air sources, excessive G maneuvers, different weapon deliveries, aircraft damage, different types of landings, barrier engagements, debriefing, and film assessment. After these specific events, there was a list of the four types of fighter missions flown most commonly in SEA: interdiction, ground support, air cover, and air-to-air interception. The respondent was asked to indicate, from his flying experience in Southeast Asia, the frequency of occurrence of each mission event (on a 1-5 scale; 1=Never/Almost Never - 5=Always/Almost Always) and the stressfulness of each event if it did occur (also on a 1-5 scale, 1=No stress - 5= Very intense Stress). Similarly for the four mission types, the respondent was asked to indicate how often he flew each type and how stressful, overall, that type of mission was. After this section of the questionnaire, the respondent was asked to indicate what type aircraft he flew, his amount of flying experience prior to combat, his total amount of flying time in fighters, his total combat flying time, and his age at combat entry. Finally, there were eight items which obtained the respondent's estimate of the number of missions that a single pilot could fly effectively in a two week period under kind of conditions encountered on missions flown into North Vietnam. These eight items were the four mission types mentioned above, but were separated into short duration (less than 3 hours) and long duration (4-6 hours) missions.

Statistical Analysis

Multiple linear regression analysis was used to quantify the impact of stress on air combat performance and identify a model which best predicted that impact. Multiple linear regression is a statistical technique often useful in assessing the relationship between various independent, or predictor variables, and a dependent, or criterion variable. In this application, the frequency of specific mission events, their stressfulness, the frequency of mission types and their stressfulness were considered potential predictors of air combat performance. While there are many indices of air combat performance, the respondent's estimate of the number of missions which could be flown effectively in two weeks (averaged across all types) was used as the criterion for this study. This is an important index of combat performance since many scenarios of future wars involve the "sortie surge" concept where

(524)

each squadron launches as many missions as possible within a relatively short period. Therefore, the use of these predictors and a criterion of estimated sortie surge rate should provide some insight relevant to the performance of aircrews in future war settings.

Results and Discussion

Using the frequency and stressfulness data regarding mission events and types, four models were built and compared for predictive validity against the criterion of the number of missions estimated that a pilot could fly in two weeks. Model 1 consisted of the frequency of mission events and types while Model 2 was composed of the stressfulness of those mission events and types. Model 3 consisted of the interaction of frequency and stressfulness for each mission event and type. Finally, Model 4, hereafter referred to as the full model, was composed of the frequency, stressfulness and interaction of frequency and stressfulness for each mission event and type. A multiple linear regression analysis using these models is summarized in Table 1. All four models were significantly ($p < .05$) related to the criterion. The model having the greatest predictive validity was the full model which accounted for 37% of the criterion variance (multiple R=.61).

Table 1

	Predictors		Multiple Correlation		F Test		
	Type	Number	R	R^2	R^2 vs 0		
					Df_1	Df_2	F ratio
Model 1	Frequency (F)	48	.38	.14	48	514	1.80*
Model 2	Stressfulness(S)	48	.38	.15	48	514	1.88*
Model 3	(FXS)	48	.35	.13	48	514	1.49*
Model 4	(F), (S), (FXS)	144	.61	.37	144	418	1.69*

Table 2

Full Model		Restricted Model		F Test		
Predictors	R^2	Predictors	R^2	df_1	df_2	F Ratio
(F), (S), (FXS)	.37	(F), (S)	.25	48	418	1.67*

*$P < .05$

To determine whether or not the interaction component in the full model was necessary, i.e. was significantly accounting for unique variance beyond

the simple effects of frequency and stressfulness, a comparison was conducted between the full model and a restricted model having only frequency and stressfulness simple effects. This comparison is summarized in Table 2. The significant ($p < .05$) F ratio obtained by comparing the full and restricted model indicated that significant criterion variance is accounted for by the interaction of frequency and stressfulness. Therefore, the full model best predicts the number of missions that a pilot could fly in two weeks.

To help describe the relationships defined by the full model, a subset of predictors was selected through statistical testing of the regression weights. All predictor regression weights were tested against zero and 16 predictors were identified as having statistically significant ($p < .05$) non-zero weights. These predictors and the sign (" + " or "-") of their weights are presented in Table 1. In describing how these factors impact on the pilot's sortie rate estimate, it would appear that as the frequency and stressfulness of arming live ordnance increases, the number of missions a pilot is able to fly in two weeks decreases. Similarly for exposure to +4.5g, as the frequency and stressfulness of this aspect of flying increases, the estimated number of sorties decreases. Other factors which had a direct negative impact on the mission estimates were the increasing frequency of minor battle damage and the increasing stress associated with air aborts. Conversely though, increasing the frequency of flying as element leader, increases the pilot sortie estimate, as does an increase in the stress associated with making one pass at ground target. The other factors involved an interaction component and therefore, their impact on sortie estimates is dependent simultaneously on both the frequency of occurrence and the stress associated with that event. In other words, the impact of the stressfulness of these events varies with how often they actually occur. Specifically, for cockpit checkout; day, instrument take off; flying as wingman, and flying interdiction type missions, the impact on mission rate is dependent on the catalytic interaction of frequency on stress.

Table 1

SIGNIFICANTLY WEIGHTED PREDICTORS

FREQUENCY	STRESSFULNESS	INTERACTION
Weapon Armament (-)*	Weapon Armament (-)	Cockpit Checkout (-)
Day, Instrument Take-off (+)		Day, Instrument Take off (-)
Flying as Element Leader (+)	Air Aborts (-)	
Flying as Wingman (-)		Flying as Wingman (+)
Exposure to +4.5g (-)	Exposure to +4.5g (-)	
	Air to Ground Exposure-1 pass (+)	
Minor Battle Damage (-)		
Flying Interdiction Missions (+)	Flying Interdiction Mission (+)	Flying Interdiction Missions(-)

* The sign within the parentheses indicates the sign of the regression weight for that predictor.

526

It is interesting to consider that the factors which were identified as having significant relationships with how many missions a pilot could fly did not include the more life threatening events which might occur during a combat fighter mission. Such factors as the frequency and stressfulness of encountering enemy interceptors, surface-to-air missiles, anti-aircraft artillery, and sustaining aircraft damage requiring either emergency action or a precautionary landing were not among the predictors which had non-zero weights. This would seem to indicate that regardless of the frequency or stress associated with these events, they do not play a part in the pilots' estimate of sortie rates.

Illustrations From the Model

To illustrate the relationships defined by the model, the effect of stress was computed for the stress experienced in association with the mission event "weapon armament." Weapon armament occurs just prior to take off, when the safety wires, etc. are removed from the ordnance which the aircraft will deliver to the target. The relationship between the stressfulness of this event and mission sortie estimates is plotted in Figure 1. Holding other values constant, an increase in one unit measure of stress reduced the estimate of missions by 2.2. Therefore, the stress experienced by the pilot during weapon armament had a substantial negative impact on that pilot's estimate of how many missions he could fly in two weeks. Interpretation of this relationship on a logical level is somewhat ambiguous. Weapon arament is not a dangerous phase of a combat mission and the arming of the weapons themselves is accomplished by a trained ground crew. Therefore, why this stress should be so closely tied to mission estimates is unclear. Several explanations are possible. It might be that the relationship expressed here is tapping the overall stress associated with combat since in the act of arming live ordnance the "raison d'etre" of combat flying is crystalized. The weapon armament stage also occurs just prior to take off and might reflect a commitment point, since as soon as the weapons are armed, the mission commences. Finally, the relationship here might also be assessing the impact of the preparatory stress response. For many high threat activities (e.g. parachuting) immediately prior to the activity, stress rises and the level of stress experienced appears somewhat related to success at that activity (Fenz, 1975; Ursin, Baade, & Levine, 1978). Future research will attempt to assess this relationship more fully.

To illustrate a frequence by stress interaction effect on performance in the air combat environment, the expected sortie rate under varying frequencies and loads of stress associated with flying interdiction missions was determined. As the significance of the interaction term implies, the impact of stress varied with how often the pilot flew interdiction missions. These relationships are plotted in Figure 2. This findings seem to indicate that, for pilots who flew interdiction missions less than half the time (frequency=1 or 2), the relationship between stress and sortie estimates was positive. However, for pilots who flew mainly interdiction mission (frequency=3,4, or 5), the relationship was negative. This is particularly true for pilots who flew interdiction missions almost always. In that case, for each unit increase in stress, the mission estimates were reduced by 1.2. Perhaps for those pilots who flew interdiction missions only rarely, the

stress associated with them represented a challenge while for those pilots whose primary mission was interdiction, the stress associated with that type mission was a daily obstacle to overcome, and the more stress, the less willing they were to fly more missions.

Finally, an interesting relationship also identified by the model concerned the frequency of flying as element leader. This relationship is plotted in Figure 1. Here, the more frequently a pilot flies as leader, the more his sortie rate estimate increases. For each unit increase in frequency, the sortie rate estimate increases by 1.35. Evidently, the opportunity to fly as element leader functions as a motivator for pilots. Within operational constraints, this opportunity might be manipulated to increase sortie rates.

Conclusions

There is a significant relationship between the stress experienced in association with combat flying and performance in the air combat environment. Using a criterion of sortie rate estimate for flying against heavily defended targets, it was found that a model consisting of the frequency, stressfulness, and interaction of frequency and stressfulness of mission events and types demonstrated substantial predictive validity. Additionally, the model proved useful in identifying specific factors which had positive and negative impacts on sortie rate estimates and provided insights into the combat flying experience.

Overall, the information generated in this study should prove beneficial both to the operations researcher and the tactical resource manager. Since stress is related to performance in the air combat environment, more work in this area should be done to provide a more complete understanding of this relationship. From the results of this study, it would appear that the procedures described here represent one approach to that goal.

References

Austin, F.H. A review of stress and fatigure monitoring of naval aviators during aircraft career combat operations: Blood and urine biochemical studies. In P.G. Bourne (Ed.), The psychology and physiology of stress. New York: Academic Press, 1969.

Fenz, W.D. Strategies for coping with stress. In I.G. Sarason & C.D. Spielberger (Eds.), Stress and Anxiety (Vol. 2). New York: Hemisphere (Wiley), 1975.

Kantor, J.E., Klinestiver, L., & McFarlane, T.A. Methodology to assess psychological stress and its impact in the air combat environment (AFHRL-TR-78-3). Brooks AFB, TX: Air Force Human Resources Laboratory, March 1978.

Roff, M.F. A study of combat leadership in the Air Force by use of a rating scale (ATI-72-052). Randolph Field, TX: USAF School of Aviation Medicine, February 1950.

528

Shaffer, L.F. "Fear in combat and its control" Washington, D.C.: The working Group on Human Behavior Under Conditions at Military Service, Department of Defense, Research and Development Board, 1951.

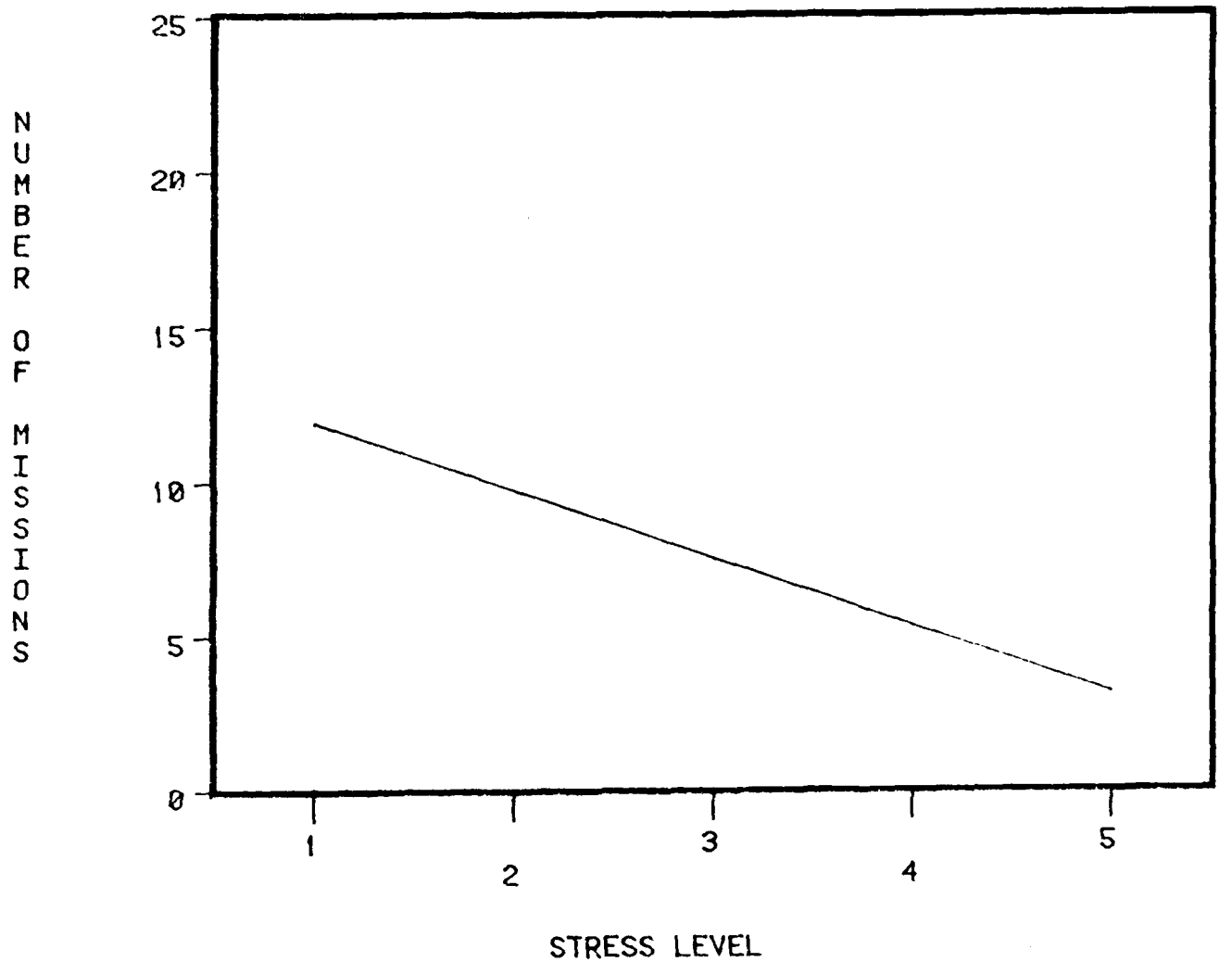
Strawbridge, D., & Kahn, N. Fighter pilot performance in Korea (IAWR Report No. 55-10). Chicago, Ill: Institute for Air Weapons Research, November, 1955.

Ursin, H., Baade, E., & Levine, S. Psychobiology of stress: A study of coping men. New York: Academic Press, 1978.

Youngling, E.W., Levine, S.H., Mochamuk, J.B., & Weston, L.M. Feasibility study to predict combat effectiveness for selected military roles: Fighter pilot effectiveness (MDC E1634). St. Louis, MO: McDonnell Douglas Corporation, April 1977.

FIGURE 1.

EFFECTS OF STRESS DURING WEAPON ARMAMENT



KA-8

FIGURE 2.

EFFECTS OF STRESS ASSOCIATED WITH INTERDICTION MISSIONS

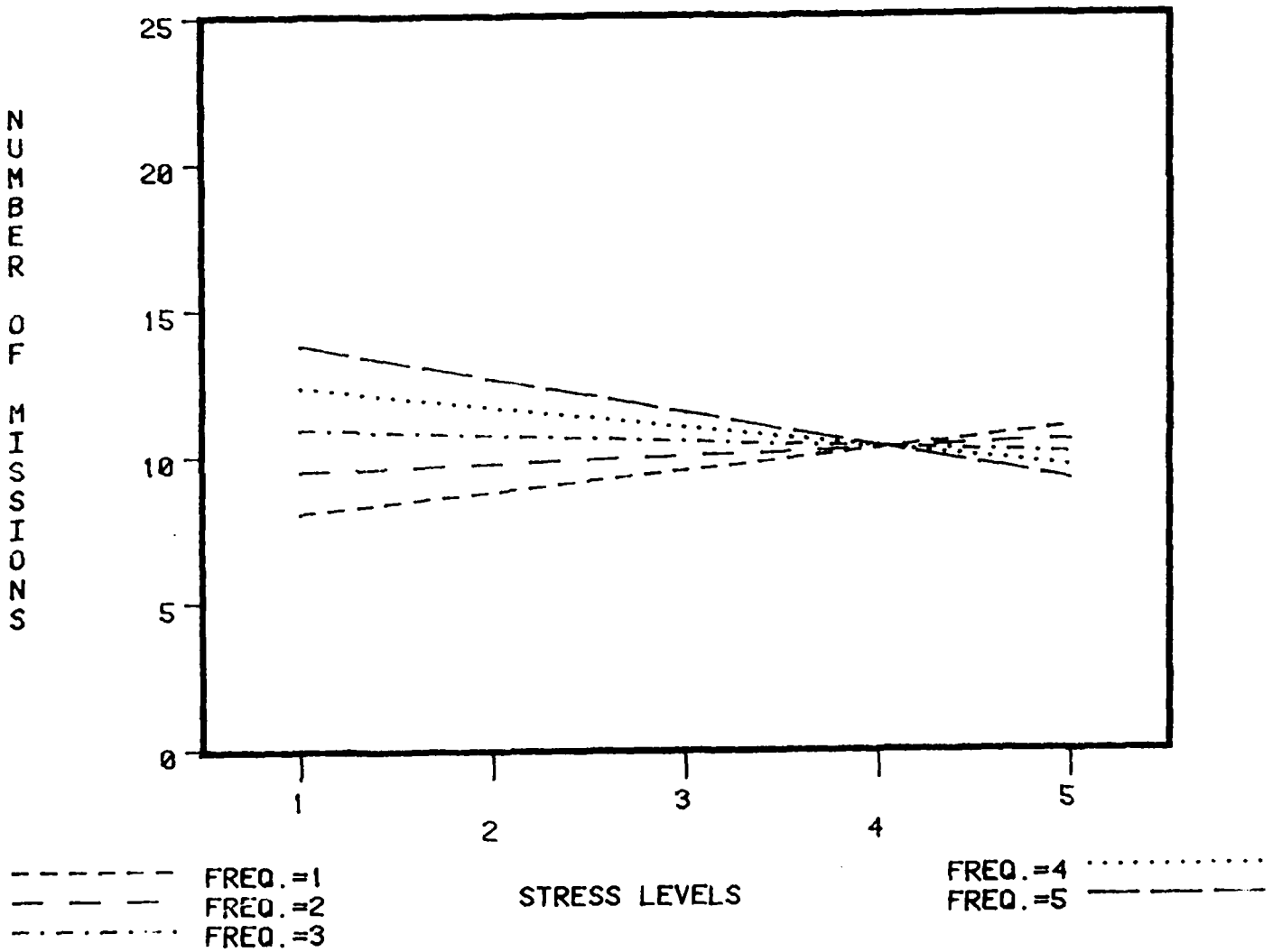
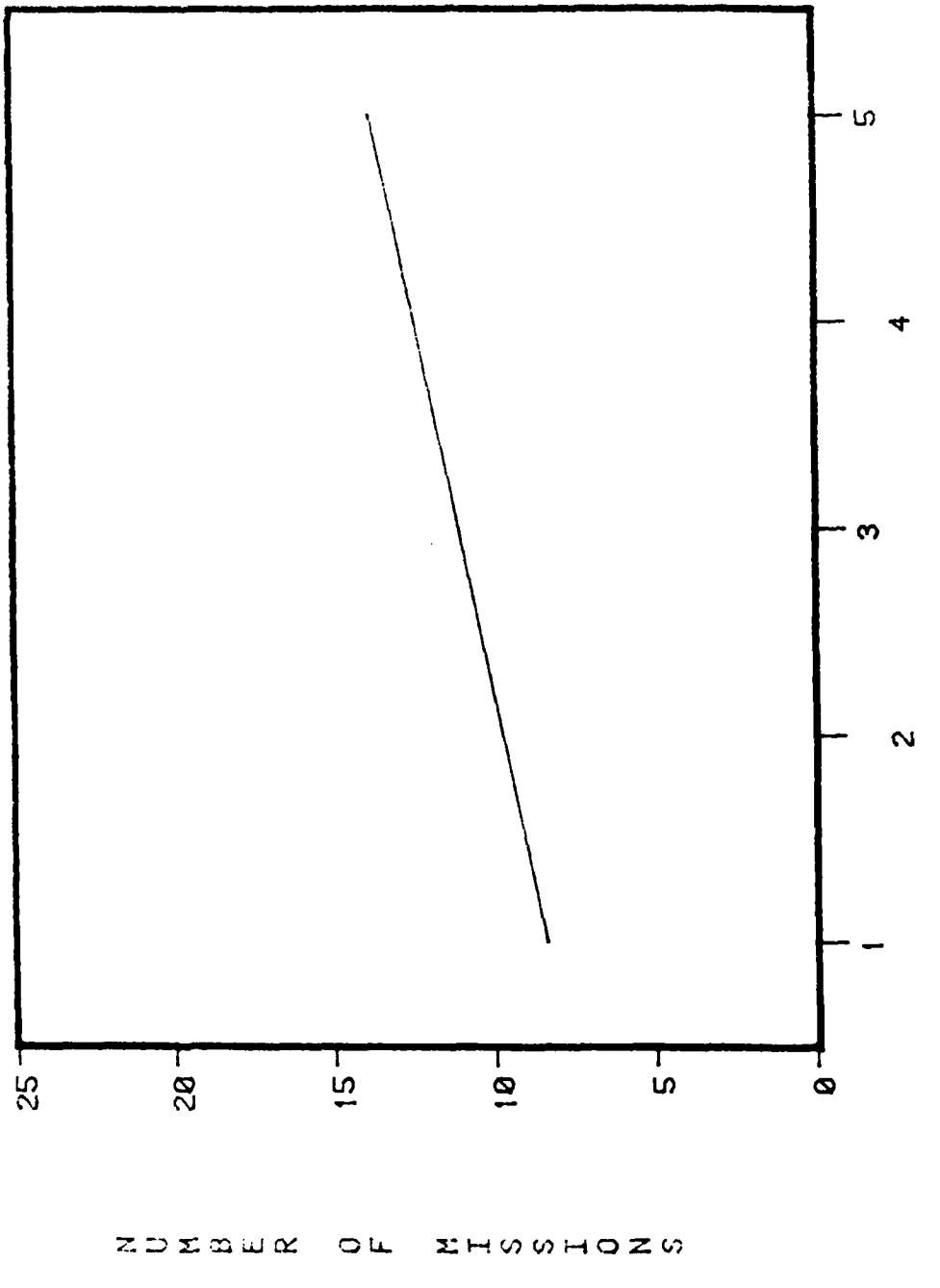


FIGURE 3.

EFFECTS OF FREQUENCY: FLYING AS ELEMENT LEADER



5-32

KELLETT, Capt. R.G., and PIGEON, R., National Defence Headquarters, Ottawa, Canada.

EFFECT OF ITEM DIFFICULTY INFORMATION ON MULTIPLE-CHOICE ACHIEVEMENT TEST PERFORMANCE (Wed A.M)

Researchers have found that test scores are influenced by individual predispositions to respond to questions in a given way and have termed the effect "response set". Assuming the existence of a response set such that many students read test questions too quickly and select the first response which seems plausible, Huck (1978) found that students altered their test-taking strategy and scored better on a three-form achievement test if told how hard each item had been for previous students.

For a homogeneous group of test-wise subjects writing difficult multiple-choice achievement examinations, the effect on test scores of providing item difficulty information was experimentally examined. Contrary to Huck's findings, score improvement did not result when subjects were given item difficulty information. In discussing the findings of the experiment, this paper suggests that the absence of a significant effect on test scores from providing item difficulty information reflects a situation in which test-wise subjects do not readily alter their test-taking strategies even when considerable advice and direction are included in the test instructions.

EFFECT OF ITEM DIFFICULTY INFORMATION ON MULTIPLE-CHOICE ACHIEVEMENT TEST PERFORMANCE

INTRODUCTION

1. In most educational achievement tests and examinations the objective has been to measure the level and type of knowledge or skill possessed by the examinees, exclusive of the influence of known or unknown extraneous factors such as physical circumstances of the testing situation, effects of test construction on results, emotional and physical condition of examinees at the time of testing, and individual differences in test-taking strategy. The first three areas have been studied in detail and considerable effort devoted to standardizing test construction and testing situations. However, the reduction in test validity resulting from individual differences in test-taking practices has received less attention, and has had little impact in the past upon educational test design (Thorndike, 1971).

2. Researchers have found that test scores are influenced by individual predispositions to respond to questions in a given way and have termed the effect "response set". Cronbach (1946) defined a response set as "any tendency causing a person consistently to give different responses to test items than he would when the same content is presented in a different form .... in this definition, 'form' includes the form of statement, the choice of responses offered, and the directions, since all of these are part of the situation to which the person reacts."

3. Various types of response sets have been hypothesized and evidence of their operation on certain measurement instruments have been accumulated. In personality tests, the emergence of response set has sometimes been stimulated because, as personality traits or correlates of traits, response set can aid in the identification and definition of personality characteristics. However, in tests of aptitude and achievement response sets have been found to dilute a test with factors not intended to form part of the test content, and so reduce its logical validity (Cronbach, 1950). Also, they have reduced the empirical validity of tests and artificially broadened or narrowed the range of individual differences in score. Thus, response sets have affected test reliability, falsely heightening reliability where they are consistent and lowering reliability where they have reduced score range. Ability to make inferences from test data has been adversely affected by response sets. Since response sets often altered the percentage of examinees passing an item, judgements about item difficulty must have been distorted. Changes in examinees' scores, originally attributed to growth in knowledge or change in interest, might have resulted instead from a change in caution, avoidance of extreme response positions, increased test-wiseness, or shift in some other response bias.

4. Even where response sets merely increase true test variance, they alter the meaning of test scores, and so make interpretations from scores questionable. According to Guilford (1954), we should strive to get rid of response sets even if to do so lowers reliability. Tests will be better diagnostic instruments even if bias variance is replaced by error variance, as the latter is not focused systematically in any direction while bias variance is.

5. Huck (1978) has suggested that the inclusion of item difficulty information (known item difficulty indices) on an examination leads to improved scores by counter-balancing negative test-taking sets such as reading questions too quickly and selecting the first plausible answer. Indeed, altering test instructions has long been known to alter examinees' judgements (Goodfellow, 1940; Gault and Goodfellow, 1940), and to alter markedly the influence of response sets (Cronbach, 1950).

6. In this empirical study, hypotheses concerning the operation of response set on multiple-choice examinations and the effects of including item difficulty information as a counter-measure were tested. A statement of the hypotheses, and description of subjects, methodology, instruments and design for data analysis used to test the hypotheses follow. Experimental results, discussion and conclusions will then be presented.

RESEARCH HYPOTHESIS

7. Response set has been defined as any tendency (habit or momentary set) which causes a person to give consistently different responses to test items than he would give if the same content were presented in a different form. Researchers have shown that where response set occurs, it is reliable and stable over time and over varying situations (Gustav, 1963; Berg, 1953; Berg and Rapaport 1954; Rapaport and Berg, 1955; Rundquist, 1950). It has been concluded that regardless of test form, instructions greatly affect the emergence of response set and the type of set that occurs (Goodfellow, 1940). The difficulty level of individual test items and entire examinations has been found to be a prime factor stimulating the onset of response set (Cronbach, 1950). The results of studies designed to demonstrate that the location of the correct response, or best distractor, on a multiple-choice test affects item difficulty due to the operation of positional response bias, have been inconclusive. The weight of evidence has indicated that multiple-choice item difficulty is not greatly influenced by position of either the correct response or best distractor (Marcus, 1963; McNamara and Weitzman, 1945; Clark, 1956). Similarly, the contention that the difficulty or ease of earlier items

alters subjects' responses to later items has been generally shown to be untrue for reasonably difficult, untimed multiple-choice tests when subjects are "test-wise" (Sax and Carr, 1962; Huck and Bowers, 1972).

8. One can conclude that Cronbach's (1950) argument that the multiple-choice test form is immune to the influence of response set may perhaps hold true for a positional response bias, a rather simplified, almost automatic tendency to select or avoid a given choice or choices (Hopkins and Hopkins, 1964; Wilbur, 1970). While there is some evidence that even such simple positional bias's may occur when multiple-choice tests are extremely difficult, there is definitely a strong indication that more complex sets or bias's in the cognitive domain influence multiple-choice test results (Gaier, Lee, and McQuitty, 1953). In fact, Huck's (1978) experiment demonstrates that such complex bias's do alter test results and have distorted item difficulty values.

9. In view of the foregoing discussion, it was hypothesized that for a homogeneous group of test-wise subjects the inclusion of item difficulty information on difficult multiple-choice achievement examinations would result in improved scores. In effect, score improvement was expected to result from the elimination of a response set. Further, it was hypothesized that score improvement would be greater on the more difficult items than on the less difficult ones. A precise expectation concerning the effect, if any, on test reliability of including item difficulty information could not be formulated, but in the study the question was addressed.

METHODOLOGY AND DESIGN

Research Subjects

10. Canadian Forces officers in the ranks of lieutenant to major (from one to fourteen years service in commissioned rank) who were commissioned in 1971 or later are required to take six professional knowledge, self-study courses deemed to be fundamental to their professional development. Successful completion of the six courses is a selection prerequisite for further professional development training. The six courses are:

OPDP 2 - General Service Knowledge
OPDP 3 - Personnel Administration
OPDP 4 - Military Law
OPDP 5 - Financial Administration and Supply
OPDP 6 - National and International Studies
OPDP 7 - War and the Military Profession

11. The Officer Professional Development Program (OPDP) has been a part of the officer development system for five years. Close to 20,000

confirmatory examinations, which is the sole means of determining success or failure in a course, have been written in the six courses since program inception. Over 800 officers have graduated from the program, that is, have completed all six required courses. At the present time approximately 5,700 officers are mandatory program participants, out of which some 2,800 wrote examinations this year. Examinations were written at over ninety centers throughout Canada and in Europe.

12. The hypotheses were tested using as subjects officers who wrote the OPDP examinations during the normal 1980 administration sessions at the two examination centers in Ottawa. A total of 186 officers wrote examinations as treatment group subjects and 170 officers formed the control group. Only officers writing in Ottawa were used as subjects because costs and travel restrictions prohibited the administration by the researcher of experimental examination versions at other centers. In fact, mandatory officers employed in the Ottawa area are representative of the target population (5,700 OPDP mandatory officers throughout the Canadian Forces) in terms of background, occupational groupings, training, experience, and current employment. Moreover, examination performance of Ottawa officers has not, in the past, differed significantly from the performance of officers at other locations, as shown by Kellett and Pigeon (1980). The number of subjects used in the study represented about seven percent of the target population and twelve percent of the accessible population (3,200 officers who write examinations annually).

13. Canadian Forces officers are, as a group, homogeneous and "test-wise". Homogeneity stems from the application of a fixed standard in recruitment and selection of officers, commonality of training, and provision of a uniform lifestyle for the officer corps. Also, many officers spend the majority of their careers living and working at geographically isolated locations. Officers are considered to be test-wise because they are subjected to intensive multiple-choice testing from the commencement of their careers, and probably even before that.

Instruments

14. To measure the effect of item difficulty information on the multiple-choice achievement test performance of Canadian Forces officers who are mandatory participants in the Officer Professional Development Program, the 1980 bilingual examinations constructed to measure the knowledge level of officers in each of the six OPDP courses were used. Content validity of the 1980 examination forms has been established by the curriculum developers (content specialists) for each of the courses. Reliability coefficients (KR-20 values) of previous versions of the examinations, calculated from populations varying in number between 487 and 1044 officers, have ranged from .82 to .91. The

1980 examination forms contained from sixty to seventy-five percent of items common to previous versions. In the study mean difficulty indices for the common items were used where items had been included on examinations more than once previously. Historically, item difficulty values have ranged from .10 to .90 with mean value of .52.

15. Examinations in five of the six courses consisted of seventy-five, four - option items covering the course study material. The examination in OPDP 4 (Military Law) had seventy multiple-choice items and twenty of the true - false type. The latter part of the Military Law examination was not considered in this study.

16. In 1980 the confirmatory examinations were written in two sessions, one in May and one in June. Parallel forms of the same examinations were administered in the two sessions, with over eighty percent of the items being common to the two forms used for each course. Although there were two administration sessions, no candidate was permitted to write two examinations in the same course in the same year. Although officers may write one examination in any number of courses in a single year, the results for each course were analyzed and interpreted separately so independence of sample groups was ensured.

Research Methodology

17. To determine if inclusion of item difficulty information on the OPDP examinations resulted in improved scores, and if score improvement was greater on the more difficult items, a two-group (posttest only) approach was followed. Although pretesting is a standard practice in educational and psychological research, it is not essential if randomization in assignment of subjects to groups exists, as pointed out by Campbell and Stanley (1966). In essence, the practical circumstances of this study precluded the application of a pretest.

18. Subjects who wrote OPDP examinations at Ottawa during the first and second sittings were assigned at random to either the treatment group or the control group. The treatment group was given regular examinations in each course but with an additional sheet included citing the difficulty indices for examination items (where known). The control group wrote regular examinations without the additional sheets.

19. According to Campbell and Stanley (1966), the design used controls for all sources of internal invalidity, but leaves two sources of external invalidity as areas of concern. The first of these is the threat from interaction of selection and treatment. Due to the representativeness of Ottawa officers in terms of the target population, and the historical consistency of their results with the results achieved by all others writing in the same year, interaction of selection and treatment was not expected to present a problem. The second threat cited as a potential concern is the reactive effect of

experimentation rendering an experimental (treatment) group no longer representative of the group of subjects from which they were selected. Since the treatment in this study did not involve application of special testing, or noticeably different processes for those in the treatment group, it is believed that reactive arrangements did not reduce generalizability of findings.

Statistical Design and Analysis

20. For each of the six course examinations, a one-way analysis of variance (ANOVA) was performed to determine if examination scores obtained by the treatment group differed significantly from those obtained by control group subjects. Since the number of subjects in each of the two groups was sufficiently large, the ANOVA was considered to be sensitive enough to detect real differences in scores, if they existed, attributable to treatment. To test the second hypothesis, examination results for subjects in both groups were rescored on two subtests: (1) the least difficult forty percent of items for which difficulty indices were cited, and (2) the most difficult forty percent of items for which difficulty indices were cited. For each subject a "difference score" (absolute value of subtest one score minus subtest two score) was computed. The difference scores for subjects in the treatment and control groups were subsequently compared using ANOVA. If the second hypothesis was true, that is, score improvement under treatment was greater on the more difficult items than on the less difficult items, then the treatment group difference scores would be significantly smaller than the control group difference scores.

21. To determine if test reliability was altered by the application of the experimental treatment, KR-20 reliability coefficients were calculated for each course examination (using first session results only) under the two conditions, with and without treatment. The two coefficients obtained were then compared using a Fisher "Z transformation", as recommended by Dayton and Stunkard (1971), to ascertain if any difference between them was significant.

22. In an attempt to establish if treatment group subjects actually used the additional sheet of instructions containing item difficulty indices, subjects in the treatment group were asked to indicate on the additional sheet if they used the indices in considering their responses to examination items. Affirmative and negative responses were tallied; however, subjects could not be identified individually according to their response, so that ANOVA could not be performed separately for subjects who used the item difficulty information as opposed to those who did not.

5-40

RESULTS

23. It was noted earlier that for each research subject a total examination score was obtained using only those items for which difficulty indices were given. As well, each subject was scored on the easiest forty percent of items (easy subtest) and on the most difficult forty percent of items (difficult subtest). A difference score was computed for each subject as the absolute value of "easy subtest" minus "difficult subtest" score. Treatment group (those who were given item difficulty information) and control group (those not given item difficulty information) performance was then compared for each set of scores (total examination, easy subtest, difficult subtest, difference score) using the SPSSM subprogram ANOVA.

24. ANOVA tables for the four sets of scores were generated separately for each of the six OPDP courses. Sources of variance in each case were "group effect", which is the difference in mean scores attributable to the presence or absence of item difficulty information, and "error variance". For all analysis of variance significance level (α) was set at 0.05.

25. Out of twenty-four F-values obtained in the analysis of variance just three were significant at the designated level of significance. Treatment group subjects achieved significantly better results than control group subjects in terms of total examination and difficult subtest scores for OPDP course 3. Treatment group performance was also superior on the difficult subtest for OPDP course 5. Treatment and control group scores were not significantly different in all other cases.

26. The KR-20 reliability coefficients for each course examination were calculated under the two conditions of treatment (difficulty indices given) and no treatment (control). KR-20 values were obtained using SPSSM Subprogram Reliability. For each OPDP course, reliability coefficients for treatment and control group scores were compared using a Fisher "Z transformation". It was found that examination reliability for all OPDP courses was not significantly different for treatment and control group subjects at a significance level (α) of 0.05.

27. In surveying treatment group subjects with respect to usage of the item difficulty information, it was found that 62.5 percent used the indices in considering their responses to items, while 37.5 percent did not use the indices.

DISCUSSION

28. In general, the results of this empirical study do not support the hypothesis that inclusion of item difficulty information on

difficult multiple-choice achievement examinations results in improved scores. In four out of the six OPDP courses treatment group subjects did not obtain significantly higher scores than control group subjects. For the two courses where significant differences were observed, it was on the most difficult forty percent of items that treatment subjects were superior to control subjects.

29. The finding that inclusion of item difficulty information resulted in improved scores on the most difficult forty percent of items for two of the six OPDP courses can perhaps be explained in terms of type of subject matter of the two courses. In both courses students are introduced to considerable new terminology. Students have reported in the past that many key words confuse them. It is likely that when treatment group subjects encountered an item reported to be difficult, they re-read the item stem and sought to clarify the terms used before responding to the item. In such a case, the provision of item difficulty indices truly reduced or eliminated the effect of a response set, or tendency to respond to an item in a preconceived manner.

30. The absence of a significant effect on test scores from providing item difficulty information for four of six courses possibly reflects a situation in which test-wise subjects do not readily alter their test-taking strategies even when considerable advice and direction are included in the test instructions. Such a suggestion was made by Smith, White and Coop (1979) in relation to item changing behavior and the impact of advice to students about item changing. The survey taken in this study of use of the additional sheet containing the item difficulty indices revealed that fully 37.5 percent of treatment group subjects chose not to use the additional information. Had analysis of variance been conducted using only treatment group subjects who actually used the difficulty indices (and thus, in reality, received the "treatment"), the results obtained might have been different.

31. Once it was found that treatment and control group performance was not significantly different for four of the six OPDP courses, it was expected that there would not be a difference in examination reliability for the two groups. In fact, there was no significant difference in KR-20 reliability for the two groups on the examinations in the two OPDP courses where treatment group performance was superior to control group performance. This finding agreed with the results obtained by Huck (1978), who explained the consistency in KR-20 reliability maintained in the face of differential performance between treatment and control subjects by noting that one source of variance (response set) was probably replaced by some other source.

CONCLUSION AND RECOMMENDATIONS

32. Contrary to Huck's (1978) findings, this empirical study showed that providing item difficulty information did not generally result in score improvement. However, when the subject matter being examined on consisted of a high percentage of new terminology and concepts frequently confused by students, the provision of item difficulty indices led to better performance on the more difficult items. Moreover, it was discovered that students most often declined to use the additional information and showed great reluctance to alter their test-taking strategies even in the light of considerable advice.

33. A large percentage of subjects in this study who formed the treatment group (those who received additional sheets containing item difficulty indices) did not respond to the modified set of instructions. Consequently, it is believed that the study results are not conclusive because many in the treatment group did not really receive a "treatment". Thus, it is considered worthwhile to replicate the experiment but with the following modifications:

a. Item difficulty indices would be included next to each item in the examination booklets rather than on an extra sheet. This should help overcome the tendency of subjects to disregard the item difficulty information as extraneous.

b. Subjects in the treatment group who actually use the item difficulty indices would be identified and analysis of variance would be conducted separately for those subjects.

KNERR, Dr. C. Mazie, and MATLICK, Richard K., Litton Mellonics, Springfield, Virginia.

COST AND TRAINING EFFECTIVENESS ESTIMATION IN ARMY MATERIEL ACQUISITION
(Wed A.M.)

The Army's life cycle system management model (LCSMM) is designed to ensure consideration of all aspects of materiel costs, human resources, and missions during the acquisition process. Estimates of the cost and effectiveness of training provide decision makers with summary information that grows increasingly important as the costs of human resources rise. Cost and training effectiveness analyses (CTEA) are required at each phase of the LCSMM: conceptual, demonstration and validation, development, and production and deployment. Each phase differs in the available data, human resource issues, and the need for CTEA information.

Litton's research for the Army Research Institute determined for each phase of the LCSMM the issues, data available, existing methods, impacts of missing or degraded data on the methods, and requirements for new methods. Litton synthesized a family of CTEA methods applicable to each phase and devised new methodology as required.

COST AND TRAINING EFFECTIVENESS ESTIMATION IN ARMY MATERIEL ACQUISITION

C. Mazie Knerr and Richard K. Matlick
Litton Mellonics
Springfield, Virginia

The Army's Life Cycle System Management Model (LCSMM) defines the process by which Army materiel systems are acquired. It was designed to ensure that all aspects of the system are considered during acquisition including personnel and training requirements. Cost and training effectiveness analysis (CTEA) provides to decision makers information about how training is influenced by and influences the characteristics of the developing system. CTEA evaluates the development of training systems to support the materiel system and the evolution of a materiel system compatible with personnel and training capabilities.

OBJECTIVE

The objective of this research was to provide Army analysts with a performance guide for CTEA at each appropriate stage of the LCSMM. The attainment of this objective required:

- Determination of points in the LCSMM at which CTEA are needed.
- Assessment of the utility of existing methods for CTEA at each point.
- Adaptation of methods to each LCSMM point where CTEA are needed.
- Development of methods for areas that existing ones do not cover or cover inadequately.
- Identification of input information required and available at each point.
- Estimation of the impacts of missing or degraded information on the CTEA.
- Synthesis of a general model for CTEA in the LCSMM.
- Production of a CTEA Performance Guide for Army analysts.

CTEA SITUATIONS AND STRATEGIES IN THE LCSMM

Several LCSMM events require CTEA information. For example, four cost and operational effectiveness analyses (COEA) are conducted during the LCSMM. CTEA support COEA by providing assessments of alternative ways to train to achieve the desired operational effectiveness of the system as well as providing the cost of each proposed alternative. At least four CTEA are needed to support the COEA and additional CTEA are beneficial to support training development decisions.

Our examination of the LCSMM decision requirements revealed the need for a more detailed analysis of the CTEA job to be performed by the Army analyst. Materiel systems do not progress through the LCSMM in lock-step fashion. Systems may be fielded with most, some, little, or none of the data described in Army doctrine, and the data are not necessarily generated in the prescribed sequence. If we wished to give the Army analyst the wherewithal to perform a CTEA, we had to use as guidance (1) the decision to be rendered on the basis of the CTEA information and (2) the data available to conduct the CTEA. These two factors are the boundaries for CTEA methodology regardless of the point in the LCSMM.

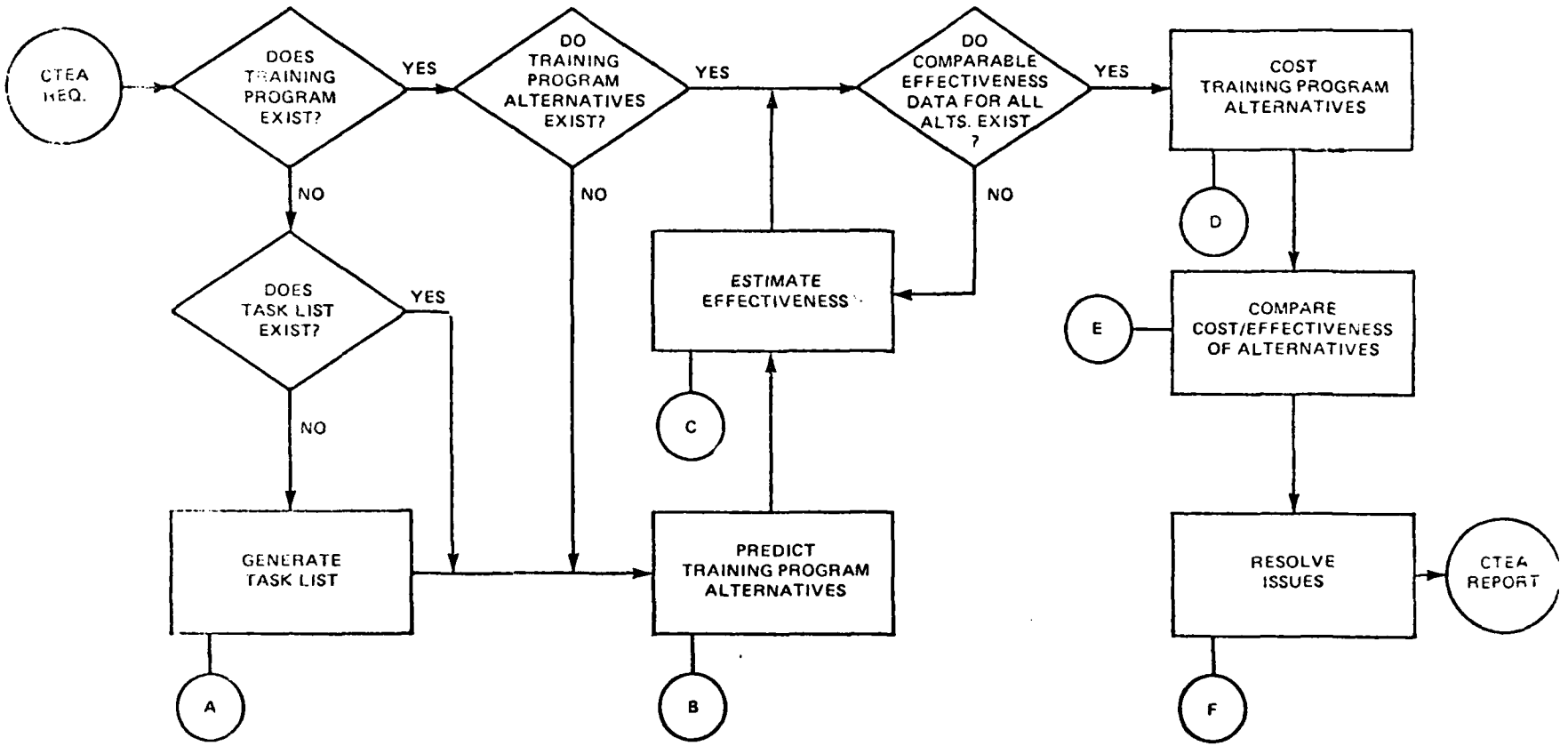
Some data conditions and sequences are more likely to occur than others. For example, the lack of a task list or training program implies the lack of training effectiveness data. Litton identified six input-data situations as:

1. No task list and no training program,
2. Task list but no training program,
3. Training program but no alternatives and no effectiveness data,
4. Training program with effectiveness data but no alternatives,
5. Alternative training programs but no effectiveness data for all alternatives, and
6. Training program alternatives and effectiveness data for all alternatives.

Litton devised a logical model for CTEA to guide the analyst in assessing the data situation and in selecting CTEA processes (Figure 1). The model leads the analyst through questions concerning the availability of task lists, training programs (including alternative training programs), training effectiveness estimates, cost analyses, cost effectiveness comparisons, and issues to be resolved. Depending upon the data situation the analyst needs a different set of CTEA processes.

None of the CTEA methods unearthed in Litton's literature review provided all information required for decisions at any point in the LCSMM. However, detailed examination of the techniques, steps, or processes within the methods (i.e., the elements from which the methods are built) showed that each method contributed to the CTEA job to be performed. Some methods contributed more elements within some data situations than others. For example, informal, expert-judgment methods in current use generate task lists if none are available. Formal analytic models for prediction of training programs, however, are superior to the judgmental models if task lists already exist.

Litton designed CTEA strategies to guide the selection of the processes and conduct of the CTEA. The strategies correspond roughly to the data situation at which the analyst starts the CTEA. Six strategies were required to perform under all conditions. Methods and processes were selected or developed to fully implement each strategy. In addition, alternative methods were selected for each process required by a strategy. Automated methods were selected whenever possible to reduce demands on personnel resources available to perform CTEA.



PROCESS A - GENERATION OF TASK LIST
B PREDICTION OF TRAINING PROGRAMS
C ESTIMATION OF EFFECTIVENESS
D COST OF TRAINING PROGRAM
E COMPARISON OF TRAINING PROGRAM ALTERNATIVES
F RESOLUTION OF ISSUES

Figure 1.
GENERAL CTEA MODEL

Extant CTEA methodology contained three weaknesses. First, use of historical data was hampered by inadequate definition of analogous tasks (i.e., tasks in a fielded system functionally and behaviorally similar to tasks in the proposed system). Second, the existing methodology did not thoroughly address the issue of trainability, one of the estimations needed in the LCSMM. Third, the most thorough cost analysis model omitted costs of training in units.

Litton devised methods to meet these methodological needs. First, our analogous task method (ATM) was developed to take advantage of information obtained in the process of training soldiers on fielded weapons systems. The method is applied on a task-by-task basis with an overall estimate of training effectiveness. The analogous task method has six steps: (1) definition of the critical tasks to be performed on the developing system (the target tasks); (2) classification of the target tasks to provide a basis for finding the analogous tasks; (3) analogous task identification and selection; (4) assessment of training for the analogous tasks; (5) generation of estmates of training for the target tasks; and (6) aggregation of the effectiveness and cost measures across all tasks to obtain a picture of training for the developing system as a whole.

The second methodology devised by Litton, trainability analysis, is needed to examine the interactions among tasks, training program alternatives, and personnel characteristics. Once tasks have been identified and alternative means of training those tasks have been predicted or developed, it is necessary to determine that, given the characteristics of the personnel who will man the system, the tasks can be trained to required levels of proficiency. Our literature review unearthed no explicit method for trainability analysis, therefore we proposed a process to use as an interim technique until further research is conducted.

Litton's third methodological development expanded the TECEP cost model (TAEG No. 16: Braby et al., 1975) to cover costs of institutional field training and training in Army units. The Litton cost model is intended to aid the CTEA analyst in preparing recommendations to the decision maker regarding choices among alternatives. It is not intended for budgetary purposes. The Litton model does, however, capture pertinent institutional and unit training costs.

PROCESSES IN THE CTEA STRATEGIES

Powerful methods applicable to CTEA in the LCSMM have been developed. The methods selected for application by Army analysts are:

- Training Efficiency Estimation Model (TEEM; Jorgensen and Hoffer, 1978),
- Training Consonance Analysis (TCA; Hawley and Thomason, 1978),
- Training Effectiveness, Cost Effectiveness Prediction (TECEP, also known as TAEG No. 16; Braby et al., 1975),
- Army CTEA Methods in Current Use:

  DIVAD Gun CTEA,
  Improved Hawk (Hawk PIP) Training Development,
  Roland Training Development,
  Improved TOW Vehicle (ITV) CTEA,
  Diagnostic Rifle Marksmanship Simulators (DRIMS) CTEA, and
  Methods for the Analysis of Training Devices and Simulators (TRAINVICE).

Analytical processes from these methods and from the three methods that Litton developed to fill gaps in CTEA technology were selected to meet the demands of the data situation in the LCSMM (Table 1). Processes appropriate to each data situation were combined in Litton's six strategies to guide the Army analyst in conducting the CTEA. In the example cited earlier, if no task list exists, informal expert-judgment methods in current Army use are recommended to generate the task list (Data Situation A in Table 1). When there is no task list the analyst must perform all processes from generation of the list to resolution of CTEA issues. Since the analyst uses expert judgment to generate the task lists, the analyst is likely to use expert judgment for prediction of alternative training programs and for estimation of training effectiveness. Formal analytical models can also be used to predict the training programs and effectiveness. However, the estimates need to be in the same terms, i.e., produced by the same method, to be comparable.

If task lists exist, the formal analytical methods are recommended to predict training program alternatives. Training consonance analysis (TCA) can be used to estimate effectiveness no matter how the programs are predicted. The training consonance ratios produced by TCA are the effectiveness estimates. The Litton cost model, modified from TECEP, is recommended for all CTEA situations in the LCSMM. It is important to note that the cost estimates must be comparable in level of detail and precision to make valid cost comparisons among alternative training programs.

The CTEA strategies guide the analyst through the analysis using processes selected to meet the needs of the data situation and to answer the questions asked of the CTEA. The strategies also consider problems that arise if some training programs or cost effectiveness estimates are global estimates while others are precise empirical data points. Litton wrote a CTEA Performance Guide that describes in detail how to conduct the CTEA under each set of conditions.

ADDITIONAL CTEA METHODS

Several other CTEA methods were reviewed but were not incorported into the model. Four of the methods were developed for the Air Force and would require substantial revision to make them suitable for Army use. These methods are automated in part or in total and use consolidated data bases. They are powerful methods that should be considered for future inclusion when they can be adapted to Army CTEA. They are:

- Coordinated Human Resources Technology (CHRT; Goclowski et al., 1978),
- Method of Designing Instructional Alternatives (MODIA: Carpenter-Huffman et al., 1977),
- Digital Avionics Information System: Training Requirements Analysis Model (DAIS/TRAMOD); Czuchry et al., 1978),
- B-1 Bomber Systems Approach to Training (B-1 SAT; Sugarman et al., 1975).

The Training Developer's Decision Aid (Pieper et al., 1979) method was developed for the Army and it is generally applicable to CTEA. It has not yet been fully developed, however, and some problems in implementation remain,

Table 1. CTEA Processes By Data Situation

A GENERATION OF TASK LIST	B PREDICTION OF TRAINING PROGRAMS	C ESTIMATION OF EFFECTIVENESS	D COSTING OF TRAINING PROGRAMS	E COMPARISON OF TRAINING PROGRAM ALTERNATIVES	F RESOLUTION OF ISSUES
DIVAD GUN HAWK PIP ROLAND	TEEM TECEP DIVAD GUN ATM	TEEM DIVAD GUN ATM TRAINVICE TCA	LITTON COST MODEL	TEEM TCA BDM/CARAF DIVAD GUN DRIMS ATM TECEP	ITV DRIMS TRAINABILITY ANALYSIS

especially the generation of task lists. Because of the extent of arbitration in the logic of the method, an extensive validation effort is necessary to demonstrate the usefulness and precision of its outputs. Available descriptions of the method are not complete enough to permit its inclusion in the CTEA Performance Guide.

CTEA methods of the future, if they are realized by the Army, will probably all be computer based. This feature promises not only greatly enhanced efficiency in the processing of information but almost certainly a degree of precision not now achievable. It is possible that such an increase in the efficiency and precision of information processing could both increase the effectiveness of systems and reduce the time required to acquire them.

REFERENCES

Braby, R., Henry, J.M., Parris, W.F., Jr., & Swope, W.M. A technique for choosing cost-effective instructional delivery systems (TAEG Rep. No. 16). Orlando, FL: Dept. of the Navy, Training Analysis and Evaluation Group, April 1975.

Carpenter-Huffman, P. MODIA: Vol. 1, Overview of a tool for planning the use of Air Force training resources, R-1700, Project Air Force Office (AF/RDQA), Washington, DC: Hq USAF, July 1977.

Carpenter-Huffman, P. MODIA: Vol. 2, Options for course design. R-1701, Project Air Force Office (AF/RDQA), Washington, DC: Hq USAF, July 1977.

Carpenter-Huffman, P., Pyles, R., & Fujisaki, M. MODIA: Vol. 3, Operation and design of the user interface, R-1702, Project Air Force Office (AF/RDQA), Washington, DC: Hq USAF, July 1977.

Czuchry, A.J., Doyle, K.M., Frueh, J.T., Baran, H.A., & Dieterly, D.L. Digital Avionics Information System (DAIS): Training Requirements Analysis Model; Users Guide, AFHRL-TR-78-58611), Wright-Patterson AFB, OH: Air Force Human Resources Laboratory, September 1978.

Czuchry, A.J., Glasier, J., Kistler, R., Bristol, M. Baran, H. & Dieterly, D.L. Digital Avionics Information System (DAIS) Reliability and Maintainability Model, AFHRL-TR-78-2, 1978.

Goclowski, J.C., King, G.F., Ronco, P.G., and Askren, W.B. Integration and Application of human resource technologies in weapons system design: coordination of five human resource technologies. AFHRL-TR-76-6. Brooks AFB, TX: US Air Force Systems Command, 1978.

Goclowski, J.C., King, G.F., Ronco, P.G., & Askren, W. Integration and application of human resource technologies in weapon system design: processes for the coordinated application of five human resources technolgies. AFHRL-TR-78-6 (II). Brooks AFB, TX: US Air Force Systems Command, March 1978.

Goclowski, J.C., King, G.F., Ronco, P.G., & Askren, W.B. Integration and application of human resource technologies in weapon system design: consolidated data base functional specifications. AFHRL-TR-78-6 (III). Brooks AFB, TX: US Air Force Systems Command, May 1978.

Hawley, J.K. & Thomason, S.C. Development of an air defense cost and training effectiveness analysis (CTEA) methodology (for the AN/TSQ-73): Vol. I - CTEA within the life cycle system management model. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, 18 December 1978.

Jorgensen, C.C. & Hoffer, P.L. Prediction of training programs for use in cost training effectiveness analysis. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, 1978.

Pieper, W.J., Elliott, T.K., Hawley, J.K. Tryout of a training developer's decision aid for optimizing performance-based training in machine-ascendant MOS. Valencia, PA: Applied Science Associates, Inc., 1979.

Sugarman, R., Johnson, S., & Ring, W. B-1 systems approach to training, final report, SAT-1, Vol. 1, Wright-Patterson AFB, OH: B-1 Systems Project Office, Data Configuration Division, July 1975.

Self-Concept, Organizational Image, and their Relationships to Organizational Choice

Lieutenant C.D. Lamerson

Abstract

One hundred and sixty-eight grade eleven students from a central Ontario secondary school described their future intentions, their selves and their most and least preferred organizations. Most and least preferred organizations were chosen from six large Canadian organizations (Canadian Armed Forces, Canadian Broadcasting Corporation, Eatons, International Business Machines, Ontario Civil Service and Royal Bank).

The results indicate that the students have definite images of both their self-concepts and of the organizations. Differences in organizational images were also noted between the sexes. The students described themselves significantly more similar to their most preferred than least preferred organization. Organizational preference groups had characteristic and uncharacteristic adjectives describing both the members and the most preferred organizations.

It was concluded that (with reference to these specific ones) images were firmly held of the organizations. Young people not only have images of organizations but of themselves (their self-concepts). The image of the organization and how it is seen as relating to the self-concept (as perceived by the student) is also related to what organizations are chosen as most and least preferred.

INTRODUCTION

There comes a time in most peoples lives, usually in mid- to late adolescence or early adulthood, when they must decide what it is that they will be doing for most of the rest of their lives. The decision to finish secondary school or quit early;to become an apprentice or get a job; to continue to community college or to university all eventually lead to the major decision of choosing an occupation. Not only must an occupation be chosen but also an organization where the occupation may be practiced must be chosen.

Ginzberg et al. (1951) proposed that the final occupational or organizational choice was made after a ten year development period. At approximately the same time as Ginzberg was proposing his theory of occupational choice, Super (1951) was proposing a slightly different one. Super's theory differed in that he did not suggest when the occupational choice would be made but rather how it would be made. The choice of occupation would cause a person to state explicitly their self-concept since, he would say, we all choose an occupation which we think will satisfy and fulfill our self-concept. Consequently, at whatever point the individual made his occupational choice he would be choosing a means of implementing his self-concept and testing it against reality.

In a later paper Super (1953) further elaborated on his self-concept theory. He suggested that occupational choice is in the nature of a compromise between interests, capacities, roles and opportunities where self battles reality with a compromise being the yielding of one or the other. Super suggested that it not be considered so much an occupational choice but a vocational development. That is, the development of a self-concept and the preference for, choice of, entry into and adjustment to; an occupation.

Although these theories and others have offered some reasons for occupational choice there is also the question of organizational choice. Is it just the same decision making processes or is there some additional process taking place? Behling, Labovitz and Gainer (1968) suggested three theories of how potential employees choose the organization they would like to work for. The first of these the Objective Factors Theory, suggested that the individual very rationally weighs and evaluates the measureable characteristics (pay, benefits, location, etc.) of each organization, then chooses the organization with the combination of characteristics that appeals to him most.

The Critical Contact Theory pointed out the relative lack of knowledge held by prospective employees about organizations. Since there was little knowledge about the organization the individual found it difficult to differentiate amongst them. The decision of which to work for was thus made on the basis of the impression formed of the organization during interviews and from application proceedings.

The last theory, Subjective Factors Theory, suggested that the individual looks for an organization with a personality that matches his own. This theory has a more personal and emotional basis. It is the organization's perceived ability to satisfy the needs of the individual and be compatible with his self-concept that is important. The individual is thought to have a fairly stable organizational image that has developed from secondary sources long before he starts deciding on an organizational choice. Objective factors brought up about the organization can supplement or detract from this preconceived image but only slightly.

Super's (1951 & 1953) theories of vocational development have much in common with the Subjective Factors Theory. The choice of the organization is an attempt to implement the self-concept. Therefore, organizational choice is seen as a relationship between organizational image and self-concept.

Tom (1971) found that his sample of college students described themselves as more similar to their most preferred to their least preferred organization when they described themselves, their most and least preferred organization using the scales of the Adjective Check List (Heilbrun and Gough, 1965).

Vroom (1966) and later Vroom and Deci (1971) showed that MBA students making organizational choices rated an organization more attractive if they perceived it as a place they would be able to attain their goals. Therefore, the attractiveness of the organization directly related to the extent the subjects believed organizational membership would be instrumental to attainment of their goals.

Englander (1960), in a similar study, found that the group of teachers he was studying had chosen their occupation because it was a means of fulfilling their self-concept.

(556)

Ziegler (1970) found that self-descriptions were more similar to a fictitious person in their most preferred occupational area than to one in their least preferred occupational area. He also found that certain terms were specifically used to describe an occupational interest area.

Wanous (1977) in an extensive review of the literature concerning organizational choice and entry concluded that most of the studies indicated a concern with satisfying personal needs through membership in an organization.

Finally, Rodgers (1976) looked at high school students and found that although they had little or no work experience they had definite images not only about the organizations but about the total workings of the organizations. These images included views on vacation time, job autonomy, pay, technicality of jobs, etc.

The purpose of this research was to continue in the same vein as the above studies in the manner of design of the study but to differ from them in the interest area studied. Steps beyond the above-mentioned research were attempted by looking at the relationship of self-concept and organizational image as they affect organizational choice;as opposed to looking at the relationship between self-concept and occupational image as it relates to occupational choice. A small, controlled number of organizations were looked at rather than letting subjects choose organizations in an open ended fashion.

While most of the previous studies have used university students or job applicants with relatively large amounts of work experience this study used secondary school students who have little or no work experience with these or any other organizations. In this manner support can be sought for the results of Rodgers (1976) who also used secondary school students as subjects while looking at the same six organizations as will be used in this study.

Students completed a questionnaire containing the Adjective Check List (Heilbrun and Gough, 1965) to describe themselves. Their most preferred organization (MPO) and least preferred organization (LPO); a rating scale of the six organizations; a scale stating their future intentions; and some questions on biographical data. The expectations of what would be found during this study were:

(1) that subjects would have very definite images of the organizations regardless of their age; and these images would differ according to the sex of the individual.

(2) that subjects would describe themselves as more similar to their most preferred than least preferred organization. Proof of this hypothesis would offer support for the results of Tom (1971) and Ziegler (1970).

In addition to these two hypothesis it was also expected that when the organizational preference group descriptions were analyzed that certain adjectives would be characteristically used to describe members or MPO of each group. It was also expected that some uncharacteristic adjectives would be found being considered non-descriptive of members and MPOs of the organizational preference groups.

Method

Subjects

One hundred and sixty-eight (eighty-seven male and eighty one female) students from a central Ontario secondary school were used in this study. Although all the students were taking a grade eleven mathematics course thirty-seven subjects were in grade twelve. The average grade for the students was 11.2. Ages of the students ranged from fifteen to twenty-two with an average age of 16.4 years. Biographical data is shown in table one.

Questionnaire

The questionnaire (appendix 1) had six major parts. The first part consisted of questions on the subjects' biographical data (age, sex and grade).

The second part asked subjects to indicate their future plans by checking off their plans from a list provided. The list had such plans as: "finish this year; finish grade twelve; finish grade thirteen; join an apprenticeship; attend community college, and; attend university."

The third section was the rating scale for the six organizations concerned. The organizations were listed in alphabetical order and subjects were asked to number them in order of preference from one to six. The organizations used were: Canadian Armed Forces, Canadian Broadcasting Corporation (CBC), Eatons (E), International Business Machines (IBM), Ontario Civil Service (OCS) and the Royal Bank (RB). These organizations were used because they are well known and because they were used in a previous study by Rodgers (1976).

The fourth part was the Adjective Check List (ACL) as designed by Heilbrun and Gough (1965) using the three hundred adjectives that make up the check list, but without using the scales that make up the check list. The subjects described themselves on the checklist by circling those adjectives they thought were descriptive of themselves.

The fifth part was the same three hundred adjectives of the ACL to be used in this case to describe the organization the subject had indicated as most preferred (rank order of one).

Finally, the subjects described their least preferred organization (with a rating of six) on the ACL.

Procedure

Ninety-six percent of the students in grade eleven math completed the Questionnaire while the other four percent were either absent or chose not to complete the questionnaire.

Instructions were included in the questionnaire but the main instructions, regarding the students' freedom to choose to complete the questionnaire, confidentiality and anonymity were read out loud to them by a tape recording. They were allowed to make their own decision: either to complete the questionnaire or to leave it blank. Teachers were instructed not to attempt to answer any questions posed by students, rather to instruct them to do their best.

Results

Although all one hundred and sixty-eight of the completed questionnaires were accepted as satisfactory, some parts of individual questionnaires had to be discarded because they were left blank or were ruined. Only four questionnaires were ruined.

The students' future plans can be seen in table two. The majority of students (33.9% or 57 students) had plans to finish grade twelve;the second most popular intention was to attend community college (27.4% or 46 students). The other intentions, in their order of choice, were: university (20.8%); grade thirteen (7.1%); grade eleven (3.6%); and, no choice made (1.8%).

The first hypothesis, concerning the organizational image and its view by the sexes, is dealt with in tables three, three A, four and four A. Table three indicates the numbers and percentages of students who rated each organization as their most preferred one. Most frequently chosen of the MPO's was the CBC followed closely by the Canadian Forces. The number of people preferring them were forty and thirty-seven, respectively. Least often chosen of the MPO's was the Ontario Civil Service (14 people chose it).

Table three A indicates the choice of MPO by sex in numbers and percentages. Males chose Canadian Forces as most popular MPO while Eatons was the least popular. Females chose Eatons as most popular while Ontario Civil Service was the least popular. A Spearman rank-order correlation coefficient of -.26 was determined indicating no similiarities between the way males and females rated the organizations.

Numbers and percentages of students considering each organization as the one they would least like to work for can be seen in table four. Canadian Forces was most frequently chosen as least preferred with 52 students rating it this way, followed by Ontario Civil Service with 35 students. The organization least often chosen as LPO was the Royal Bank with 13 people rating it thus.

In table four A the sex by LPO ratings can be seen with numbers and percentages of students rating them. Males most often said Canadian Forces for LPO. The females also most often said the Canadian Forces.

In order to determine whether students described themselves more similarly to their MPO than LPO, their individual ACL scores were analyzed. For each subject the ratio of SELF-MPO was compared to the ratio of SELF-LPO. This ratio is the same as the one used by Ziegler (1970) in his study of occupational choice and was determined by taking the number of adjectives that were the same for self-description and description of MPO over the N of adjectives for self-description to produce the SELF-MPO ratio. For the SELF-LPO ratio the N of adjectives that were the same for self-description and description of LPO over the N of adjectives for self-description was used. The sums of the two ratios were then computed separately and their respective means were determined. The mean SELF-MPO equals .27 while the mean SELF-LPO equals .13 (table five). A t-test was used to determine the difference and see if it was significant, with a resulting $t = 5.406$. This is significant at the $\alpha < .99$ level. Therefore, the SELF-MPO ratio was significantly greater than the SELF-LPO ratio. Thus the significant difference indicates that subjects described themselves more similarly to their most than least preferred organization.

To determine the descriptive adjectives that were characteristic of an organizational preference group each preference group was considered separately. For an adjective to be considered characteristic an arbitrary criterion of fifty percent was set. This meant that fifty percent of the preference group had to use an adjective before it was considered characteristic. To prevent adjectives that were popular, as opposed to being characteristic, from influencing the results a correction factor was devised whereby the percentage of that adjective's use across all organizational preference groups was determined. Therefore, to be characteristic an adjective not only had to be used by fifty per-cent or more of the preference group, but its percentage of use within the group had to be higher than its percentage of usage across groups.

Using these two criteria the adjectives that characteristically described the members of each organizational preference group were determined (table six).

The second use of the above-mentioned two criteria was to determine the adjectives characteristic of each organization when it was described by the people most preferring it (table seven).

Adjectives that were never used by the subjects of an organizational preference group to describe themselves can be seen in table eight. The adjectives that were never used to describe any subject of any preference group can also be seen in table eight. The adjectives with asterisks (*) behind them are those considered as unfavorable by the manual of the ACL. The percentage of uncharacteristic adjectives that are unfavorable in each organizational preference group was also computed and is shown in table eight.

Table nine indicates the adjectives never used to describe the organization by those who rated it as MPO. Seventeen adjectives were never used to describe any of the most preferred organizations. Asterisked (*) adjectives were those considered unfavourable by the manual of the ACL. The percentage of uncharacteristic adjectives that were unfavourable was computed for each organizational preference group and can also be seen in table nine.

Discussion and Conclusions

This study demonstrates support for the hypothesis that secondary school students in the study would have very definite images of the six organizations. It also supports the expectation that the ratings of the organizations by the different sexes would not be similar (as indicated by the Spearman rank order correlation coefficient).

The results of Ziegler (1970) and Tom (1971) were supported by the results of the second hypothesis indicating that subjects described their SELF as more similar to their MPO than their LPO. The results can also be considered an extension of the work of Englander (1960), Blocher and Schutz (1961), Vroom (1966) and Vroom and Deci (1971).

In an extension of Ziegler's (1970) research indicating that occupational preference groups do use characteristic adjectives to describe the organizations this research indicated the presence of characteristic and uncharacteristic adjectives used when describing both the members and the MPO of the organizational preference groups. This suggests that not only are the self-concept and organizational image well formed but that they can be easily verbalized by the use of an ACL.

The theory (Super, 1951 & 1953) that organizational image and self-concept are firmly defined by the individual and are related to each other is supported. Their relation to organizational choice has also been indicated. The support for this theory comes from the findings that students rated the organizations according to their preference (hypothesis one) and that they described their SELF as more similar to their descriptions of MPO than LPO (hypothesis two).

The findings of hypothesis two also offer support for the Subjective Factors Theory (Behling, Labovitz and Gainer; 1968). The students chose as MPO an organization they perceived as a place to implement their self-concept. Thus their descriptions of SELF and MPO were similar.

On the more practical side, the high percentage (33.0%) of students planning to leave school after grade twelve indicates that perhaps the majority of studies in this area, using university students and people already in the work force, were missing a large area of the population with already formed organizational images who would and could be making organizational choices. This was also found by Rodgers (1976) in a similar study using high school students.

It is interesting to note the patterns of how subjects rated the organizations as MPO and LPO. The Canadian Forces was second most popular as MPO but it was also most popular as LPO. This would indicate that subjects had a bimodal opinion; either strongly liking or disliking the Canadian Forces. In comparison the Ontario Civil Service was the least popular MPO and second most popular LPO indicating that the subjects' view of the Ontario Civil Service was a majority disliking it. The view of Eatons was also fairly unimodal; it was third most popular as both MPO and LPO. The Royal Bank and IBM seem to both be viewed with little definiteness. On both scales they are rated as near the least popular end in percentages. This seems to indicate that they are neither highly liked or disliked and are just slightly bimodal. The last of the six organizations, CBC, holds a definite positive image in the eyes of the subjects. It was rated as most popular of MPO's while being second least popular of LPO's.

These results offer further support for the expectation that students would have very definite images of the organizations (hypothesis one). Although they are quite young (average age was 16.4 years) and had little or no experience in the work world (especially with these organizations) they showed that their organizational images are already formed.

Sex of the students (hypothesis one) did make a difference when it came to which of the organizations were rated as MPO or LPO. By within group percentages the Canadian Forces is least popular with females as an MPO.

Changes occur when the sexes are looked at in relation to how they rated LPO's. The Canadian Forces is the most striking having an equal rating of LPO between males and females.

The results indicate that the males and females did have very different images of the organizations. The organizations that might commonly be considered female oriented (the bank and the large department store) were, predictably, highly preferred by the female subjects. The military organization was, perhaps predictably, preferred by males more than females. No explanation as to why the CBC, Ontario Civil Service and IBM were so equally preferred by males and females can be offered.

Although, the purpose of this research was not ~~to~~ look at specific adjectives used to describe the SELF and MPO of the preference groups, it is interesting to note certain aspects of these results. First, the number of characteristic adjectives used to describe SELF as compared to the number used to describe MPO. In all organizational preference groups the SELF had more characteristic adjectives than did the MPO; indicating a more definite view of SELF than MPO.

Secondly, it can be noted that the descriptions (using characteristic adjectives) of any one organization and of the people most preferring it is distinctive from the same descriptions of any other organizational preference group. Although some adjectives are characteristic for more than one group there are still very different combinations of adjectives. Attempting to state that these adjectives definitely describe either an organization or the people attracted to it would cause difficulties because of the different semantic concepts of each adjective that different people would have. Using the descriptions as a method of selection or placement would be far too dangerous due to this semantic problem. However, the adjectives might be more subtly utilized in recruiting and publicity campaigns that would appeal to people who would describe themselves (therefore, have a self-concept) similar to the individuals in the preference groups or would indicate what the organization is like. The uncharacteristic adjectives could also be used in this type of campaign to subtly indicate that the organization attracts people who are not like this (this is not part of their self-concept) and what the organization is not like. For example, a recruiting and publicity campaign of the Canadian Forces, based on this research, might indicate that as an organization it is an active, adventurous, alert, daring and tough place to work (table seven) while not being a contented, initiative, handsome, good-looking or soft-hearted place (table nine). The campaign would probably also attempt to appeal to people who would consider themselves ambitious, adventurous, easy-going and active (table six) while not being commonplace, distractible, feminine, hostile, prejudiced, slow, timid or preoccupied (table eight).

The uncharacteristic adjectives offer a third interesting point about descriptions of organizations and members or organizational preference groups. It could be expected that these uncharacteristic adjectives would also be unfavourable adjectives; however, in no organizational preference group was the percentage of unfavourable adjectives more than fifty percent of the uncharacteristic adjectives. This held true for descriptions of both organizations and members of the preference groups. It was not the case in the groups of adjectives that were never used to describe any organizations or members of the preference groups. These findings seem to speak to the fact that subjects saw themselves and the organizations in a very definite manner. There were certain adjectives that did not suit this definite organizational image and they became the uncharacteristic adjectives irregardless of their being favorable or unfavourable.

Looking at a specific adjective used in describing members of a preference group provides the fourth and last interesting point. This is an uncharacteristic adjective used to describe the members of an organizational preference group. Subjects who most preferred the Canadian Forces as a place to work did not consider themselves feminine; although the results showed that twenty-seven percent of the preference group were female. From this it may be necessary to conclude that the Canadian Forces has not only a very definite image in the eyes of the public but that this image is not feminine in nature.

In conclusion, it can be seen that the image of organizations, even when the person has little or no contact with them, is well formed. Not only do secondary school students have definite organizational images but, their self-concepts are firmly delineated.

At that point in time when organizational choice must be decided upon it will be based on the organization's image and how similar it is perceived to be to the self-concept. Thus, the relation of self-concept and organizational image are also related to organizational choice.

TABLES

TABLE ONE

Table 1A - age (N, % and X)

age	N	% of total
22	1	.6
18	8	4.7
17	46	27.4
16	106	63.1
15	7	4.2
Total	168	100.0

Mean age = 16.4 years

Table 1B - sec (N and %)

sex	N	% of total
males	87	51.8
females	81	48.2
Total	168	100.0

Table 1C - grade (N, % and X)

grade	N	% of total
11	131	78
12	37	22
Total	168	100

Mean grade = 11.2 grades

TABLE TWO

Subjects' future intentions

(by N and %)

Intentions	N	% of total
Finish this year	6	3.6
Finish grade twelve	57	33.9
Finish grade thirteen	12	7.1
Apprenticeship	9	5.4
Community college	46	27.4
University	35	20.8
No choice	3	1.8
Total	168	100.0

TABLE THREE

Ratings of MPO

(by N and %)

Organization	N	% of total
CF	37	22.0
CBC	40	23.8
E	26	15.5
IBM	25	14.9
OCS	14	8.3
RB	25	14.9
No choice	1	.6
Total	168	100.0

TABLE THREE A

Ratings of MPO by sex

(by N and %)

Organization	Males (87)			Females (81)		
Total saying	N	% in org.	% of sex	N	% in org.	% of sex
CF (37)	27	72.9	31.0	10	27.1	12.3
CBC (40)	22	55.0	25.3	18	45.0	22.2
E (26)	7	26.9	8.2	19	73.1	20.4
IBM (25)	13	52.0	14.9	12	48.0	14.8
OCS (14)	8	57.1	9.2	6	42.9	7.5
RB (25)	9	36.0	10.3	16	64.0	19.8
No choice (1)	1	100.0	1.1	0	0	0
Total	87		100.0	81		100.0

TABLE FOUR

Ratings of LPO

(by N and %)

Organization	N	% of total
CF	52	31.0
CBC	16	9.5
E	28	16.7
IBM	23	13.7
OCS	35	20.8
RB	13	7.7
No choice	1	.6
Total	168	100.0

TABLE FOUR A

Ratings of LPO by sex

(N and %)

	Organization	males (87)		females (81)		
Total saying	N	% in org.	% of sex	N	% in org.	% of sex
CF (52)	29	55.7	33.3	23	44.3	28.4
CBC (16)	5	31.3	5.7	11	68.7	13.6
E (28)	17	60.7	19.6	11	39.3	13.6
IBM (23)	11	47.9	12.7	12	52.1	14.8
OCS (35)	15	42.9	17.3	20	57.1	24.7
RB (13)	9	69.2	10.4	4	30.8	4.9
No choice (1)	1	100.0	1.0	0	0	0
Total (168)	87	/	100.0	81	/	100.0

TABLE FIVE

Ratios for SELF-MPO AND SELF-LPO

Equations

SELF-MPO ratio= N of same adjectives SELF + MPO / N SELF

SELF-LPO ratio = N of same adjectives SELF + LPO / N SELF

Computations

Total for all subjects (168)

SELF-MPO ratio = 46.26
SELF-LPO ratio = 21.67

Mean for all subjects (168)
Mean SELF-MPO ratio = .27
Mean SELF-LPO ratio = .13

T-test for significance
$t = (X1 - X2) / Sx1 - x2$
$= 5.406$
$df = 334$
$o \quad .99 = 1.960$

Therefore there is a significant difference between X1 and X2.

TABLE SIX

Characteristic adjectives corrected to describe members of organizational preference groups.

Organization	adjectives
RB	cheerful, cooperative, emotional, good-natured, mature healthy, honest, kind, forgiving, warm, friendly, helpful, reasonable, responsible, reliable, trusting.
CF	adventurous, ambitious, easy-going, active.
CBC	adventurous, ambitious, cooperative, easy-going, excitable, active, emotional, dependable, good-natured, humorous, imaginative, healthy, friendly, responsible, sensitive, sociable, trusting, warm.
IBM	cooperative, mature, responsible, understanding.
E	ambitious, excitable, curious, emotional, dependable, gentle, mature, healthy, honest, forgiving, friendly, helpful, reasonable, sensitive.
OCS	adventurous, ambitious, appreciative, cautious, cheerful, easy-going, active, argumentative, dependable, energetic, good-natured, humorous, imaginative, mature, healthy, logical, helpful, intelligent, pleasant, reasonable, responsible, self-controlled, sensitive, soft-hearted, trusting.

TABLE SEVEN

Characteristic adjectives corrected to describe MPO of organizational preference groups.

Organization	adjectives
RB	cheerful, cooperative, mature, honest, sociable, pleasant, trusting.
CF	adventurous, daring, active, alert, tough.
CBC	adventurous, artistic, active, imaginative, interest-wide, friendly.
IBM	efficient.
E	cheerful, mature, honest, friendly, helpful, sociable, reliable.
OCS	adventurous, cheerful, cooperative, logical, outgoing, friendly, organized, thorough, trusting, understanding.

TABLE EIGHT

Uncharacteristic and unfavourable adjectives describing members of organizational preference groups.
(* denote unfavourable adjectives)

Organization	adjectives (not said)	*% unfavourable adjectives
RB	cruel*, enterprising, boastful*, fault-finding*, peculiar, reflective, severe, self-pitying*, show-off*, stingy*, thorough, thrifty.	50.0
CF	commonplace, distractable, feminine, hostile*, prejudiced*, slow, preoccupied, timid.	25.0
CBC	formal, hasty.	0.0
IBM	absent-minded, jolly, spontaneous, reckless.	0.0
E	dignified, shrewd, spend-thrift.	0.0
OCS	interests-narrow*, rebellious, reserved, self-centered*, suggestible.	40.0
Never said	despondent, snobbish*, queer*, slipshod*, smug, unscrupulous*.	66.6

TABLE NINE

Uncharacteristic and unfavourable adjectives describing MPO of organizational preference groups.
(* denote unfavourable adjectives)

Organization	adjectives (not said)	*% unfavourable adjectives
RB	conscientious, energetic, nervous optimistic, opinionated*, sensitive, simple, self-seeking, pleasant, polished.	10.0
CF	contented, initiative, handsome, good-looking, soft-hearted.	0.0
CBC	moderate, sexy.	0.0
IBM	argumentative*, impulsive, peculiar, sentimental, spunky, unemotional.	16.6
E	foresighted, spontaneous, sympathetic.	0.0
OCS	artistic, excitable, cold*, cool, mischivious, moody, noisy, outspoken, shrewd, touchy.	11.0
Never said	absent-minded, despondent, effeminate, cowardly*, distrustful*, foolish*, infantile*, inhibited, quitting*, slipshod*, smug, spineless, unscrupulous*, vindictive*, weak*, sulky*, whiner*.	70.5

(570)

References

Behling, G., Labovitz, G., and Gainer, M. College recruiting: A theoretical base. Personal Journal 47,13-19,1968.

Blocher, D.H. and Schutz, R.A. Relationships among self descriptions, occupational stereotypes and vocational preferences. Journal of Counselling Psychology. 8, 314-317, 1961.

Englander, M.F. A psychological analysis of occupational choice: Teaching. Journal of Counselling Psychology. 7, 257-264, 1960.

Ginzberg, E., Ginsburg, J.W., Axelrod, S., and Herma, J.L. Occupational Choice: An Approach to a general theory, Columbia University Press: New York, 1951.

Heilbrun, A.B. and Gough, H.G. The adjective check list and manual, Consulting Psychologists Press, Inc., 1965.

Morrison, R.L. Self-concept implementation in occupational choices, Journal of Counselling Psychology, 9, 255-260, 1962.

Oppenheimer, E.A. The relationship between certain self-constructs and occupational preferences. Journal of Counselling Psychology, 13, 191-197, 1966.

Rodgers, M.N. Organizational image: A determinant of organizational choice. Unpublished masters thesis, University of Waterloo, 1976.

Super, D.E. Vocational adjustment: Implementing a self-concept. Occupations, 30, 88-92, 1951.

Super, D.E. A theory of vocational development. American Psychologist, 8, 185-190, 1953.

Super, D.E., Starishevsky, R., Matlin, N. and Jordan, J.P. Career development: Self-concept theory, New York: College Entrance Examination Board, 1963.

Tom, V.R. The role of personality and organizational images in the recruiting process. Organizational Behavior and Human Performance, 6, 573-592, 1971.

Vroom, V.H. Organizational choice: A study of pre- and post- decision processes. Organizational Behavior and Human Performance. 1, 212-225. 1966.

Vroom, V.H. and Deci, E.L. The stability of post-decision dissonance: A follow-up study of the job attitudes of business school graduates. Organizational Behavior and Human Performance, 6, 36-49, 1971.

Wanous, J.P. Organizational entry: Newcomers moving from outside to inside. Psychological Bulletin, 84, 601-618, 1977.

Ziegler, D.J. Self-concept, occupational member concept, and occupational interest area relationships in male college students. Journal of Counselling Psychology, 17, 133-136, 1970.

LAWTON, George W., U.S. Army Research Institute, Alexandria, Virginia.

SCALING THE VALUE OF INCENTIVES FOR ARMY ENLISTED MI PERSONNEL
(Tue P.M.)

Incentives are an important tool in personnel management. But some incentives are more effective than others, and some incentives cost more than others. The author has developed a technique using functional measurement theory, to measure the value of nonmonetary and monetary incentives using the same metric, namely dollar value. The paper will describe the development of a questionnaire for use in scaling the value of incentives, and the results obtained from administering the questionnaire to a group of enlisted military intelligence personnel. Implications of these results for job satisfaction, personnel management, and monetary incentive programs will be discussed.

SCALING THE VALUE OF INCENTIVES FOR ARMY ENLISTED MI PERSONNEL

George W. Lawton
U. S. Army Research Institute
Alexandria, Virginia

I have conducted some research concerned with the value, to soldiers , of existing and potential incentives. The practical reason for doing this research was to find out what soldiers in Electronics Warfare-Cryptologic MOSs like and don't like about their jobs, and what the Army might offer them to make their jobs, and the Army, more desirable.

A second reason for doing the project was methodological. Current theories of work motivation (Campbell & Pritchard, 1976) and career decision making identify the value of various outcomes of work and career decisions as important determinants of choice, level of effort, and job satisfaction. In order to predict these dependent variables, or to predict the effects of changes in existing incentives, we need measures of the value of different job outcomes.

It is not uncommon in research on work motivation, to simply present verbal descriptions of different incentives, and ask job incumbents to rate their desirability or importance. Measures of value derived from these methods are indirect for two reasons. First, the stimuli involved, usually in the form of a questionnaire, are verbal descriptions of incentives rather than actual incentives. Second, the responses measured are verbal responses and not usually the responses by which the incentive is obtained.

It quite appropriate to raise questions about the kinds of inferences that can be made from such indirect measures. There are at least two potential sources of error in using them (Opsahl & Dunnette, 1966). First, there are large individual differences in people's ability to identify the effective incentives in their behavior. Second, the reinforcement contingencies are different for verbal behavior about the desirability or importance of incentives than for the behavior actually involved in obtaining the incentives. In particular, the social reinforcement for verbal reports may make them inaccurate by biasing them in the direction of social desirability. People may be reluctant to report that incentives like pay are important in comparison to socially acceptable incentives like job autonomy or intrinsic job satisfaction.

The use of such indirect measures of value assumes that some characteristic of the subject's response, e.g. the location of the mark on a rating scale, is functionally related to the verbal description of an incentive, and that it is not functionally related to irrelevant factors. Such irrelevant factors include social desirability, irrelevant content in verbal description of the incentive, formal characteristics of the verbal descriptions like syntax,

This paper was prepared for presentation to the Military Testing Association, Toronto, Canada, 1980. The views expressed in this paper are those of the author and do not necessarily reflect the views of ARI, the United States Army, or of the U. S. Department of Defense.

675

and physical characteristics of the stimuli, like type face, orientation of the rating scale, and order of presentation.

In addition, indirect measures of value assume that the functional relationships between the verbal description of incentives and judged or rated value of incentives are analogous to the functional relationships between actual incentives and the behavior they maintain.

A great deal is known about the relationship between behavior and the conditions of reinforcement. Killeen (1972) has suggested that the concept "value" be regarded as an intervening variable which summarizes the way organisms integrate the conditions and parameters of reinforcement like rate, amount, delay, quality, and instrumentality. We will use the work "value" here in that sense.

It is my contention that the relationships between described incentives and judged value should parallel the relationships between actual incentives and behavior found in laboratory and field experiments. Judged value should be a function of the described amount, duration, and delay of incentive in indirect measurement, just as actual behavior is a function of actual amount, duration, and delay of reinforcement.

The following specific relationships should hold.

1. Value should be an increasing function of amount or duration of positive incentive and a decreasing function of negative incentive or incentive loss.
2. Value should be a decreasing function of delay of positive incentive and an increasing function of delay of negative incentive, or incentive loss.

Where amount and delay of incentive are both allowed to vary, their joint effects should be described by equation 1, which says that value is directly proportional to amount of incentive and inversely proportional to delay of incentive. Equation 2 shows an equivalent logarithmic form.

Equation 1 $$V = a(1/d)$$

Equation 2 $$\log V = V_s = \log a - \log d$$

In summary, this project had two objectives: a theoretical-methodological one concerned with the relationships between rated value, and described parameters of incentive; and a practical one, concerned with producing a scale of value which would allow Army policy makers to choose incentives that would be effective in helping retain enlisted Army intelligence personnel.

METHOD

To evaluate the extent to which judged value is a function of described parameters of incentive like amount and delay, and unrelated to irrelevant variables like source, I included a number of experimental items in a questionnaire. All items were in rating scale format, with the scale ranging from -10 to +10. Subjects were asked to rate each incentive according to its desirability. In some cases, the items described incentives which varied in only a single parameter of incentive like amount or delay. In other cases, two or more parameters such as amount and delay, were described. To construct items that could vary in the desired ways, I followed Galanter's (1975) suggestion that items describing monetary losses and gains be included in the questionnaire. There were three reasons for including these items. First, monetary items provided extremely useful anchors for scaling nonmonetary incentives. For example, in making decisions about the management of incentives, it is potentially useful to know that a promotion is rated as equally desirable to a $1000 cash bonus. Second, the Army has opted to use a system of variable monetary incentives to attract and retain personnel to hard-to-find job categories. To evaluate this program, we need to be able to measure the differential value of different amounts of money. Third, I needed items that could describe incentives that clearly varied in amount. Monetary incentives of different types, from different sources, satisfied this requirement. Sample items of different types are shown in Table 1. The lottery format was adopted from Galanter (1975) except that the Maryland State Lottery was specifically identified since all participants were stationed at Ft. Meade, Maryland.

Items describing other possible incentives were written after I became thoroughly familiar with the job and lifestyle of Army intelligence personnel, and interviewed Army and civilian policy makers, job incumbents, tehir supervisors, and commanding officers. Every effort was made to write items which were appropriate and stated in the language of Army intelligence personnel. A total of 210 items were combined into a questionnaire which was administered to 77 Army Intelligence personnel, in two groups.

RESULTS

1. Rated value of gains and losses of different amounts of money from different sources.

Figure 1 shows the rated disirability of recieving different amounts of money from different sources. Amounts of money varied from one dollar to $500,000. Sources of monetary gain were finding money, winning the Maryland State Lottery, and receiving a one-time bonus from the Army. The curves relating rated value to amount for each source run close together. A plot of rated value for different amounts of money from lottery and bonus is shown in Figure 2. Table 2 shows the results of an analysis of variance on these items. The analysis confirms what visual inspection of the figure shows. A within-subjects analysis, using source and amount as independent variables, showed that variations in ratings due to source, amount, and their interaction,

were all statistically significant. This rather curious result says that soldiers prefer to win small amounts of money in the state lottery, but to receive large amounts from the Army.

Ratings of items describing different ways of losing different amounts of money also show that the rated negative desirability of such loss depends on more than just the amount of money involved. Figure 3 shows the rated value of losing different amounts of money in different ways. The ways of losing money included paying a fine, paying off on a lost bet, and losing money from your pocket. Losing money from your pocket and paying a fine appear to be far less desirable than paying off on a lost bet.

2. Monetary and nonmonetary incentives of equivalent value.

The highest rated items all described large monetary gains. The only nonmonetary item to receive a very high positive rating described a six-month educational leave of absence after every 18 months of Army service. This incentive was rated as roughly equivalent to a $500 pay increase and a $50,000 lottery prize. The nonmonetary incentive receiving the next highest rating was a three month leave-of-absence to attend college at the Army's expense. This was roughly equivalent to a cash bonus of $1000 and less desirable than a $50 pay increase. A cluster of nonmonetary incentives rated between $100 and $500 cash and slightly more desirable than a $20 pay increase inclues promotion, getting the duty assignment you want, and being recommended for promotion.

Nonmonetary incentives equivalent to different monetary losses include shiftwork at -$10, undesirable assignments at -$25, longwork days and group living arrangements at -$100. All of these are currently part of the Army intelligence soldiers lifestyle.

3. Ratings of the value of receiving different amounts of money after varying delay periods.

In the laboratory study of reinforcement, it has been known for sometime that delay of reinforcement can result in severely reduced performance. This is a practical problem in many work organizations, where incentives like promotion may be announced weeks or even months before becoming effective. Items describing incentives with different associated delays were included to try to assess the impact of delay of incentive.

Lottery questions varying in amounts of money ranging from $10 to $10,000 and with delay periods ranging from zero to two years, were randomly embedded in the questionnaire. Four amounts and five delay periods were included in a 2 x 2 factorial design. The subtractive model for the combination of amount and delay of monetary incentive was tested, using a within subject analysis of

variance. Figure 4 shows the mean ratings for combinations of monetary amount and delay period. In Figure 4 these ratings are shown as a function of delay period, with monetary amount as parameter. Table 5 shows the results of a two way repeated measures ANOVA. Effects of both amount and delay are highly significant, as was the interaction between amount and delay. To examine the subtractive hypothesis for monetary loss, items showing three different amounts of fine and five different delay periods were analyzed in the same way. Visual inspection of Figure 5 shows the rated negative value does not change much as a function of delay, but that the curves are hardly parallel.

4. Ratings of nonmonetary incentives as a function of delay period.

Figure 6 shows the rated desirability of Army promotion, as a function of delay period. It was my expectation that the slopes of this function would be roughly the same as the delay-of-money functions. This is not the case. These curves drop off more sharply than the monetary functions. Figure 6 shows the promotion data plotted for only those soldiers who would be in the Army for at least two years. The pattern of the results is the same for this group as for the total group. The same result holds for other Army incentives like getting a desirable assignment. Delay severely reduces the value of such incentives.

5. Rated value of shiftwork and workday length for different periods of time.

Figure 7 shows the value of assignment to swing or midshift for a specified or indefinite period of time. There is a small decline in rated value with increasing length of assaignment for midshift. For swing shift, there is no apparent difference between a three month assignment and an indefinite assignment.

Figure 8 shows the rated value of workdays ranging from eight to twelve hours long. Our raters show mild positive ratings for an eight hour day, and negative ratings for ten and twelve hour days. Figure 9 shows the rated value of getting two hours off work as a function of the length of time for which the shortened work day is in effect. In view of the strong aversion expressed for long workdays, it is not surprising that this receives a positive rating. It is surprising that getting two hours off work once is just as desirable as having a permanently shortened workday.

DISCUSSION

The results demonstrate that judged value is a function of described amount and delay of incentive. But they also demonstrate that the rating response is functionally related to other, irrelevant variables. There is nothing in our theory of value to lead us to believe that $10 from the lottery is better than $10 from the Army. Likewise, there appear to be systematic changes in the ratings as a function of position of the item in the questionnaire. In cases where the rated value is not monotonically increasing function of dollar amount, the larger dollar amount appeared later in the questionnaire, suggesting that ratings were either decreasing or becoming more centered, i.e. tending to zero, as the subject responded to more items. The data do not allow examination of thie hypothesis, since order of presentation was not varied. The monetary equivalents of nonmonetary items appart to be high. Possible reasons for this include ceiling effects due to the rating scale, and contamination of th

ratings by expectancy effects. Soldiers may rate educational opportunities more highly than lottery winnings simply because they regard the educational incentives as more likely than the lottery prize. At the very highest levels of monetary incentives, some subjects give negative ratings. Why this is so is not clear, but it accounts for the reversal of the $100,000 and $500,000 lottery items.

It was expected that the slopes of delay functions for monetary and nonmonetary incentives would be similar. That this was not the case also suggests the operation of other variables in determining the rating response.

From the practical standpoint, the primary result is that we have demonstrated that the value of an incentive can be eroded severely when there is a large dela in the description of the incentive. Behavior theory suggests that this is also true of the effect of delay of incentive on actual performance and satisfaction. This suggests that organizations should act to reduce or eliminate delays of incentive of more than a week or two.

The extremely high value these soldiers place on educational opportunities suggests a viable kind of incentive for the Army to use in retaining them. It is likely to be cost effective when compared to the levels of cash bonus which are rated at equivalent desirability.

REFERENCES

Anderson, N. H. (1977), Note on functional measurement and data analysis. Perception and Psychophysics, 1977, 21, 201-215.

_____ (1978), Methods and designs: measurement of motivation and incentive. Behavior Research Methods and Instrumentation, 1978, 10, 360-375.

Campbell, J. P. and Pritchard, R. D. Motivation theory in industrial and organizational psychology. In M. D. Dunnette (ed.) Handbook of Industrial and Organizational Psychology. Chicago, Ill.: Rand McNally College Publishing Co., 1976.

Galanter, E. Scaling utility and disutility of monetary and nonmonetary events. Unpublished Technical Report, Columbia University, 1975.

Killeen, P. The matching law. Journal of the Experimental Analysis of Behavior, 1972, 17, 489-495.

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

FIGURE 1. Rated desirability of different amounts of money received from different sources. Amounts ranged from $1 to $500,000. Sources included finding money, winning the state lottery, and receiving a bonus from the Army.

FIGURE 2. Desirability ratings in response to questionnaire items of the form shown in Table 1 for lottery winnings and Army bonuses.

FIGURE 3. Rated undesirability of losing different amounts of money in different ways. Amounts ranged from $1 to $100. Ways of losing money included losing it from your pocket, paying off on a lost bet, and paying a fine.

FIGURE 4. Desirability ratings in response to questionnaire items of the form: "How deisrable or undesirable would it be if...you win $_____ in the Maryland State Lottery which will be paid to you (after a specified delay). Mean ratings are shown as a function of delay period, with monetary amount as a parameter.

FIGURE 5. Ratings of undesirability in response to questionnaire items of the form ...you have to pay a fine of $_____ (after a specified delay period). Mean ratings are shown as a function of delay period, with monetary amount as a parameter.

FIGURE 6. Desirability ratings in response to questionnaire items of the form: "(After specified delay period) you receive a promotion" and "... your name will appear on a promotion list (after a specified delay period). Mean ratings are shown for both types of item as a function of delay period. These data are for soldiers who have at least two years to separation.

FIGURE 7. Rated value of being assigned to work swing shift or mid shift, shown as a function of the duration of the assignment.

FIGURE 8. Rated value of workdays of different lengths.

FIGURE 9. Rated value of getting 2 hours off work, shown as a function of the duration for which the shortened workday is in effect.

581

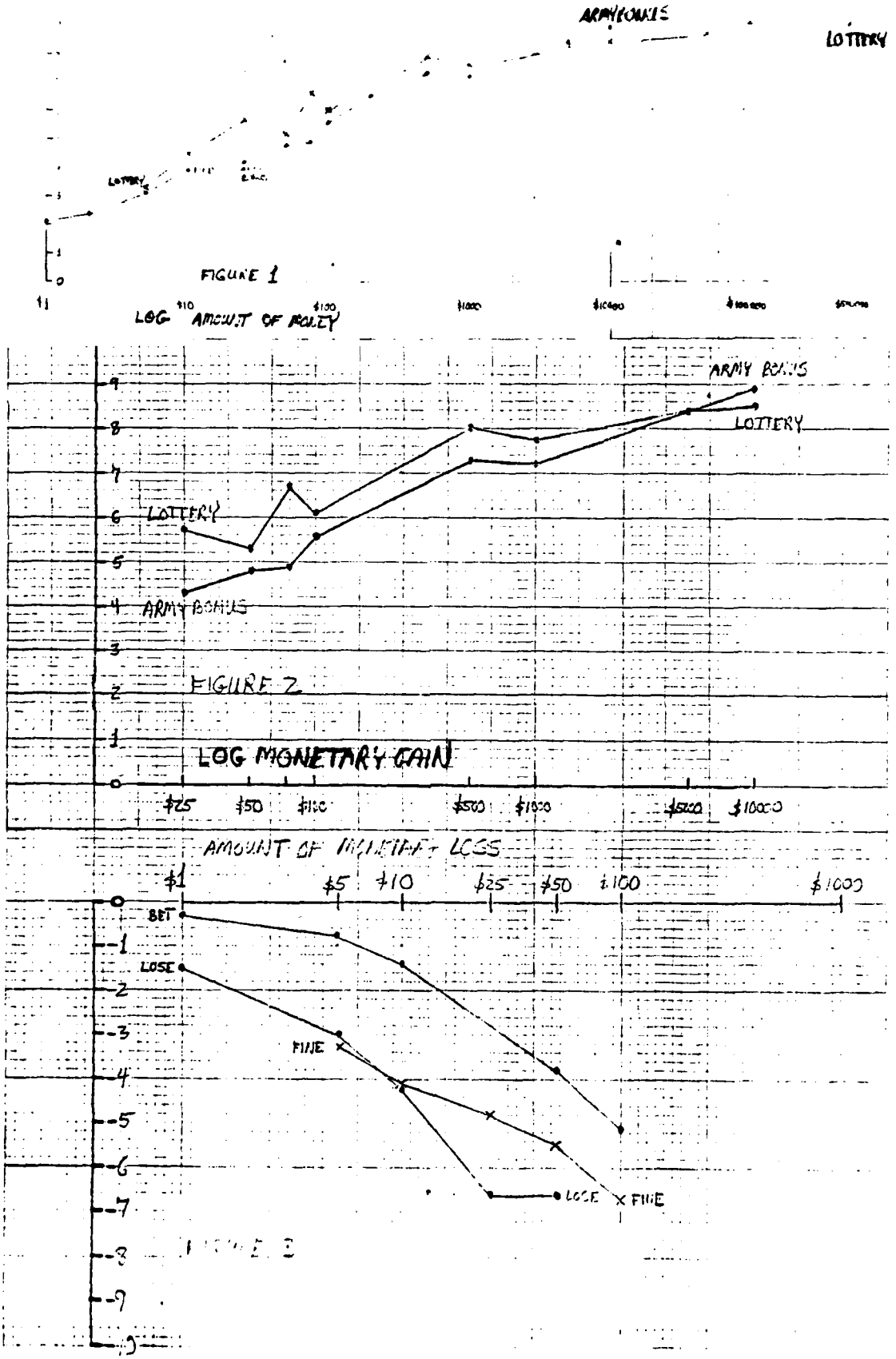
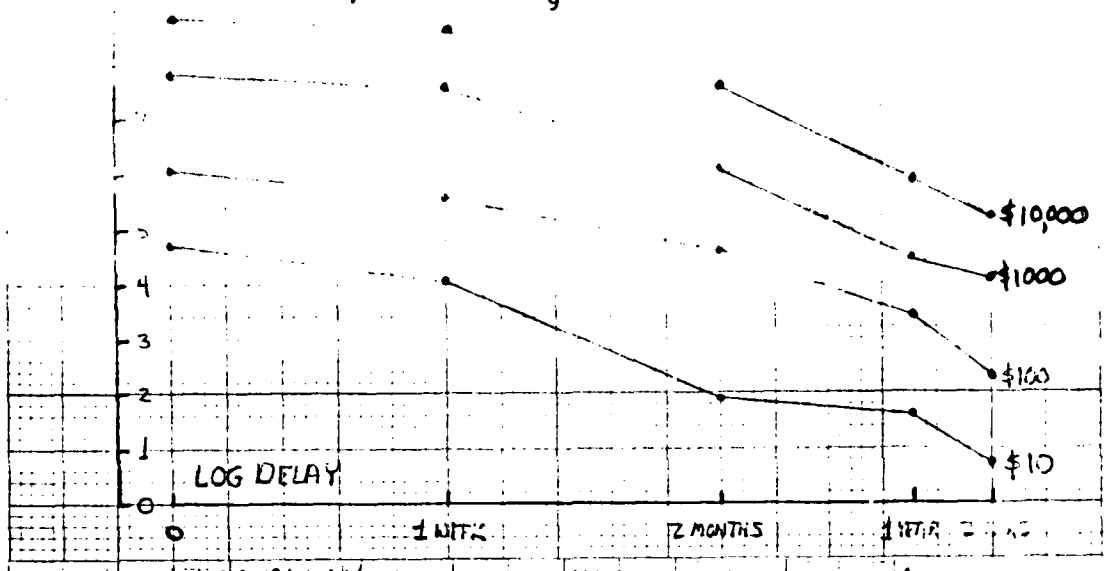


FIGURE 1

FIGURE 2

FIGURE 3

FIGURE 4



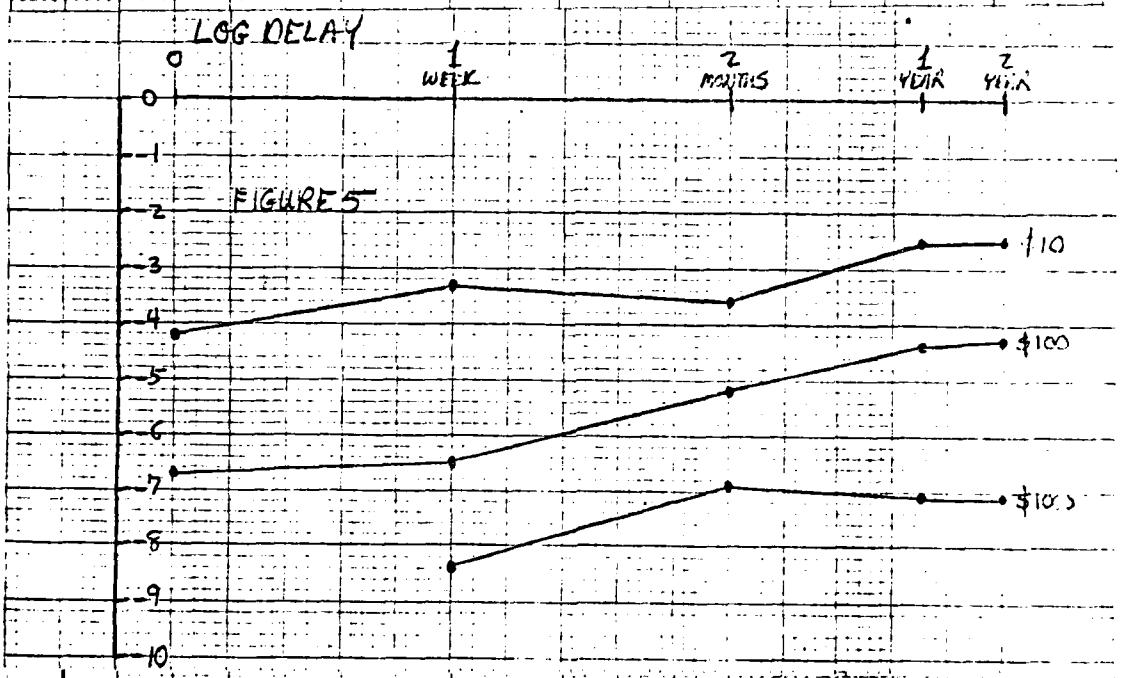
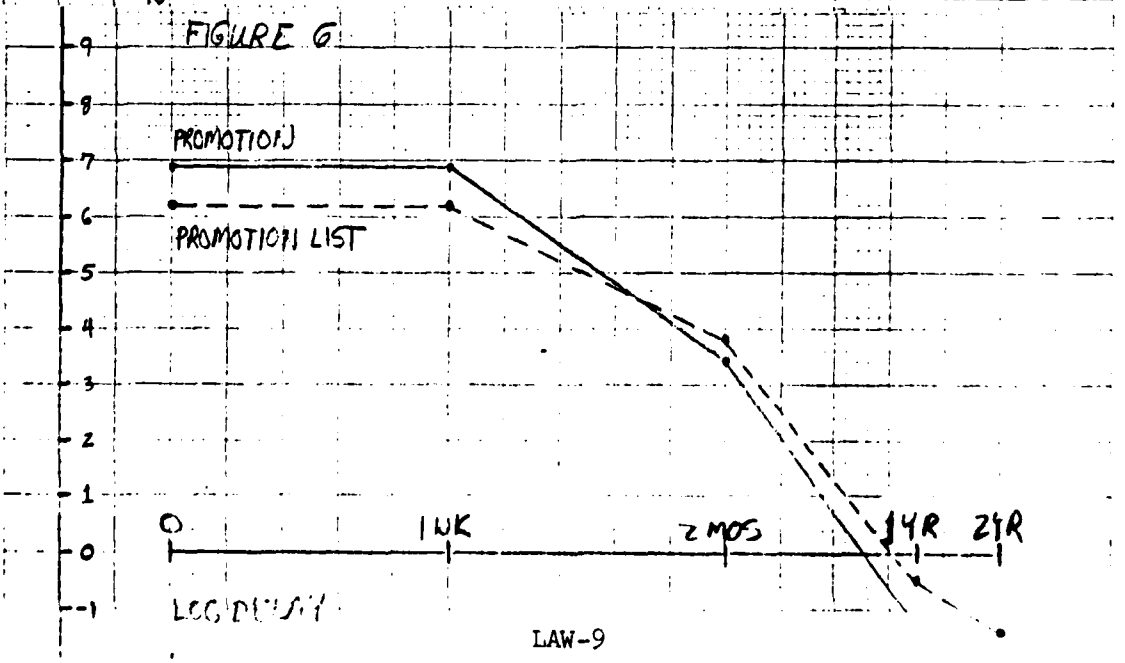


FIGURE 5

FIGURE 6



583

FIGURE 7 FIGURE 8 FIGURE 9

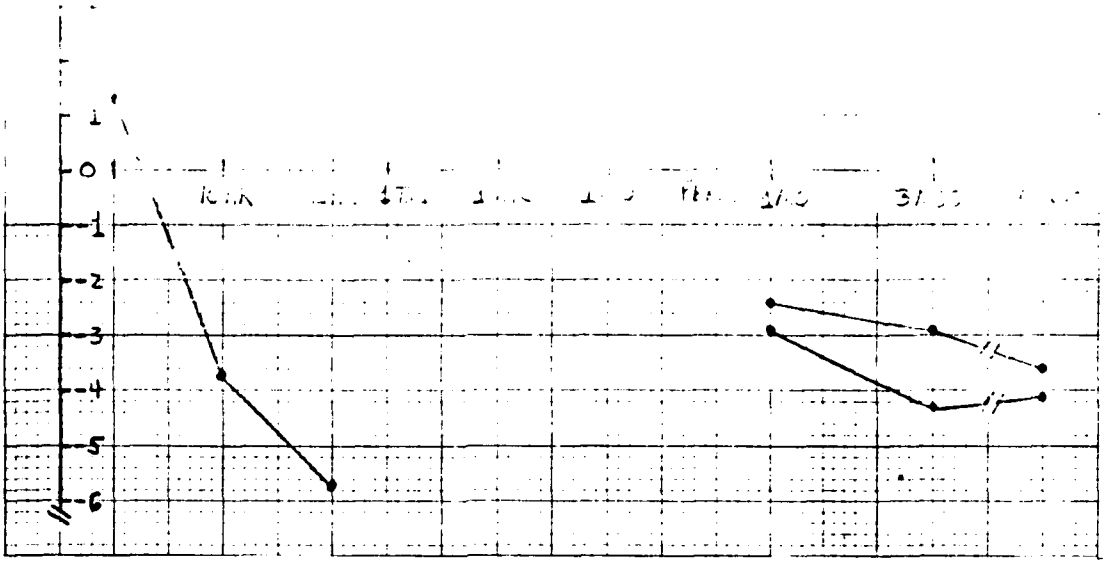


TABLE 1

Sample Monetary Items

How desirable or undesirable would it be if ...

1. You win $____ in the Maryland State Lottery which will be paid to you (after a specified delay)?

2. (After a specified delay) the Army pays you a one-time bonus award of $____?

3. You have to pay a fine of $____ (after a specified delay)?

TABLE 2

ANOVA for monetary items varying in source and amount

Source	df	ms	F	p
subjects	1	55377.00	598.26	.0000
error	75	92.56		
source	1	121.26	8.58	.0045
error	75	14.13		
amount	7	340.19	76.57	.0000
error	525	4.44		
source x amount	7	20.00	6.90	.0000
error	525	2.90		

TABLE 3

ANOVA for monetary items varying in amount and delay

Source	df	ms	F	p
subjects	1	38087.88	183.59	.0000
error	73	207.46		
delay	4	794.84	38.69	.0000
error	292	20.55		
amount	3	1508.65	103.01	.0000
error	219	14.65		
delay x amount	12	7.13	1.75	.0521
error	876	4.07		

(684)

A DESIGN FOR VALIDATING SELECTION PROCEDURES FOR GROUPS OF JOBS

Richard A. Lilienthal[1] and Theodore H. Rosen
U.S. Office of Personnel Management

This paper presents the methodology developed for a U.S. Office of Personnel Management test validation project. The goal was to design a procedure that would enable a large number of jobs to be validated in one study. A molar definition of work behavior was developed so that jobs could be grouped together according to similar work behaviors. A work behavior inventory was chosen as the most appropriate job analysis technique for analyzing many jobs having large numbers of incumbents. A method of generalizing validity from a few criterion-related validation studies to a number of related jobs was developed. It allows generalizing from jobs for which criterion-related validation studies were conducted to other jobs to the extent that they share constructs and work behaviors.

In recent years, organizations have increasingly realized the need to document the validity of their selection techniques. For organizations containing a large number of jobs, separate validation studies for each job are prohibitively expensive in time and money. They may also be impractical from a research viewpoint because of the difficulty in obtaining large enough sample sizes and in developing acceptable criteria for every job. This paper presents a methodology for documenting the validity of selection devices used for a large number of related jobs. The methodology described in this paper was designed for use by the U.S. Civil Service Commission (now the U.S. Office of Personnel Management), the central personnel agency for the executive branch of the Federal government. Parts of the methodology are considered original (e.g., the method of generalizing validity from criterion-related studies) while others are adaptations of existing techniques.

Type of Validity

Of the three major strategies available for validating selection procedures--content, criterion-related, and construct--the construct strategy was chosen as the most appropriate. The other strategies could be employed, but the construct strategy was felt to be best for supporting a selection procedure to be used for a large number of jobs. In our construct validity approach, a job analysis employing a work behavior inventory is used to insure that the selection devices are job related. In addition, several criterion-related studies are conducted to empirically

---

Reprint requests should be sent to Richard Lilienthal, HQDA, PECC-FSS, 200 Stovall Street, Alexandria, VA 22332.

[1]Now with the U.S. Department of the Army.

demonstrate the relationship between the selection devices, constructs, and work behaviors. Construct validity is a more complex strategy than either criterion-related or content validity. However, the personnel selection literature indicates that a number of similar cognitive abilities (constructs) are present in most jobs in an occupational area. For example, 27 professional and administrative jobs were found to share the same six cognitive abilities (McKillip, Trattner, Corts, & Wing, 1977). Thus, the strategy is practical for developing and validating broad-band examinations. Furthermore, it contributes to knowledge in psychology by helping us identify the abilities that underlie successful work performance.

Job Analysis--Writing Work Behaviors

The development of any selection procedure properly begins with a job analysis, and the first phase in most job analyses is the identification of the activities performed to achieve the objectives of the job or group of jobs being studied. Job activities can be written at various levels of specificity. Consider a continuum of specificity ranging from activity statements so specific or molecular that description of a given job would require 1000 statements, to statements so general or molar that a half dozen describe a job. Activities written at a level towards the specific end of the continuum are usually called tasks and those written at a level towards the general end of the continuum are usually called duties. For the purpose of grouping jobs for a validation study, a definition of job activity at a level between task and duty seems most appropriate. The term "work behavior" is used here to describe such an activity level. Table #1 presents tasks, duties, and work behaviors for a given job to show the relationship between them. (The duties and work behaviors listed there are not as general as could be written. More specific statements were written because of the relative simplicity of the job used in the example.)

Using a moderate level of specificity, one can write activity statements that are behaviorally oriented, avoiding the criticism of duties that they are so general as to be meaningless. At the same time, several task statements can be combined into one work behavior statement. Many jobs that appear to be different at the task level are found to be related at the work behavior level. The use of a more molar level is supported by mounting evidence that differences between jobs in task makeup produce, at most, only trivial variations in test validity (Pearlman & Schmidt, Note 1).

Job Analysis--Inventory

An inventory approach was selected for the job analysis. The ease in obtaining job analysis data with an inventory of work behaviors allows a much more extensive sampling of incumbents than would be possible with other job analysis techniques. Scoring is similarly uncomplicated--there is a choice of having incumbents respond on sheets that are machine scored or of using ordinary forms and having the responses keypunched. Statistical packages for analyzing work behavior inventory data already exist (e.g., Christal, Note 2).

The size of the inventory is determined by the number of jobs being described, the similarity between the jobs, the complexity of the jobs, and the level of specificity of the statements written. Preliminary work on writing work behavior statements for aid/technician jobs indicates

a large amount of overlap between jobs; even aids in such seemingly different fields as nursing and engineering have many work behaviors in common. It appears that one inventory of six or seven hundred work behaviors can be constructed for some two dozen aid/technician jobs (see the Appendix for the job titles).

Construction of a work behavior inventory is time consuming. Initial information about work behaviors is obtained from printed material: classification and qualification standards, recent position descriptions, training materials, and the Dictionary of Occupational Titles (U.S. Department of Labor, 1977). Supplementary information is obtained during job-site visits where incumbents are observed performing their jobs and are interviewed about their jobs. The resulting draft inventory is reviewed in a group situation by five to ten supervisors and incumbents who serve as subject matter experts (SME's). They correct technical wording and include additional work behaviors which they know are being performed. The expanded inventory is then distributed for a field review by SME's throughout the country. To make responding to the final inventory easier, work behaviors are organized into groups. Since some work behaviors are more important than others for successful job performance, and since some occur more often than others, measurements of these variables are included in the inventory. Instructions for filling out the inventory form are relatively simple. The incumbent reads the list of work behaviors and checks those work behaviors performed as part of the normal job ("normal" meaning that it is a responsibility of one's being in that occupation); writes in any significant work behaviors that he or she performs which are not in the work behavior list; and then rates the work behaviors checked, using relative-time-spent and importance scales. Instructions for constructing task inventories are available (Melching & Borcher, 1973; Christal, Note 2; Morsh & Archer, Note 3). A practical example of the use of the technique is presented by van Rijn (1977).

The inventory proposed here is similar to these task inventories except for its use of work behaviors instead of tasks.

Identification of Constructs

Constructs can be identified by either SME's or psychologists. The SME's know the jobs but need to be taught the constructs. On the other hand, psychologists know the constructs but seldom have a complete understanding of the jobs. In our procedure, SME's link the work behaviors and constructs since it appeared that it would be easier to teach the constructs to the SME's than to teach all the jobs to the psychologists. Constructs are selected from those identified in the professional literature (e.g., Ekstrom, French, & Harmon, 1976; Marquardt & McCormick, 1974; Theologus & Fleishman, 1971). Each construct is named and defined so as to distinguish it from other constructs. The group of SME's learns the meaning of the constructs and then rates each construct on the list on its relevance to each critical work behavior.[2] They also note which abilities are needed for entry-level positions

[2]Analysis of the time spent and importance ratings obtained earlier allows the development of a much shorter list of "critical" work behaviors. Critical work behaviors are those rated highest on the importance and/or time spent scales.

(minimum standards) and which abilities differentiate superior from barely acceptable workers (to be used for the ranking of applicants). This procedure is similar to that employed by Payne and van Rijn (1978) with a task inventory.

Criterion-Related Studies

The relationship between the construct, as measured by the selection procedure, and the related work behaviors must be supported by empirical evidence from one or more criterion-related studies. A concurrent validation strategy is likely be employed for this purpose. The number of studies will depend upon the degree of similarity among jobs, the availability of resources, and technical considerations such as the number of incumbents. A performance criterion must be developed for each job.

A separate criterion-related study does not have to be performed for each job. Since the ability to generalize validity among jobs is a controversial topic, we refer to a source considered by most to be rather conservative on it, the Uniform Guidelines on Employee Selection Procedures (USEEOC, USCSC, USDOL, USDOJ, 1978). The Guidelines state that if a study pertains to a number of jobs having common critical or important work behaviors at a comparable level of complexity, and if evidence of criterion-related validity is presented for one job, the construct validity can be generalized to the other jobs in the study [Guidelines, Section 14D(4a)]. Therefore, occupations will be compared to determine the extent to which they share similar critical work behaviors. Although the Guidelines say only that jobs must have one or more work behaviors in common [Guidelines, Section 14D(4b)], we did not interpret this to mean that just one work behavior in common between a job with validity demonstrated through a criterion-related study and a job without such evidence entitles one to generalize validity from the former job to the latter. We interpreted this more strictly to mean that one or more common work behaviors must be present for each construct. Thus, the demonstrated validity for one job can be generalized to another job to the extent that the jobs share constructs and have similar work behaviors associated with each construct. As an example, assume that job #1 is found to require verbal, quantitative, and spatial abilities. (Use Table 2 in conjunction with the following text.) Job #2 is found to require verbal, quantitative, and reasoning abilities. Job #3 is found to require verbal and quantitative abilities. Assume also that a criterion related study has shown that the verbal, quantitative, and spatial predictors are valid for job #1. If jobs #2 and #3 share work behaviors with job #1 for verbal and quantitative abilities, then the criterion-related validity found for those predictors for job #1 can be generalized to jobs #2 and #3. In no circumstances could the criterion-related validity found for those precictors for job #1 be generalized to the use of a reasoning predictor for job #2. To support the validity of a reasoning predictor for job #2, another criterion-related study would have to be done either on job #2 or on another job using the reasoning predictor and having one or more similar work behaviors with job #2 for the reasoning construct.

Conclusion

The authors have described a methodology for documenting the validity

of selection procedures for a number of similar jobs. This methodology can be applied in any organization having a large number of jobs. The authors recognize other possibilities for conducting a job analysis and validating a test. In addition, parts of the described methodology are interchangeable with other techniques. Obviously each situation must be examined in order to determine the applicability of this or any methodology.

REFERENCE NOTES

1. Pearlman, K., & Schmidt, [illegible]. L. Task differences as moderators of test validity: A disconfirmation. Paper presented at the annual conference of the International Personnel Management Association Assessment Council, San Diego, June 11-14, 1979.

2. Christal, R. E. The United States Air Force occupational research project (Report No. AFHRL-TR-73-75). Lackland Air Force Base, Texas: Occupational Research Division, Air Force Human Resources Laboratory, 1974.

3. Morsh, J. E., & Archer, W. B. Procedural guide for conducting occupational surveys in the United States Air Force (Technical Report 67-11). Lackland Air Force Base, Texas: Personnel Research Laboratory, Aerospace Medical Division, Air Force Systems Command, September, 1967.

REFERENCES

Ekstrom, R. B., French, J. W., & Harmon, H. H. Manual for kit of Factor-referenced Cognitive Tests. Princeton, New Jersey: Educational Testing Service, 1976.

McKillip, R. H., Trattner, M. H., Corts, D. B., & Wing, H. The Professional and Administrative Career Examination: Research and Development (PRR 77-1). Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center, April, 1977. (NTIS No. PB 268 780/AS)

Marquardt, L. D., & McCormick, E. J. The job dimensions underlying the job elements of the Position Analysis Questionnaire (PAQ) (Form B) (Report No. 4). West Lafayette, Indiana: Occupational Research Center, Purdue University, 1974.

Melching, W. H, & Borcher, S. D. Procedures for constructing and using task inventories (Research and Development Series No. 91). Columbus, Ohio: The Center for Vocational and Technical Education, The Ohio State University, March, 1973.

Payne, S. S., & van Rijn, P. Development of a written test of cognitive abilities for entry into the District of Columbia fire department: The task-ability test linkage procedure (TM 78-5). Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center, 1978. (NTIS No. PB 294 028)

Theologus, G. C., & Fleishman, E. A. Development of a taxonomy of human performance: Validation study of ability of scales for classifying human tasks (Technical Report No. 10). Washington, D.C.: American Institute for Research, 1971. (NTIS No. AD 736 184)

U.S. Department of Labor. Dictionary of Occupational Titles (4th ed.). Washington, D.C.: U.S. Government Printing Office, 1977.

U.S. Equal Opportunity Commission, U.S. Civil Service Commission, U.S. Department of Labor, & U.S. Department of Justice. Uniform guidelines on employee selection procedures. Federal Register, 1978, 43(166), 38295-38309.

van Rijn, P. Job analysis of entry-level firefighting in the District of Columbia fire department: A duty/task approach (TM 77-5). Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center, June, 1977. (NTIS No. PB 274 106/AS)

(590)

TABLE 1

Partial Description of Pathology Technician Job in Three Levels: Task, Work Behavior, and Duty

Tasks	Work Behaviors	Duties
Conduct donor medical history interviews. Prepare donor medical history cards. Check and record donor's weight. Check and record donor's pulse. Check and record donor's blood pressure. Check and record donor's hemoglobin/ hematocrite.	Obtain and record medical data from patient/donor.[a]	Perform blood banking activities.
Brief donors on blood collection procedures. Position donors for blood collection. Conduct phlebotomies. Clamp and cut blood collection tubes. Remove needle assemblies from donors. Apply bandages to donor venipuncture sites.	Draw blood samples from patient/donor.	
Conduct RH typing. Conduct DU tests. Conduct antiglobin tests. Conduct genotyping. Conduct cold antibody tests. Conduct antibody presence/identification tests.	Perform blood analysis.	
Select hematology analysis procedures.	Select hematology analysis procedures.	Perform hematology analysis procedures.
Prepare bone marrow smears. Label bone marrow smears. Stain bone marrow smears. Examine bone marrow smears.	Prepare and conduct analysis of bone marrow smears.	
Conduct macroglobin screening tests. Conduct fibrinogen screening tests. Conduct cryoglobin screening tests.	Conduct screening tests.	

Note. These lists are examples only. They are not complete descriptions of the job.

[a]The term "patient/donor" is used because the same work behavior is performed on patients and blood donors. Use of the slash shows that the behavior is not restricted to one situation.

TABLE 2

Example of Relationships Among Job Abilities, Work Behaviors, and Validity of Predictors

	Ability			
Job	Verbal	Quantitative	Spatial	Reasoning
	Work Behaviors			
1	A,B,C,D	E,F,G,H	I,J,K,L	
2	A,C,M,N	E,F,O,P		Q,R,S
3	C,D,T,U	G,V		
	Validity of Predictor[a]			
1	Criterion	Criterion	Criterion	
2	Genl.	Genl.		None
3	Genl.	Genl.		

[a]Criterion: Validity demonstrated through a criterion-related study.
Genl.: Validity demonstrated through generalization from the criterion-related study on job #1 because of overlapping work behavior(s).
None: Validity not yet demonstrated.

514

APPENDIX

Aid/Technician Jobs

Park Technician
Radio Operator
Biological Technician
Nursing Assistant
Medical Aid (sterile supplies)
Rehabilitation Therapy Assistant
Nuclear Medicine Technician
Medical Technician
Pathology Technician
Medical Radiology Technician
Therapy Radiology Technician
Medical Machine Technician
Pharmacy Technician
Orthotist and Prosthetist
Dental Assistant
Dental Laboratory Aid and Technician
Health Aid and Technician
Engineering Technician
Surveying Technician
Engineering Drafting
Museum Specialist
Physical Science Technician
Hydrologic Technician
Meteorological Technician
Cartographic Technician
Geodetic Technician
Math Technician
Cryptanalysis

LIPSCOMB, M. Suzanne., Air Force Human Resources Laboratory, Manpower & Personnel Division, Brooks AFB, Texas.

UTILIZATION OF WOMEN IN THE AIRCRAFT MAINTENANCE CAREER FIELD (Fri A.M.)

As current Air Force policy has opened many traditionally all-male specialties to women, it has become increasingly important to the Air Force to have detailed management information concerning how females are actually being utilized in the operational environment. In late 1977, the Air Force Human Resources Laboratory conducted an exploratory study of female aircraft mechanics. The study involved analysis of on-hand data collected during a routine occupational survey conducted by the Air Force Occupational Measurement Center. The analysis identified numerous unanswered questions and problem areas regarding femal utilization. Presently there are approximately 2000 women working in the Aircraft Maintenance specialty. This makes possible one of the most definitive studies ever conducted on females working in non-traditional areas. This study will analyze data pertaining to all individuals entering the aircraft maintenance career field since it was opened to females in 1972 and evaluate utilization patterns of those currently working in the specialty.

UTILIZATION OF WOMEN IN THE AIRCRAFT MAINTENANCE CAREER FIELD

M. SUZANNE LIPSCOMB

Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235

During the past decade, the number of women in the Air Force has increased from approximately 12,000 to more than 60,000. This has taken place while the total active duty force strength has declined by approximately 30%. There are now over 20,000 women working in job specialties which were once considered traditionally male jobs, reflecting a substantial shift in the distribution of sexes across jobs. For example, the Air Force currently has more women working in aircraft maintenance than in personnel specialties.

The substantial increase in numbers of women and their assignment into many traditionally all-male career fields has made it increasingly important for the Air Force to have detailed management information concerning the characteristics of successful and unsuccessful women, the expectations and attitudes of women, and the performance and utilization patterns of women in the operational environment.

While women are being assigned to many different specialties, there are presently 1,700 women in the Tactical Aircraft Maintenance and Airlift/Bombardment Aircraft Maintenance specialties, comprising approximately 4% of the population of these two specialties. The large number of women in these traditionally male career fields provides a medium for a comprehensive study of women in non-traditional areas. In 1978, Bergmann & Christal reported a preliminary study investigating the utilization of women in the Aircraft Maintenance career field (Bergmann & Christal, 1978). Using occupational data collected by the Air Force Occupational Measurement Center (AFOMC), a job-type analysis and an analysis of aptitude distributions were conducted. Even though the data had not been gathered to investigate the utilization patterns of women and, the sample of females was small, the study did produce some interesting findings.

Results from this study indicated that there were differences in task assignment as a function of gender. In the same specialty, a higher percentage of the males were found to be doing actual maintenance tasks while a higher percentage of females were doing support tasks. The data also suggested that during the first enlistment there was a movement of individuals from maintenance to support tasks. However, this movement appeared to be much larger for females than males. While few differences were found in the tasks performed by males and females working in maintenance functions, differences were found in the tasks performed by men and women working in support functions. Women spent more of their time than did men performing administrative or clerical tasks. These tasks were rated by supervisors as being more difficult than those tasks performed by men in either maintenance or support jobs. Finally, no significant differences were found in job attitudes of males and females in either support or maintenance functions.

The analysis of aptitude distributions conducted in this study suggested that the mechanical aptitude requirements, used for initial classification and assignment and historically predictive of success for males, might not be appropriate for females. Many of the women in the sample did not qualify for entrance into the career field with their scores on the mehanical aptitude test but were admitted into the career field during the period when entry was allowed based on either mechanical or electronic aptitude scores. Nevertheless these individuals had graduated from technical training school and were functioning in the career field and peforming maintenance jobs. However, no data were available on their performance levels or on individuals no longer in the specialty.

The findings of this exploratory research identified requirements for further investigation including the necessity of obtaining task performance data based on larger and more representative samples of males and females, information on the individuals entering the career field, those leaving the career field, and the jobs, attitudes and performance levels of those currently in the career field. Information obtained in these areas would provide Air Force management with a comprehensive picture of the utilization of women in the aircraft maintenance career field and allow planning to insure the optimun return on personnel investments. To provide that information, a comprehensive research program has been initiated covering five basic areas of concern: 1) the characteristics of the career field input population; 2) characteristics of those leaving the career field; 3) on-the-job utilization patterns; 4) job expectations and attitudes; and 5) on-the-job performance. This paper will outline the planning and progress of the research to date.

Method and Progress

In order to investigate the characteristics of those entering the aircraft maintenance career field, data available on historical tape files at the Air Force Human Resources Laboratory (AFHRL) will be analyzed. Education levels and aptitude scores will be examined as well as technical school success and promotion test scores. The variables of race or ethnic group, marital status, number of dependents, height, weight, and age will give an overall description of the men and women entering the career field. Those same variables will be examined for the men and women exiting the career field either by cross-training or by leaving the Air Force. For those cross-training out of the aircraft maintenance career field, their new specialty will be examined to determine its similarity or difference to aircraft maintenance and its traditionalism for males or females. Length of service in the career field and the Air Force will also be examined for gender differences.

In order to examine on-the-job utilization patterns within the aircraft maintenance career field, an occupational survey has been administered in conjunction with the routine survey of the career field by the Air Force Occupational Measurement Center. The sample selected for analysis of male/female differences included 100 percent of the women available and two men for every woman, matched on length of service.

The occupational survey, mailed out in April 1980, contained a background section in which job incumbents provided demographic information about themselves. Also included in this section were items concerning job satisfaction, reenlistment intent, days TDY, shift work and equipment useage. This section was followed by a task list covering 23 duties and 1045 tasks to assess the relative time spent on tasks performed by the respondent. The task list was followed by a final section addressing job expectations and experience. Data from this survey are in the initial analysis phases.

Responses to the task inventory will be analyzed along several dimensions. Job typing will be done to determine what jobs are being performed in the career field. The jobs which are identified will be classified as either maintenance or support and will then be examined to see if there are any differences in the utilization patterns of males and females. The tasks within each job type will be analyzed to see if there are any differences in the tasks being performed by males and females in the same job type. Routinely collected supervisors' ratings of task difficulty will be used to determine any differences in the average level of difficulty of the tasks being performed by males and females in each job type and across the career field. Other analyses concerning such factors as skill level distribution, average grade and amount of supervision will be performed. In all, these analyses will give an in depth view of how individuals are being utilized in the career field and identify any gender specific differences.

While this occupational information will in itself be very useful, it is also important to know how these utilization patterns change over time. In order to capture these changes, the job inventory will be administered a second time, 12 months after the first administration, to the same sample of males and females. A Time 1-Time 2 analysis of the original data and the data collected in the second survey will be performed to discover any changes in the utilization patterns of personnel during the 12 month period. If there is, in fact, a more rapid movement of women than men from maintenance to support jobs the Time 1-Time 2 analysis will define that movement.

A questionnaire concerning job expectations, experiences, and attitudes was included as the final section of the occupational survey. The purpose of this section was to identify why the individual entered the career field, past interest and experience in the mechanical area, expectations about the job before entering the career field, and experiences since entering the career field. It is important to know if men and women enter the career field for the same or different reasons, have differing expectations, have different experiences on the job and how these factors effect such things as attitudes and intention to reenlist. The first items in the questionnaire were concerned with the reasons the respondent entered the Air Force, and the reasons they entered the aircraft maintenance career field. Respondents were also asked whether or not aircraft maintenance was their first career field choice and their career plans upon entering the Air Force. These questions were raised by a study conducted in 1974 in which over half of the women in a sample of technical school trainees reported they had chosen aircraft

maintenance because it was the only field open when they enlisted (Longridge, 1974). Also relating to attitudes held before entering the career field, respondents were asked about their prior experience and interest level in the mechanical area. As an indication of current interest in the area, they were asked if they would try to find a job in a mechanical field if they were to leave the Air Force.

In order to examine the expectations held by men and women before they entered the career field, a series of questions were asked concerning how their work compared to what they had expected. The items covered technical difficulty, physical strength requirements, workload, and environment. These same areas were also covered by questions relating to how the respondents' job had changed since entering the career field. A series of questions concerning job satisfaction were asked followed by questions about the amount of assistance required on their job and specific problems with technical tasks and physical requirements. The final items in the questionnaire concerned supervisor attitudes and a section concerning the male/female composition of the work group and the perceived effectiveness of that work group.

The responses to this questionnaire will be analyzed for overall male/female differences. The data will also be analyzed by job type wherein respondents will be grouped according to the kind of job they perform as defined by responses to the task inventory. Thus, any male/female differences in questionnaire responses can be analyzed separately for those in maintenance functions and those in support functions. This should prove particularly useful in the analysis of those items directly influenced by the type of work performed.

The final phase of this research effort will be to assess on-the-job performance and identify predictors of performance in mechanical career fields. An integrated system of performance assessment applicable to mechanical career fields will be developed and applied in the field. Job knowledge, task performance, and supervisory rating data will be used to assess levels of job performance. Analysis of these data is intended to identify predictors of performance for all individuals in a mechanical career field and to identify any differences attributable to gender.

The overall research effort outlined here will provide a comprehensive view of the utilization of women in a non-traditional career field, and answer many questions about the men and women entering and leaving such a career field. In many ways, the study will provide meaningful information to those managing this and similar career fields and provide information relevant to the more general issue of integrating the female workforce into non-traditional specialties.

Reference

Bergmann, J. A. and Christal, R. E. Female utilization in non-traditional areas. Proceedings, 20th Annual Conference of the Military Testing Association, Oklahoma City, October - November 1978, 444-463.

Longridge, T. M., Jr. Training of Enlisted Women as Jet Aircraft Maintenance Specialists. (Sheppard PR 74-4). Sheppard Air Force Base, Texas: Training Research Applications Branch, USAF School of Applied Aerospace Sciences, August 1974. (AD/A - 006434).

LOOPER, Larry T., Air Force Human Resources Laboratory, (AFHRL/MDMD), Brooks AFB, Texas.

MODELLING APPROACHES FOR THE OPTIMAL ALLOCATION OF RECRUITING RESOURCES (Tue A.M.)

This paper will look at the development of an optimization model for Air Force Recruiting Service to aid in their recruiting resource allocation decision process. Two optimization models have been developed. One of these is a contractor effort using a Markov process and the other, which will be the primary focus of this paper, is an in-house model using a nonlinear market response function coupled with a dynamic programming routine to allocate recruiter effort at the office level. An extensive data base has been constructed consisting of demographic, economic, and recruiting information. Presently, Air Force Recruiting Service is supporting an update of the data base and intensive evaluation of the optimization model for possible use as an operational resource utilization decision aid. This paper will focus on the model development as well as actual and proposed applications of the model to the resource utilization decisions necessary to meet Air Force personnel input requirements.

RECRUITING RESOURCE AND GOAL ALLOCATION DECISION MODEL

L.T.Looper, AFHRL and C.A.Beswick, U.of S.Carolina

I. BACKGROUND

This report discusses a research effort conducted by the Air Force Human Resources Laboratory (AFHRL) in support of a formal request by the Market Analysis Directorate of the Air Force Recruiting Service. The request was to conduct a research program exploring the possible uses of mathematical algorithms and allocation techniques to serve as aids in the resource and goal allocation decision process.

The initial phase of this project involved the examination of the applicability to the Air Force recruiting problem of the non-linear regression/dynamic programming allocation model developed by Beswick, 1973. Under contract, Dr. Beswick assisted in the application of his model and provided much invaluable technical input to the research and the writing of this report. The model, after it was delivered to AFHRL, was considerably modified with respect to both function and output. These changes resulted from suggestions made by the Air Force Recruiting Service.

Currently, the model is being further tested with guidance from the Recruiting Service. The crucial area of data determination and collection is in a state of flux, and of course, any validity or reliability of the model depends on having an accurate, relevant, and up-to-date data base. Closely related to this line of research is a contractual effort supported by AFHRL which is attempting to take a different approach toward the resource and goal allocation problem. In that contractual effort, a Markov-type probability state flow model is being developed as a decision aid, and a prototype model should be completed by December 1979. This report will concentrate on the need for, the development of, and the use of the model developed at AFHRL (using Beswick's model as a prototype) as part of the overall effort to meet the needs of the Recruiting Service.

II. NEED FOR THE RESEARCH

The reality of the All-Volunteer Force and the accompanying decreased benefits available to enlistees (e.g., the "G.I. Bill"), along with strong recruiting efforts by other services, require the Air Force to consider carefully the most efficient use of all its resources in order to attain the force posture levels necessary to carry out its mission. The primary resource is manpower and its allocation. The size and scope of Recruiting Service operations (some 1,600 recruiters and 1,000 officers) make it imperative that at least one type of quantitative tool be analyzed to determine its applicability in making the resource and goal allocation decision.

Such a need does not mean that no previous work exists in this field or that the Air Force does not currently use any formal method. On the contrary, Bennett and Haber (1972, 1973, 1974), Kelly (1972), Babiskin, Grissmer, and Sterrett (1974), and Hamblin (1974) have explored factors relevant to Army and Navy recruiting resource allocation, especially that of recruiter themselves. Arima (1976, 1978a, 1978b) has performed an excellent systems analysis of Navy recruiting, and his recently-published study (1978a) looks at the effectiveness of Navy advertising.

The Air Force itself is using a well-developed, formal approach to the determination of recruiting goals. The model and the goals for all its various recruiting programs are published periodically in the Air Force Recruiting Service Market Analysis Handbook. The approach is a straightforward linear regression, but it does not attempt to directly allocate resources. The resource allocation decision is made at several organizational levels based on on-site visits and other recruiting data collected by Recruiting Service.

The increasingly stringent budget limitations being placed on the Air Force make it imperative that methods be utilized which will result in more efficient allocations and more effective attainment of goals. This report presents a decision model developed for Air Force Recruiting Service and discusses its use as a decision aid in making the following decisions:

1. Recruiter Allocation
2. Total Recruiter Strength
3. Office Location and Boundaries
4. Assignment of Recruiting Goals
5. Forecasts and Performance Evaluation

III. DEVELOPMENT OF THE RESPONSE FUNCTION

In the most general sense, the task of an Air Force recruiter is one of "selling" the Air Force to prospective enlistees. Like private-enterprise sales personnel Air Force recruiters have many administrative matters to handle, but just as their civilian counterparts, the Air Force recruiters' main job is contacting individuals (clients) about joining the Air Force. If Air Force recruiting is similar to commercial individual sales marketing, then some of the research on sales force allocation in the business world should be adaptable to the Air Force sales force decisions.

Each of these decisions is based on an understanding of how the recruit market *responds* to recruiting effort and what the determinants of this response might be. Sales force response functions are typically based on small segments of the market called control units. A control unit is the smallest sub-unit of the market used for analysis. Depending on the nature of the selling situation, it may be an individual customer, a group of customers, or a small geographic area. For the Recruiting Service, an ideal control unit might be the county or school district. Because of data availability, however, the model considered in this report is based on Air Force office structure. Estimation of a control unit's response to varying levels of selling (recruiting) effort presents a major challenge because response is influenced by a variety of interacting factors, including market potential, recruiting effort and experience, Air Force advertising effort, and other variables such as the economy.

Some sales force researchers have used subjective estimates to develop separate response functions for each control unit (Little, 1970; Lodish, 1971; Montgomery, Silk, & Zaragoza, 1971). Subjective estimates of sales response make allowance, at least implicitly, for the interacting factors just mentioned. Statistically developed multivariate response functions have been employed in the areas of setting sales force size (Lambert & Kniffin, 1970), evaluating salesman performance (Cravens & Woodruff, 1973; Cravens, Woodruff, & Stamper, 1972), and in allocating selling effort (Beswick, 1977).

The way a control unit responds to recruiting effort is determined by the many interacting factors earlier. If the number of reservations is chosen as the measure of market response, the response relationship may be written in functional notation as:

$$R = f(E, W, M, A, X, PR, 0) \quad (1)$$

This is, response (number of reservations) in a control unit (recruiting office) is a function of recruiting effort (E), workload (W), market potential (M), advertising effort (A), recruiting experience (X), prior reservations (PR), and other variables (0).

Two fundamental problems face the researcher attempting to determine the *specific* function which describes office response. First, the choice of appropriate measures for the variables indicated in equation (1) is by no means obvious. For example, what measure should be used for

reservations; Should it be total reservations or net reservations (after training attrition) or quality reservations? Secondly, the problem of identifying the nature of the functional interaction between these measures is a difficult one: Which of the infinite number of possible functions f should be investigated empirically?

In particular, a specific functional form of equation (1) would be desired which would be applicable in a wide variety of situations and yet specific enough to incorporate knowledge gained from prior sales force work and from conceptual insights into the effectiveness of the recruiting decision variables in the context of Air Force recruiting (e.g., diminishing returns).

This report hypothesizes a functional relationship of the type described in the previous paragraph. This relationship is called the multiplicative factor model because it treats recruiting response as the product of a series of factors representing the relative strengths of the underlying causal variables of equation (1). This response function can be written as follows:

$$r_i = z_i t_i^a + C \qquad (2)$$

where

r_i = number of reservations in office i

z_i = f(W, M, A, X, PR, 0) as defined earlier in equation (1)

t_i = work-months of effort (E) in office i

a = response elasticity for E

C = constant term (from regression)

The effort factor (E) describes market response to recruiting effort; it is a response function of the general form shown in Figure 1, although S-shaped curves, or curves which decline beyond saturation levels of recruiting effort are also possible. When combined with the variables of z, as in equation (2), E defines an effort-response relationship which considers control unit differences in recruiting effort and in office and environmental variables (z) which are superimposed on the effort-response relationship be means of equation (2) to give a complete estimate of control unit response.

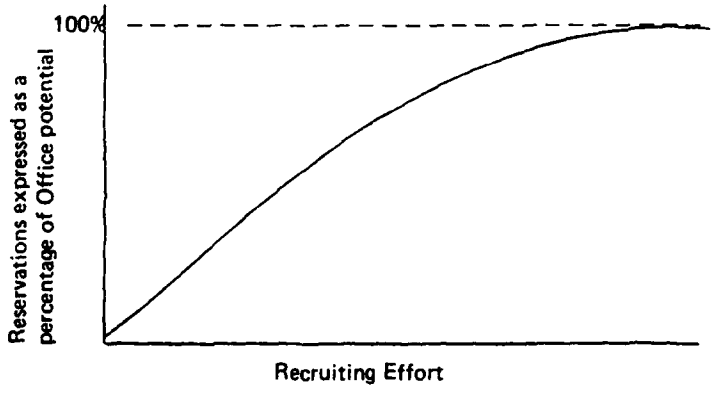


Figure 1. **Recruit market response to recruiting effort.**

Many data items and variables were examined as possible candidates for inclusion in the model of equation (2). It was decided to include all offices (subject to data availability) in building and testing the model. After merging data from several different sources, complete records were available for 807 offices. These offices were split into two groups: 538 were used to build the response function, and 269 were reserved for test purposes. Table 1 lists these data items and variable measures. Because of the basic functional form of the response function chosen in equation (2), non-linear regression analysis was then used to determine the precise functional form.

Table 1. **Variables Used in Model Development**

Factor	Variable	Definition
Performance	RES	Male Non-prior Service Reservations (April 77 – March 78)
Recruiting Effort	MME	Work-Months of Effort (estimate)
	RAS	Number of Recruiters Assigned
	RAUT	Number of Recruiters Authorized
Market Potential	HSM	Number of High School Seniors (Male)
	QUAL[a]	Percent Passing ASVAB [Armed Services Vocational Aptitude Battery] Mental Ability Test
	INT[a]	Percent Stating Interest in Military Career (ASVAB Test)
	UNEMP[a]	Percent Perceiving Difficulty in Employment (National Sample of 10,000)
	POP[a]	Population – Ages 16 to 21
	MINOR[a]	Minority Percent (ASVAB Test)
Advertising Effort	LEADS[a]	Number of Qualified Leads from National Advertising
	AFB	Number of Air Force Bases
Prior Experience (Reservations)	ACC	Male Non-Prior Accessions (April 76 – March 77)
Recruiting Experience	EXP	Average Recruiter Experience (Months)
Workload	SQMI	Square Miles Covered by Office
	NHS	Number of High Schools
	AHSS	Average High School Size

[a]These variables are available only by flight (i.e., data are not available at the office level).

In the initial stages of analysis, multiple linear regression was employed to investigate linear relationships among the variables in the data base and to aid in selecting variables to be included in the non-linear analysis. The complexities inherent in using a non-linear least squares procedure and the difficulty in using the program to select variables for inclusion in the model make it highly desirable to limit the number of variables analyzed by the procedure.

In an attempt to limit the variables, the first step in non-linear analysis was to use stepwise linear regression to analyze a logarithmically transformed version of equation (2). Although the models developed by this method may not be the best (in the least squares sense) fit of equation (2), they are close approximations to the best model, and it was felt that the capability of the stepwise procedure to select appropriate variables for inclusion in the model outweighed the non-optimal nature of the models developed. The Gauss-Newton non-linear least squares procedure was employed to develop the final recruiting response function which will be used in the decision analysis model for the Air Force Recruiting Service. This response function, which is the product of the non-linear regression analysis, is of the following form (variables are from Table 1):

$$R_i = (MME_i)^{.649} * 1.96(HSM)^{.13} * (INT)^{.14} * (AHSS)^{-.09} * (LFADS)^{.04} * (ACC)^{.69} * (EXP)^{.11} + 5.85(AFB) + 21.8 \quad (3)$$

Where:

1. The constant term value 21.8 has been added to improve the empirical fit of the multiplicative factor model. It simply raises the level of the entire response function shown in Figure 1 by a constant amount. Alternatively, it is the minimum number of recruits expected from any office.
2. AFB is included as a linear term because for many offices this variable is zero and hence is not suitable for inclusion in the multiplicative factor model. Other exogenous factors which affect only a limited number of offices could be treated in a similar fashion.
3. All of the multiplicative variables, except MME and HSM, have been indexed by dividing by the respective means for all offices; this allows reservations to be predicted under assumption of average levels of any of these variables by substituting a "1" for the indicated variable.
4. Only the variables indicated in equation (3) were investigated as part of the non-linear least squares analysis. Other variables were excluded from this final stage of analysis because they did not contribute significantly in the log-linear analysis or because they were highly correlated with another variable in the model (e.g., RAS was highly correlated with MME). The highest pairwise correlation between variables remaining in the model was .59 (MME, HSM).

The model given by equation (3) achieved an R^2 of .72. When model predictions for the test sample (269 observations) were compared to actual reservations, an R^2 of .68 resulted. These results are a good indication that the Multiplicative Factor Model does provide the estimates of market response necessary for quantitative analysis of recruiting service decisions and serves as proper input to the allocation phase of model development.

The problem of allocating recruiting effort for a fixed number of recruiters (N) can be formulated as follows:

$$\text{Maximize} \quad \sum_{i=1}^{n} r_i$$

$$\text{Subject to:} \quad \sum_{i=1}^{n} t_i = 12N \quad (4)$$

That is, maximize the total reservations in all n offices subject to the constraint that a fixed number of recruiters is available. This process can be repeated for increasing values of n until the total Air Force reservation goal for any recruiting program is reached. That is, repeat expression (4) for increasing n until

$$\sum_{i=1}^{n} r_i \leq G \quad (5)$$

where G is the goal for total reservations. Thus, this process will determine the minimum number of recruiters required to achieve a given reservation goal and will indicate how these recruiters should be allocated to offices.

Beswick (1977) describes a dynamic programming algorithm which can be employed to perform the procedure specified in the previous paragraph. This algorithm with several modifications was used to allocate recruiting effort for the 807 offices for which data were available. The total number of recruiters

was initially fixed at current levels. Model output for two squadrons (71 offices) is presented in Tables 2 and 3. The columns in these tables (from left to right) are the office, the work-months of recruiting effort for the current year, the optimum level of recruiting effort specified by the algorithm, the number of reservations predicted by equation (3) using the current level of effort, the number of reservations predicted at optimal levels, the actual reservations during the past year, the current level of recruiter strength, and the optimal levels of recruiter strength by office, flight, and squadron.

IV. USE OF THE MODEL

Several modifications were made to the model developed by Beswick (1977) to align it more closely to Air Force recruiting needs and to simplify its use as a problem solving and decision making tool. First, the model was made interactive via the use of a computer terminal. The model asks which of several recruiting accession programs the user is interested in studying. These accession programs are non-prior service male, non-prior service female, prior service male, prior service female, Officer Training School – male, Officer Training School – female, health services – doctor and health services – nurse. The Air Force has specific recruiting goals in each of these programs and also dedicates a portion of its recruiting force to each program.

The model then asks for the estimated number of recruiters, the recruiting goal for the particular program under analysis, and the number of recruiting offices. It should be noted here that the number of offices is fixed for any one model run. The allocation takes place among existing offices and is not concerned with opening new offices other than by indicating that certain areas need increased effort allocation. This will be seen in the output analysis section.

The user is then requested to input a factor of recruiting difficulty which is a perceived percentage of increasing or decreasing returns to recruiting effort, an upper and lower percentage limit on office manpower increase or decrease, and a percentage factor indicating amount of total recruiting effort spent on this accession program. After these data are input by the user, the model then performs the dynamic allocation and, upon completion, indicates to the user whether the goal could be achieved with the input number of recruiters or, if not, how many recruiters it would take to achieve the accession goal.

Once the model output is complete, the user may then examine the printout in office level, flight level, or squadron level form. As an example, consider the output presented in Tables 2 and 3 for Squadron 16 and Squadron 18, respectively. Squadron 18 shows an excess of 282 work-months (23 recruiters). Moreover, the response function predicts that if these changes can be implemented, 282 additional reservations (2271 – 1989) can be gained in Squadron 16, while Squadron 18 will lose 287 (2423 – 2136). Changes in expected reservations are always calculated between current and desired predicted reservations. Comparisons with actual figures can be misleading due to random model variation. Most importantly, the model output provides an office level analysis which can be used to plan these desired changes.

According to the model, office 16AA needs three more recruiters, while 16BE, 16BF, and 16CE should be considered as candidates for closings. Other Squadron 16 offices show substantial untapped market potential. The model indicates that at least double the current levels of effort is desirable in offices BA, BB, CC, CD, and EE.

While Squadron 18 shows the need for substantial net reduction in effort, there are still five offices (FC, FE, FG, FH and GB) which require at least one more recruiter. In general, though, most Squadron 18 offices require some reduction in effort levels and several should be considered as possible closeouts.

Many of the desired allocations can be achieved by reassigning or adding new recruiters as indicated previously. Another method is adjustment of office boundaries. This technique would, in

608

Table 2. **Office Level Analysis of Recruiting Effort – 1600 Recruiters**
(Squadron 16)

Office	Effort		Reservations			Recruiter Strength	
	Current	Desired	Current	Desired	Actual	Current	Desired
Flight 16A							
16AA	12	49	63	123	90	1	4
16AB	12	2	35	26	42	1	0
16AC	12	23	59	76	65	1	2
16AF[a]	10	19	48	62	57	1	2
16AG	6	5	33	32	32	1	0
16AH[a]	10	19	48	62	57	1	2
Totals	62	118	286	381	343	5	10
Flight 16B							
16BA	24	52	87	130	85	2	4
16BB	36	75	119	177	86	3	6
16BC[a]	20	17	61	57	59	2	1
16BD[a]	20	17	61	57	59	2	1
16BE	6	0	24	0	31	1	0
16BF	12	0	23	0	41	1	0
16BG	24	13	62	49	56	2	1
Totals	142	174	437	471	417	12	14
Flight 16C							
16CA	30	16	72	55	58	3	1
16CB	12	12	47	47	55	1	1
16CC	12	60	66	147	91	1	5
16CD	24	80	98	189	72	2	7
16CE	12	0	32	0	50	1	0
16CF[a]	18	31	67	86	65	2	3
Totals	108	199	381	523	391	9	17
Flight 16D							
16DA	46	42	114	109	139	4	3
16DB	12	10	45	43	44	1	1
16DC	24	10	58	42	73	2	1
16DD	24	13	62	49	55	2	1
16DE	18	9	51	41	58	2	1
16DF	12	20	51	63	35	1	2
16DH	12	3	38	28	30	1	0
16DI[a]	20	18	62	59	62	2	2
Totals	168	125	483	434	496	14	10
Flight 16E							
16EA[a]	16	26	62	77	66	1	2
16EB	12	0	22	0	58	1	0
16EC	14	10	54	49	75	1	1
16ED	12	12	47	47	41	1	1
16EE	32	77	113	182	113	3	6
16EH	18	4	44	30	46	2	0
16EJ[a]	16	26	62	77	66	1	2
Totals	120	156	402	462	465	10	13
Squadron 16 Totals	600	771	1,989	2,271	2,112	50	64

[a]These are new and contain flight-average data.

Table 3. Office Level Analysis of Recruiting Effort – 1600 Recruiters
(Squadron 18)

Office	Effort		Reservations			Recruiter Strength	
	Current	Desired	Current	Desired	Actual	Current	Desired
Flight 18A							
18AA	26	29	78	83	90	2	2
18AB	36	25	88	74	75	3	2
18AC	14	14	51	51	41	1	1
18AD	36	10	70	43	82	3	1
Totals	112	78	287	251	288	9	7
Flight 18B							
18BA	20	30	70	84	83	2	2
18BB	24	8	55	38	66	2	1
18BC	42	14	81	51	134	4	1
18BD	12	9	44	41	36	1	1
Totals	98	60	250	213	319	8	5
Flight 18C							
18CA	18	17	59	58	72	2	1
18CB	22	8	54	38	87	2	1
18CC	12	4	38	30	29	1	0
18CD[a]	16	7	47	36	55	1	1
18CE	12	0	26	0	34	1	0
Totals	80	36	224	162	277	7	3
Flight 18D							
18DA	24	8	55	38	57	2	1
18DB[a]	26	11	62	44	53	2	1
18DC	34	12	71	47	69	3	1
18DD	22	5	49	32	33	2	0
18DE[a]	26	11	62	44	53	2	1
18DF[a]	26	11	62	44	53	2	1
Totals	158	57	360	249	318	13	5
Flight 18E							
18EA	24	3	47	29	56	2	0
18EB[a]	18	5	46	33	39	2	0
18EC	14	4	41	30	28	1	0
18ED	20	2	41	26	19	2	0
18EE	24	14	64	52	65	2	1
18EF	12	0	26	0	31	1	0
Totals	112	29	264	170	238	9	2
Flight 18F							
18FA	52	39	119	102	95	4	3
18FB	24	11	60	44	72	2	1
18FC	36	82	122	192	61	3	7
18FE[a]	36	48	105	122	76	3	4
18FG[a]	36	48	105	122	76	3	4
18FH[a]	36	48	105	122	76	3	4
Totals	220	276	615	705	456	18	23
Flight 18G							
18GA	30	31	85	87	65	3	3
18GB	24	33	77	90	71	2	3
18GC	22	11	36	31	16	2	1
18GD	36	17	79	56	61	3	1
18GE[a]	28	19	72	61	53	2	2
18GF[a]	28	19	72	61	53	2	2
Totals	168	129	423	385	319	14	11
Squadron 18 Totals	948	666	2,423	2,136	2,215	79	55

[a]These offices are new and contain flight-average data.

(610)

effect, "create" new offices. For example, if 18GB and 18GC are adjacent, it may be desirable to move responsibility from GC to GB, thus increasing effort in GB while reducing the effort in GC by one recruiter.

Absolute agreement with desired effort levels is not necessary, and in some cases not even desirable. Small differences (1 or 2 work-months) between actual and desired levels of effort are relatively unimportant (as well as being difficult to eliminate). Large changes in effort may also be difficult to achieve in a short period of time, especially since Air Force policies (and common sense) preclude excessive numbers of recruiters at a specific location. (Note: This run was made with an unlimited increase in office strength, which was done only to show model global optimality.) These difficulties may prevent full achievement of optimal allocation and hence prevent achieving the increase in reservations predicted by the model, but it does not lessen the desirability of moving in the direction of model allocations.

Output from the allocation model provides a convenient worksheet which can be used to plan recruiting operations for the following year. The use of the model in allocation and office boundary decisions has been discussed previously. When these effort decisions (which may differ from the model's desired allocations as discussed previously) are entered as "current" levels, the model predictions will then represent expected reservations for the office given planned effort levels.

As the check of model validity, allocations for the 18th Squadron were compared with a recent staff report on these offices prepared after extensive field study and analysis. The detailed recommendations in this report were compared line by line with model output. In every office where a change was indicated, the direction of change suggested by the report was the same as the model recommendation. Changes in the magnitude of effort also concurred with model output, although in three offices where large changes were indicated, the staff recommendations were slightly more conservative than the model.

As for possible additional uses of model output, recruiting goal assignments could be determined through use of the predicted reservations based on management's planned effort level for that office. If the predicted level is below Recruiting Service's goal, then some adjustment, either through the goal or through increased manpower, should be made. In the area of performance evaluation, model predicted output could be compared with actual productivity by office. Of course, no individual recruiter data are in the model, so performance would be by office, flight, and squadron comparisons.

In both performance evaluation and goal assignment, it is advisable to adjust the goal or standard to reflect situations not considered by the model. This process involves managerial analysis and judgement of specific office situations and hence is time consuming. However, this type of involvement greatly enhances model effectiveness.

Total recruiter strength could be determined simply by setting model input parameters at various levels and "gaming" the model until the desired goal is achieved under various manpower and difficulty factors. Such factors could be exogenously determined by Recruiting Service management. As a forecasting tool, the model would use next year's *planned* effort levels to determine predictions and allocations.

V. CONCLUSIONS

Air Force Recruiting Service is strongly supportive of this model and its related research, as well as the contractual effort previously alluded to, which will enlarge the scope of the resource and goal allocation decision model. As a prototype, this model has been sanctioned by Recruiting Service. Data collection and especially data base update stand as major obstacles to its effective

utilization. It is believed, however, that with continued support such a quantitative scheme for assisting in the resource and goal allocation decision process is one viable means for meeting Air Force accession goals and assuring the ability of the Air Force to remain mission ready.

REFERENCES

Arima, J.K. *A systems analysis of Navy recruiting.* Special Report 76-9. San Diego, CA: Navy Personnel Research and Development Center, April 1976.

Arima, J.K. *Advertising budgets, advertising effectiveness, and the Navy's recruiting advertising program.* NPS 54-78-009. Monterey, CA: Naval Postgraduate School, December 1978. (a)

Arima, J.K. *Determinants and a measure of Navy recruiter effectiveness.* NPRDC-TR-78-21. San Diego, CA: Navy Personnel and Research Development Center, June 1978. (b)

Babiskin, R., Grissmer, D., & Sterrett, R. *An analysis of the Gilbert youth survey for utilization in recruiting resource allocation.* McLean, VA: General Research Corporation, Operations Research Division, September 1974.

Bennett, J.T., & Haber, S.E. *On the application of marginal productivity analysis to the allocation of recruiters within the military service.* Washington, D.C.: George Washington University Program in Logistics, January 1972.

Bennett, J.T., & Haber, S.E. *Selection, deployment, and evaluaation of marine recruiters.* Washington, D.C.: George Washington University Program in Logistics, June 1973.

Bennett, J.T., & Haber, S.E. *The allocation of recruiters among spatial areas.* Washington, D.C.: George Washington University, 1974.

Beswick, C.A. *An aggregate multistage decision model for sales force management* (Doctoral dissertation, University of Tennessee, 1973).

Beswick, C.A. Allocating selling effort via dynamic programming. *Management Science*, March 1977, **23**(7), 667-678.

Cravens, D., & Woodruff, R. An approach for determining criteria of sales performance. *Journal of Applied Psychology*, 1973, **57**, 240-247.

Cravens, D., Woodruff, R., & Stamper, J. An analytical approach for evaluating sales territory performance. *Journal of Marketing*, January 1972, **36**(1), 31-37.

Hamblin, T.R. *Optimal allocation of Coast Guard District recruiting funds.* AD-A003 524. Monterey, CA: Naval Postgraduate School, December 1974.

Kelly, J.N. *Allocation of available recruitment resources and establishment of recruitment production goals.* Cambridge, MA: Management Analysis Center, Inc., December 1972.

Lambert, Z., & Kniffin, F. Response functions and their application in sales force management. *Southern Journal of Business*, January 1970, **5**, 1-11.

Little, J. Models and managers: The concept of a decision calculus. *Management Science*, April 1970, **16**(8), B466-B485.

Lodish, L. Call plan: An interactive salesman's call planning system. *Management Science*, December 1971, **18**(4), Part II, 25-40.

Montgomery, D., Silk, A., & Zaragoza, C. A multiple-product sales force allocation model. *Management Science*, December 1971, **18**(4), Part II, 3-24.

(612)

POST TRAINING SURVEYS IN IBM'S FIELD ENGINEERING DIVISION

Ed Magdarz, IBM Corporation, Dept. 817, Raleigh, NC

ABSTRACT

IBM's Field Engineering division has many field personnel servicing computer systems at customer locations across the country. They receive frequent training because of changes in products, technology, and maintenance methods. It is desirable to measure the effect of training on their job performance.

The Post Training Survey system is a recent approach to improving that measurement. The surveys use 'paper and pencil' questionnaires. They have both 'general' and 'job specific' questions. Question and response format has been carefully designed to increase feedback validity. A computer file with 'work experience' data helps select survey candidates. Two computer programs analyze the returned questionnaires. One program does an initial detailed analysis. The other summarizes data and maintains history. This paper describes the survey system's rationale, operation and early results.

RATIONALE

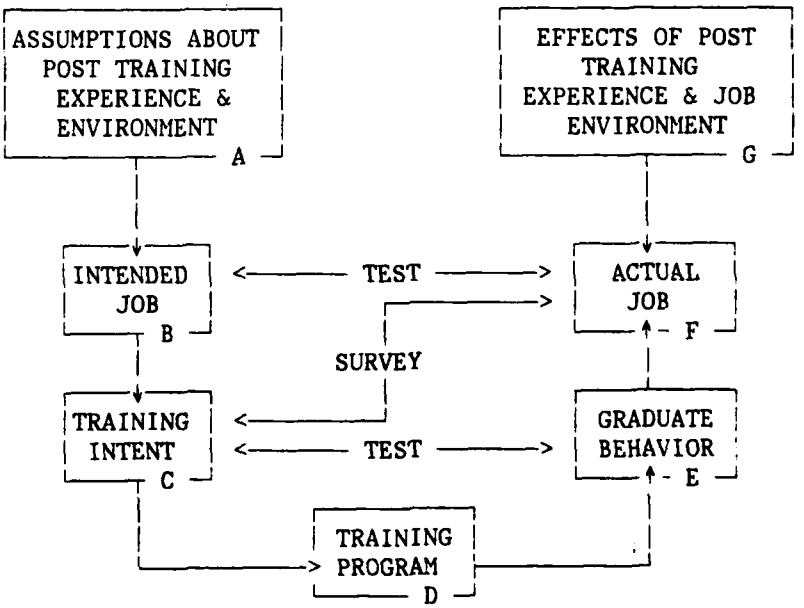


FIGURE 1

A conventional method of measuring training effectiveness is

to test graduates as they exit the training program. Their performance is compared to the training objectives/intent (Figure 1, Blocks C, D and E). This method cannot adequately account for the effects of post training job experience and of the job environment (Blocks F and G). Yet many training programs, by design, depend on the effects of post training experience and the environment (Blocks A, B and C).

It is more accurate to measure training effectiveness by testing graduates in the job environment after they gain job experience (Blocks B, C, D, E and F). Unfortunately this is not always practical. There are problems of cost and of trying to control for non-training variables that exist on the job. The Post Training Survey system was developed as a substitute for actual testing in the job environment. Instead of performance testing it measures the perceptions and opinions of the graduates (Blocks C, D, E and F).

OPERATION

INITIATION OF TRAINING

As a new product is being developed, a training program that teaches 'the job of maintaining the product' is also developed. Training begins when the new product is released to the marketplace and usually continues for several years.

The number of people trained to maintain a product depends on its size, complexity, distribution and quantity installed. Therefore the volume of maintenance training ranges from programs with a few graduates to programs that have several thousand graduates.

INITIATION OF SURVEY

The Post Training Survey is conducted early in the life of a training program. This allows timely identification of problems. The result is improved training for the remainder of the students. Ideally the survey occurs after thirty to fifty people have graduated and acquired job experience.

The amount of job experience needed to ensure a valid survey depends on the unique characteristics of each job and therefore varies from survey to survey.

Only programs that will produce a significant number of graduates are surveyed. We estimate that approximately forty surveys will be conducted each year.

SURVEY POPULATION

A record of each training program graduate is kept on a computer file. As the graduate accumulates experience on the new job, a record of his/her service incidents are also put in a computer file.

Survey candidates are selected according to their service experience. The computer files are searched for graduates that fall within maximum and minimum experience limits.

An attempt is made to obtain from thirty to fifty candidates for each survey.

QUESTIONNAIRE CONTENT

Figure 2 shows an example questionnaire from a typical survey. The questionnaire for any survey has three sections.

The general section has three questions covering the respondents overall perception of training effectiveness. These questions are the same for every survey. This allows comparison of the perception of effectiveness from one training program to another. Thus training programs can be ranked.

The specific section has up to twenty questions that cover job tasks. These questions, unlike those in the 'general' section, are different in each survey because they cover the tasks of the specific job being surveyed.

This section uses a unique response format that solicits the respondents':

- Perception of task importance.
- Opinion of preparedness provided by training.

The format allows the analysis program to identify six different results for each question as will be described later under 'Initial Analysis and Report'.

An optional section allows inclusion of up to five non-task related multiple choice questions.

SURVEY ADMINISTRATION

The department responsible for a new training program monitors the computer files for availability of survey candidates. When approximately thirty to fifty candidates are available, survey questionnaires are mailed to each

candidate's manager. The candidate completes the questionnaire and mails it back to the training department.

INITIAL ANALYSIS AND REPORT

Refer to Figure 3. Weekly, small groups of completed questionnaires reach the training department and are keyed-in to the computer system. Then the system prints an analysis of the questionnaires accumulated thus far. The report is designed for use by people with little or no experience in statistics. Therefore it calculates results as simple percentages and averages.

The example report in Figure 3 shows data typical of our early results and is interpreted as follows.

- The analysis is of thirty respondents.

- General Items: The overall perception of training effectiveness (ALL TRG) for the job is positive (mean = 2.54). However, the education center portion is perceived as more effective (mean = 2.27) than the CAI portion (mean = 2.80).

- Specific Items: There are six 'results' columns for each of this questionnaire's eight specific questions (items). The columns are normally shown side by side across the report. However in this example, because of space limitations, they are shown as two groups, the first with Columns 1 and 2 and the second with Columns 3 thru 6. Descriptions of the columns follow.

- Notice that below each column title, in parenthesis, is the response combination that the column counts. Refer to Figure 2 to review the 'two response per item' format. For example report Column 1, 'IMPORTANT', counts the response combination where 'yes' was circled for (Important?) and anything was circled for (Prepared?) thus (Y + ANY). Report Column 2 (ANY ?/BLANK), counts any respondent who circled a '?' or had a 'blank' for either (Important?) or (Prepared?). Report Columns 3 thru 6 count the various 'yes' and 'no' response combinations such as (Y + Y).

- Column 1 'Important': A low percentage for a item suggests that the data in the remaining five columns may not be significant. Notice that 100% of the respondents indicated that the job task related to Item 5 is important. Therefore the remaining five columns for Items 5 probably have significant data.

- Column 2 'Unsure of Importance or Preparedness': A high percentage for an item indicates that the data for the

remaining four columns is a reduced sample and therefore less valid. Respondents who are included in this column are not used in the analysis of the remaining four columns. Notice that none of the items had more than 1/3 of the respondents indicate 'unsure'.

- Column 3 'Important and Prepared': A high percentage in this column suggests that training for the related job task is effective. Notice Item 8 where only 30% of the respondents thought they were adequately prepared for a job task they considered important.

- Column 4 'Important but Not Prepared': A high percentage in this column suggests that training for the related job task is not adequate. Notice Item 8 where 50% of the respondents thought they were not adequately prepared for a job task they considered important.

- Column 5 'Not Important but Prepared': A high percentage in this column suggests overtraining on an unimportant job task.

- Column 6 'Not Important and Not Prepared': A high percentage in this column suggests that the item should not have been included in the questionnaire.

- Overall, the job tasks related to Items 5 and 7 appear to have the most effective training. Items 4 and 8 suggest that their related job tasks have the least effective training. Item 2 suggests that there may be overtraining on the related job task.

HISTORY ANALYSIS AND REPORT

The general section data from each survey is transferred within the computer from the Initial Analysis file to a History Analysis file where it remains for several years.

Refer to Figure 4. The example shows a report obtained from the History Analysis. It contains only two surveys. It is possible to request and analyze any amount of the surveys on the file. The first data line is interpreted as follows.

- The data is for Training Program 01234 which is the responsibility of Department 701. The survey was initiated on 01/23/80 and is still being administered (still 'on').

- For General Question 1 (Also refer to Figures 1 and 2) 30 respondents are on file. 14 of them responded favorably (i.e., good or very good) and 2 responded unfavorably (i.e., poor or very poor). This means that

46% responded favorably and 7% responded unfavorably.

- The complete results for General Questions 2 and 3 are not shown on this example because of space limitations.

The second data line is for another training program and is interpreted in a similar manner.

The third data line shows totals and averages. For example, the graduates perception of the overall effectiveness of Department 701's training programs, on the average, is 62% favorable. The History Analysis allows comparison of training programs and departments.

This concludes the description of the 'operation' of the Post Training Survey system. The results of early use of the system follow.

EARLY RESULTS

The results of three 'pilot' surveys are very encouraging. Each has revealed instances of unimportant job tasks, under-training and over-training that were not detected by other measures of effectiveness. The questionnaire return rate is very high, over 90%.

SUMMARY

The Post Training Survey system might be referred to as 'just another survey system'. However I believe it has the following unique features that broaden the use of surveys for measurement of training.

- Use of computer systems to select survey candidates and analyze initial and historic results.

- Concept of measuring 'job tasks' rather than training components (e.g., objectives) to evaluate training effectiveness.

- Concept of comprehensively selecting candidates according to multiple criteria (e.g., experience, etc.), thereby increasing the validity of response data.

- Use of two column response format that yields six types of results.

- Two levels of analysis - Initial (detailed) Analysis that includes task by task results, and History (summary) Analysis that allows ranking of training programs and departments.

** FIGURE 2 (START) ***** EXAMPLE QUESTIONNAIRE ***********

IBM FE EDUCATION - POST TRAINING SURVEY

Our records indicate that you completed 1234 Printing Subsystem training in Course Number 11322, Class Number ___.

Now that you have had time to apply the training that you received on the 1234 Printing Subsystem, you can help us by answering the following questions.

---

1. Considering all aspects of the training you received via CAI AND EDUCATION CENTER how do you rate the overall effectiveness of training? (Circle one.)

   a.Very Good b.Good c.So-So d.Poor e.Very Poor n.?

2. Considering all aspects of the training you received via CAI, how do you rate the overall effectiveness of training? (Circle one.)

   a.Very Good b.Good c.So-So d.Poor e.Very Poor n.?

3. Considering all aspects of the training you received at the EDUCATION CENTER, how do you rate the overall effectiveness of training? (Circle one.)

   a.Very Good b.Good c.So-So d.Poor e.Very Poor n.?

For each of the following items, please circle TWO answers on the right. Circle the ? if you are unsure about the Yes or No.

	Is this an IMPORTANT part of the job? (Important?)			Did training adequately PREPARE you? (Prepared?)		
1. Run and use results of 'tape M' microdiagnostics.	Y	?	N	Y	?	N
2. Retrieve and use error log and sense information.	Y	?	N	Y	?	N

** FIGURE 2 (CONTINUES) **********

** FIGURE 2 (CONTINUED) **********

	(Important?)	(Prepared?)
3. Use CE Panel functions to assist in diagnosing failures.	Y ? N	Y ? N
4. Develop and use simple microprogram routines to assist in diagnosing failures.	(Important?) Y ? N	(Prepared?) Y ? N
5. Diagnose failures not solved by use of the MAPs.	(Important?) Y ? N	(Prepared?) Y ? N
6. Perform removals, replacements, and service checks on 'coronas' as per FEMM.	(Important?) Y ? N	(Prepared?) Y ? N
7. Perform removals, replacements, and service checks on 'fuser backup rolls' as per FEMM.	(Important?) Y ? N	(Prepared?) Y ? N
8. Install 1234 as per FEIM.	(Important?) Y ? N	(Prepared?) Y ? N

---

1. In what percentage of the failures did the MAPs assist you in diagnosing the problem? (Circle one.)

   a. 100% b. 80% c. 60% d. 40% e. 20% n. ?

2. Approximately how many 01-18 calls have you taken on the 1234. (Circle one.)

   a. 0 b. 1-5 c. 6-10 d. 11-15 e. 16+ n. ?

If you wish to remain ANONYMOUS detach the cover page (if present) that contains your name and number.

THANK YOU for your cooperation. Please use return envelope provided and mail.

** FIGURE 2 (END) **********

620

** FIGURE 3 (START) ***** EXAMPLE REPORT ******************

POST TRAINING SURVEY - INITIAL PROCESSING REPORT
TRAINING PROGRAM: 01234
INITIAL INPUT DATE: 11/01/79 LAST INPUT DATE: 11/25/79
CURRENT ACCUMULATED NUMBER OF RESPONDENTS: 30

== GENERAL ITEMS (%) ==========

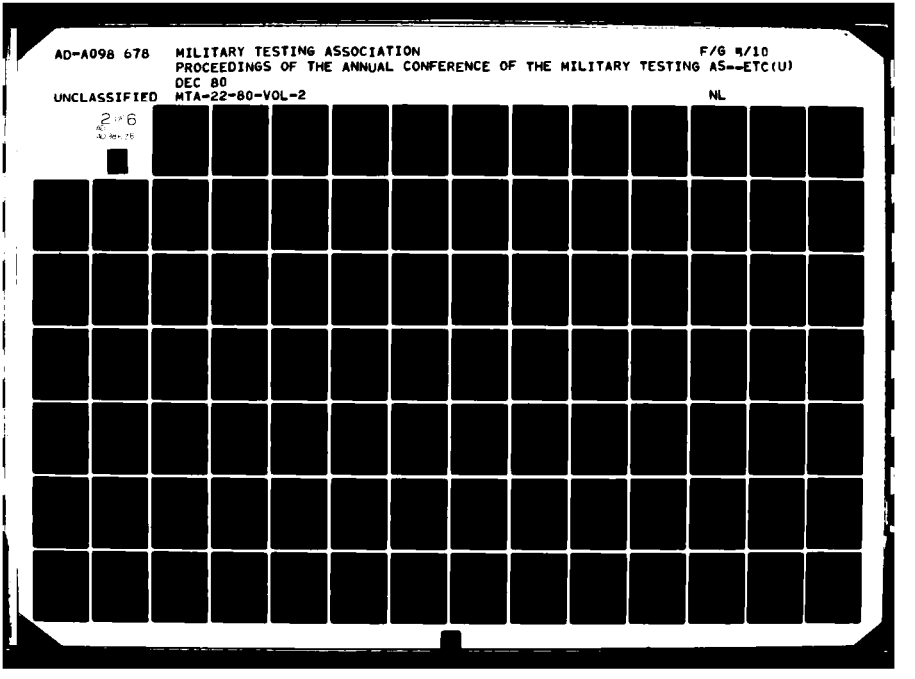
	V-- % OF PEOPLE WITH A THRU E --V						V-% OF ALL-V
TOPIC	VERY GOOD A(1)	GOOD B(2)	SO-SO C(3)	POOR D(4)	VERY POOR E(5)	MEAN OF A THRU E	PEOPLE NO OPINION N(0)
1 ALL TRG	7	39	47	7	0	2.54	0
2 CAI TRG	7	32	47	7	7	2.80	0
3 EDC TRG	13	53	27	7	0	2.27	0

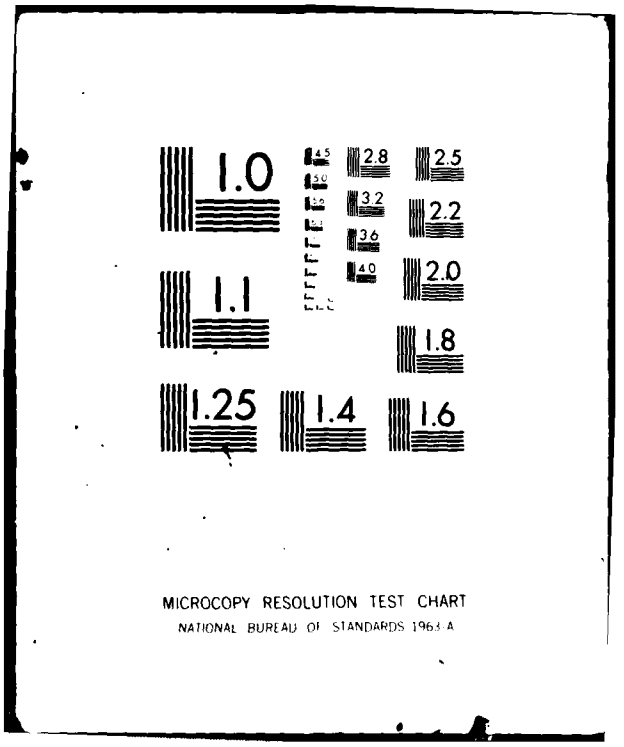
== SPECIFIC ITEMS (%) ==========

	V--- % OF ALL PEOPLE ----V	
ITEM NO.	IMPORTANT (Y + ANY)	UNSURE OF IMPORTANCE OR PREPAREDNESS (ANY ?/BLANK)
01	67	33
02	67	27
03	73	27
04	80	33
05	100	20
06	87	33
07	87	33
08	73	27

	V-------- % OF PEOPLE W/O ?/BLANK --------V			
ITEM NO.	IMPORTANT AND PREPARED (Y + Y)	IMPORTANT BUT NOT PREPARED (Y + N)	NOT IMPORTANT BUT PREPARED (N + Y)	NOT IMPORTANT AND NOT PREPARED (N + N)
01	56	22	11	11
02	36	0	52	8
03	70	20	10	0
04	40	60	0	0
05	92	8	0	0
06	60	40	0	0
07	91	0	9	0
08	30	50	10	10

** FIGURE 3 (CONTINUES) **********


AD-A098 678
MILITARY TESTING ASSOCIATION
PROCEEDINGS OF THE ANNUAL CONFERENCE OF THE MILITARY TESTING AS--ETC(U)
DEC 80
UNCLASSIFIED
MTA-22-80-VOL-2
F/G 5/10
NL


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

** FIGURE 3 (CONTINUED) ************************************

== OPTIONAL ITEMS (%) ==

ITEM NO.	% OF ALL PEOPLE A	B	C	D	E	NO OPINION N
01	0	33	67	0	0	0
02	0	0	47	53	0	0

** FIGURE 3 (END) **

** FIGURE 4 (START) ******** EXAMPLE REPORT ****************

```
POST TRAINING SURVEY - HISTORY PROCESSING REPORT/
SCEQA PTS%F/U LOC = P701 LIMIT 50.
FILE UPDATED ON 02/16
PROGRAM LOC   PTS      ON?  R1 F1 U1  F1%  U1%     R2 F2 U2
NUMBER        DATE
01234  P701 01/23/80   Y   30 14  2   46    7     30 20  0
02327  P701 08/15/80   N   50 33  5   66   10     50 37  2
                           80 47  7   62    9     80 57  2
RECORDS  2
```

** FIGURE 4 (END) **

Comments by the Discussant Bruce A. McFarlane at the Women in the Military Symposium at the Plenary Session of the 22nd Annual Meeting of the Military Testing Association, Toronto, 31 October 1980

We have all been very fortunate this morning to hear these three presentations. Perhaps the authors have not been so fortunate in the choice of me as the discussant for I am a sociologist and they are psychologists, hence we may well tend to emphasize different facets of social life.

The topic of all three papers, women in the military, is a very timely one because as most of us realize women will have to become a major target for recruitment if the all-volunteer basis for our armed forces is to continue hence any knowledge of their experience in the services is invaluable.

The three papers really fall into two categories, viz. (i) the Boyce and Belec paper which is a report of a completed project, albeit part of a larger on-going research programme; and (ii) as presented, the Simpson paper on Canada and the Lipscomb paper in the USA represent what might be called a pre-empirical data state of the research.

But before discussing the papers I would like to make a few general comments which may help to place the papers in a wider context.

The problem with which we are concerned, the sexual division of labour, has very strong underpinnings. I quote from something which I wrote some

time ago (McFarlane in Prewis, 1968: 84) In Canada

> This situation is not a singlular one because wherever human society exists a division of labour based on gender may be found. Ethnographers and anthropologists have documented the wide variety of occupations which may be defined as either masculine or feminine or both, depending upon the particular society under observation. We also know that taboos develop to support these definitions of occupations as being either masculine or feminine, and that these taboos are enforced with some vigour. Professor Hall (1964) in an article has described the situation amusingly in the following manner:
>
> > The taboos that rule a society in regard to man's work and women's work are usually fortified by a set of deeply rooted beliefs about the innate characteristics of men and women. The beliefs may have a derogatory character, as in the notion that women are fundamentally unimportant, or lack the courage for battle, or the stamina for the chase, or that they are too catty for impersonal administration, or too unimaginative for art and architecture, or too personal for working with machinery. Or the beliefs may be of a flattering sort, as that of the Indians who believed (probably accurately) that women were stronger and tougher and therefore better equipped to carry loads over portages, or the belief that women have more solid skulls than have men, and therefore run lesser risks by carrying heavy baskets on their heads.

Within the Judeo-Christian belief system the "proper" roles for men and women have taken many forms but the division of labour on sex lines is so well entrenched than even occupational titles can become sex based so that in order to describe a non-traditional encumbent in an occupational role one has to make a compound noun of the occupational title, e.g., male nurse, female engineer. (A note of caution is in order here, because at any one point in time what may be deemed the "proper" encumbent of the traditional role may not be as long-standing a tradition as one believes; e.g., elemen-

tary school teachers a hundred years ago in Canada were traditionally male - ex-Army officers or clergy). In the present context it would appear that two fundamental aspects of the belief system appear to be that the proper roles for women are nurturant ones (nurses, teachers, social workers)and those for men protective ones (Segal, 1978: 119). Hence, it is not suprising that the acceptance of women in non-traditional roles (e.g., the military) is very slow to come about. As Segal (119) notes

> The role of warrior is obviously one that violates traditional values about the ideal woman....There is perhaps no greater antithesis to this (nurturant role) than a role which involves not only competition, but the outright goal of inflicting physical harm on another. This makes the role of warrior even less acceptable than a woman being a police officer.

Thus it is not suprising that not only is there controversy surrounding the recruitment of women to the military but also the possiblility that they may be called upon to play combat or near combat roles. But perhaps this is not as serious as we male "protectors" may think. At a Geneva Conference in 1976 it was pointed out that in all conflicts since the end of WW II 13 civilians were killed for every combattant who died whereas during WW I 20 combattants were killed for every non-combattant. Hence, those at greatest risk are not necessarily those playing combattant roles.

The foregoing suggests then that it will not be an easy task to bring about a shift in occupational roles and the attendant societal attitudes towards the 'proper' work for women and the 'proper' work for men in the armed forces.

Now to return to the papers. Let us look first at Simpson's paper wherein she has put forward a model for understanding the performance appraisal of

women in the Canadian Forces as well as the role of superordinate and subordinate expectations on relationships and actual behaviour. Of particular interest is Simpson's formulation (following Jones) of the role of labelling and self-fulfilling prophecies, and as I view this area, the role of stereotypes.

I would like to suggest that any group of 'newcomers' would be in the same position as Simpson's women; that is, her problem may not be sex-specific. For example, I suspect that as Blacks moved into an integrated military in the USA, or Francophones were recruited for more positions in Canadian Industry, or North American Indians and Inuit moved into newly developing industry in Canada's northland the traditional definitions of the 'proper' role for these groups would have operated in much the same fashion. Any who seemed to "work out better than expected" would be viewed as exceptions that proved the rule that most were not very good in roles deemed to be non-traditional for them. In most instances group definitions would be applied to them rather than individual definitions; as I have illustrated elsewhere (McFarlane 1968) when northern Indians were hired and left their job after a short stay, they did so according to management because "they were Indians", when whites hired in the south of Canada for work in the north followed the same work pattern individual excuses were found for their reasons for quitting (McFarlane, 1972). Hence it might be very useful and instructive if Simpson could locate another group, perhaps the members of a non-Francophone non-Anglophone ethnic group in the forces to aid in her analysis.

I would also like to add a word of caution when the material from the two sets of questionnaires is analysed. The researcher should be very careful

when using a questionnaire designed to elicit information from one group about another group, then simply shifting the wording so that it is believed to be set from the second party's perspective. Each group may have quite different definitions of the situation, hence, be really answering two different questions although the wording is the same.

Simpson plans to look primarily at those at the top third of her scale (the more egalitarian group) and the bottom third (the traditional group), that is, at the extremes, I would recommend that since she believes that there is a practical application for the expected findings, that she examine the findings of Cotton (1980, 308-344) wherein his "Ambivalents" or middle group (neither the traditional military types nor those with an occupational orientation - the military as a 9 to 5 job) seemed to be the liaison group between the two extremes and may, in fact, be the group who held the whole thing together.

Let us now turn to Lipscomb's paper and her comments on the remarkable shift which has taken place in women's work in the arm of the service which she is examining: "...the Air Force currently has more women working in aircraft maintenance than in personnel specialties". This is, indeed, remarkable. Of singular interest to us too is her interest in the findings of Bergmann and Christal that task assignment within an occupational category appears to be a function of sex. That is, in the same specialty (aircraft maintenance) a higher percentage of males are found doing the actual mainenance task while a higher percentage of females are doing support tasks (administrative and clerical). The main question to ask, I presume, is, what is it that causes this shift? Is it related to the actual mechanical aptitude

of the males and females? To their "liking" or "disliking" aircraft maintenance work per se? Is it "something" in the system which causes them to shift.

Before attempting to answer these questions a few general comments are, I believe, in order.

Most standard mechanical aptitude tests, it seems to me, are really inappropriate for females in our society (for example, those showing pictures of specialized wrenches, or using words such as monkey wrench, spanner, etc., yet not showing sewing machine parts, bobbins, spindles for weaving sets and looms, or "kitchen" machine words). Under these conditions it is not suprising that women do not show as much mechanical aptitudes as do males, especially when added to this is the whole socialization process in North American society where from birth males are directed towards things mechanical and girls steered away from them. Even when, as is the case today, young boys begin to play with dolls we have a remarkable way of changing the definition of the situation so that the "dolls" become masculine prototypes, e.g., G.I. Joe, Action Man, etc., and the functional equivalent of the dolls house becomes a "space station". Added to this, of course, is a much more fundamental question, how well do women do in mechanical training situations? What proportion of their training actually requires that they have considerable knowledge of tools etc., before they begin their training? What role does stereotyping play in "keeping females away from the machinery"?

A further comment may be made about the aircraft maintenance personnel's educational experience before enlisting. Studies in Canada (Hall and McFarlane, 1962; Breton, 1972) have shown that girls do better in class-room situations than boys - even among drop-outs most girls who do so drop out at the end of a

successful school year, whereas the boys drop out mid-year when failing or only marginally coping with the work (Hall and McFarlane, 1962).

Given these two conditions (stereotypes concerning female mechanical aptitude and a certain academic competence) it is not suprising that after they have received their trades training in the service their supervisors soon call upon the women to carry out those necessary important tasks requiring what are essentially white collar skills - hence, shift them from maintenance to support tasks, record keeping, etc. It was also true that until a short time ago in the Air Force (USA) the educational requirements for entry for females was higher than for males. If the above explanation holds then it is not suprising that the shifts occurred as they did; a responsible supervisor (sergeant?) had little choice.

At this point I would like to pose some questions and make some further comments which may help Lipscomb after her data arrive. Who are the women who shift? Who are the men? Are they more like the women who shift or the men who do not shift in terms of mechanical aptitude and/or educational background? I would suspect too that the men who make the shift or are shifted are more likely to come to the attention of those who will be able to help make their service career a successful one, and it is more likely to be white collar skills than manual skills which play a role in this process.

I would also like to caution Lipscomb when treating the job satisfaction data. In many instances respondents do not differentiate between <u>job</u> and <u>work</u>, particularly on questionnaires, so that the following could be a possible answer which one might receive in an interview situation: "I like my work but I hate my job"! Similarly caution is needed when using Supervisory rating data

when assessing levels of job competence, especially if Simpson's findings support her hypotheses. That is, the orientation of the supervisors must be taken into account when assessing their evaluations.

Now for a few brief comments on the Boyce and Belec study. Unlike the other two studies these researchers had already gathered a vast amount of data for use in their analysis. It is a very full study and has the added advantage of having used considerable comparative data.

The authors note that the scores of the cadets at the Canadian Military Colleges and the men in the Canadian Forces "reflect significantly more egalitarian attitudes toward women's roles in society than those expressed by male cadets at the US Military Academies" and in their caution they note that there was a difference of four years in the administration of the Attitudes *Towards Women Scale (AWS) in the two countries* hence, societal changes may have taken place. I would like to put it to these researchers that since attitudes of this type are culturally determined there is every reason to accept the fact that there may well be attitudinal differences between young people in Canada and the USA, despite our seeming similarity. (For example, a higher proportion of dentists, physicians and lawyers are women in Canada than in the United States). In addition they also found differences between Anglophone and Francophone servicewomen when the AWS was administered to these two groups. In this instance it could be that a test standardized for an Anglophone population has cultural biases within it when applied to another population, the Francophones. This may help to explain their seeming anomaly that despite the more "traditional" attitudes of the Francophones as measured by the test "the servicewomen who expressed traditional attitudes toward women's

roles were willing to assure non-traditional employment within the Canadian military as those servicewomen who expressed egalitarian/profeminist attitudes". I pose this as an alternative explanation to the hypothesis derived from Cotton's study which they put forward, wherein the difference in willingness to serve in non-traditional occupations and/or settings is related to the women's orientation to the military, that is, as either "vocational" or "occupational".

One aspect of this study which I found particularly fascinating and suprising were the findings which indicated that there were "no significant differences between CF servicemen and civilian male students" in AWS test scores and that Prociuk found "no significant AWS mean score differences between CMC cadets and civilian college male students". It seems to me that these results are so striking that it calls the AWS test into question as far as its usefulness on a Canadian population is concerned. It is rather hard to accept such homogenity of attitudes among these disparate groups. Could it be that it does not measure for Canadians what it purports to measure?

In this brief overview of each of the three papers I have tried to cast the research in a broader perspective and in doing so I hope that I may have assisted the researchers with their tasks of interpretation.

Breton, Raymond, Social and Academic Factors in the Career Decisions of Canadian Youth, Ottawa: Information Canada, 1972.

Cotton, C.A., The Divided Army: Role Orientations Among Canada's Peacetime Soldiers, Ph.D. Thesis, Carleton University, Ottawa, 1980.

Hall, Oswald, "Gender and the Division of Labour" in Implications of Traditional Divisions Between Men's Work and Women's Work in Our Society, Report of a Round Table Conference, Women's Bureau, Ottawa: The Queen's Printer, Department of Labour, 1964.

Hall, Oswald and Bruce A. McFarlane, Transition From School to Work, Ottawa: The Queen's Printer, 1962.

McFarlane, Bruce A., "Education and Manpower: Some Sociological Aspects of Growth", in T.N. Brewis (ed.) Growth and the Canadian Economy, Toronto: McClelland and Stewart Ltd., 1968.

McFarlane, Bruce A., "Manpower Problems in the Canadian Mining Industry: One Possible Solution:, Canadian Mining and Metallurgical Journal, 1972, vol. 65, no. 725.

Segal, Mady Wechsler, "Women in the Military": Research and Policy Issues", Youth and Society, vol. 10, no. 2, December 1978.

MCKENZIE, Robert C., United States Office of Personnel Management, Personnel Research Development Center, Research Section, Washington, D.C.

A NEW PROCEDURE FOR ESTIMATING THE STANDARD DEVIATION OF JOB PERFORMANCE (Tue A.M)

Failure to find economically feasible methods of estimating the standard deviation of job performance (SDy) has been the main reason why the Brogden (1949) Cronbach and Gleser (1965) utility models have not been used. This report presents a new procedure for estimating the SDy. This method was first introduced by Frank L. Schmidt et al. (1977) of the U.S. Office of Personnel management. In this study the method was applied to two occupations within the Federal government (budget analyst and computer programmer). In the first occupation, supervisors of budget analysts were asked to estimate the dollar value of the superior (at the 85th percentile) and average (at the 50th percentile) performing employees. In the second study, supervisors of computer programmers were asked to give estimates of the low performing employees (at the 15th percentile) as well as the superior and average performing employees. Estimates of the low performer were included in the second study primarily to test for the assumption of normality. The difference between the supervisor's estimates of superior and average performer and average and low performer was defined as the standard deviation of job performance (SDy_2 and SDy_1). In both studies, supervisors were asked to estimate the percentage of employees who were making a positive contribution to the agency and those who were breaking even. Results of the second study showed a nonsignificant difference between the two SDy's, thus supporting the normality assumption. It was also found that experienced supervisors of programmers estimated higher standard deviations of job performance than inexperienced supervisors. Those supervisors who supervised a large number of employees also made higher estimates of SDy; however, none of these findings were supported in the first study with budget analysts. When comparing the two studies, supervisors of both budget analyst and computer programmers estimated the percentge of employees contributing positively to the organization to be about the same (77.8% vs 77.6%)

A NEW PROCEDURE FOR ESTIMATING THE STANDARD DEVIATION OF JOB PERFORMANCE

Perhaps one of the most important problems facing personnel psychologists today is the development of a meaningful criterion by which to measure employee performance. Traditionally, the approach has been to assess employee performance in terms of supervisory ratings, peer ratings, and self-ratings. The validity of a selection device is then described by the correlation between these kinds of ratings and the test. This relationship is assessed by such statistics as the validity coefficient, the index of forecasting efficiency, and the coefficient of determination. However, none of these statistics, except for the validity coefficient, are very useful when used to show the value of a test in increasing productivity.

The index of forecasting efficiency, $E = 100(1-\sqrt{1-r^2})$, was heavily emphasized in early texts (Kelley, 1923; Hull, 1928) as the appropriate means for evaluating the value of a selection procedure. The index of forecasting efficiency is the percentage reduction in errors of prediction that would result from predicting that performance of all individuals will be at the mean of the criterion measure. The forecasting efficiency of a test or battery is therefore inversely proportional to the amount of error resulting when the test is used to forecast a criterion (Hull, 1928). Cronbach and Gleser (1965) have pointed out that the index of forecasting efficiency evaluates a test by absolute error of estimate, relative to error of a chance estimate. This type of evaluation is appropriate only in situations where payoff declines linearly with error.

Perhaps the most popular interpretation of validity has been the coefficient of determination (r_{xy}^2). The coefficient of determination is found by squaring the correlation coefficient, for example; if r_{xy} is .60, squaring will result in a correlation coefficient of .36. This squared correlation is defined as the proportion of variance in the criterion accounted for by variance in the test. Schmidt et al. (1979) have pointed out that neither of these interpretations has recognized that the value of a test varies as a function of the parameters of the situation in which it is used. Both E and r_{xy}^2 are general interpretations of the validity coefficient, and lead to the erroneous conclusion that only tests with high correlations have practical significance.

Brogden and Taylor (1950) have proposed what is known as a dollar criterion to assess the value of employee performance. The practical application of this approach is most popularly demonstrated by Cronbach and Gleser's utility model. Brogden (1949) Cronbach and Gleser (1965) have pointed out that the utility of a test depends not only on the validity coefficient, but also on the cost of testing, selection ratio, number of selectees, the cutting score, and the standard deviation of job performance (in utility units). A modified version by Schmidt et al. (1979) of the Brogden-Cronbach-Gleser equation for total gain in utility can be expressed as:

$$\Delta U = n_s r_{xy} SD_y \bar{z}_{xs} - n_s C/r$$

Where:

ΔU = total gain in utility

n_s = number of selectees

The opinions expressed in this paper are those of the author and do not reflect official policy of the Office of Personnel Management.

C = cost of testing one applicant

P = selection ratio

$\bar{z}_{xs}$ = average standard score on the test of those selected (in applicant group standard score units)

r_{xy} = test validity

SD_y = the standard deviation of job performance in dollar terms.

Note that the first part of the equation ($N_s r_{xy} SD_y \bar{z}_{xs}$) is the Brogden formula for overall gain in utility; the last part ($N_s C/P$) is derived from Cronbach and Gleser (1965). A simple derivation of this formula can be found in Schmidt, Hunter, McKenzie, and Muldrow (1979).

Since the standard deviation, selection ratio, cost of testing, and number of selectees are constant for any one treatment, utility is a linear function of validity, and this relationship will hold at all selection ratios. The first five terms of the Brogden-Cronbach-Gleser formula, as listed above, have not been difficult to estimate in the past. The sixth, the standard deviation of job performance (SD_y) has been virtually impossible to estimate.

Estimating SD_y

Cronbach and Gleser (1965) described the standard deviation of job performance (SD_y) as the magnitude and practical significance of individual differences in payoffs among randomly selected applicants. The evaluated outcome of an individual's performance is referred to as the payoff, and is determined for each individual in a particular information category (e.g., test score) by weighting the value of each outcome by its probability for individuals, and then summing across all outcomes. Expected payoffs have little meaning unless they can be averaged over a large number of individuals in the same information category. Since gains in utility depend on the product of the standard deviation with other terms in the equation, a large standard deviation indicates a large practical significance of individual differences in payoff. A large SD_y also indicates that predictors which have low validities may still yield considerable benefit.

Traditionally, the procedure advocated for estimating the standard deviation has been cost accounting. However, this procedure of estimating SD_y involves the laborious and tedious process of costing out every detailed expense related to the job, such as rent, direct and indirect labor cost, and tool usage. In an earlier review, Hunter and Schmidt (1979) were able to locate only two studies which have attempted to estimate SD_y using cost accounting methods.

Roche (in Cronbach and Gleser, 1965) illustrated how detailed and complex this procedure can be. He conducted a study in a large midwestern plant of a heavy equipment manufacturer. Roche used a cost accounting procedure known as "standard costing" to estimate the standard deviation for radial drill operators. The standard cost method is an effective tool for assessing volume production. Standard cost is determined by procuring cost data on three basic factors, material used to produce products, direct labor hours used to alter the size, shape, quality, or consistency of material, and facility use required to perform direct labor. In order to use this method, all standards usually remain frozen for a period of five years, or until a general cost revision is authorized. Individual performance is assessed by a productivity measure called performance ratio. For example, for each machine operation that is performed, there is a time standard (the time it would take the average worker to perform the operation) which is compared to the time of actual production. The performance ratio is determined by dividing a person's actual production per hour by the standard hourly production rate for whatever piece of equipment he or she is working on. For example, if the standard

hourly production rate is twelve pieces per hour, and an operator produces seven pieces in an hour, his or her performance ratio for that hour is .58. Roche calculated cost estimates for each piece of machinery, direct and indirect labor cost, overhead, tool usage, rent, etc. He also made adjustments for performance below standard.

Cronbach and Gleser (1965) in commenting on Roche's study have pointed out that accountants perhaps did not understand the utility estimation problem, so that they may have underestimated or overestimated the expenses. Also, they calculated contribution to profit rather than to productivity. Only a small percentage of a worker's production goes to profit, and the rest goes to expenses of various kinds. Even when a complex and detailed procedure such as cost accounting is employed, one is still not certain as to whether estimates are accurate. And if this approach is used at higher occupational levels, uncertainties about accuracy are even greater.

Schmidt et al. (1977) have recently developed a method of rationally estimating the standard deviation of job performance (SD_y). This method is based on the following reasoning: If job performance in dollar terms is normally distributed, then both the difference between the value to the organization of products and services produced by the average performer (at the 50th percentile) and the superior performer (at the 85th percentile); and the difference between the average performer and the low performer (at the 15th percentile) are equal to SD_y.

This paper describes two studies which attempt to demonstrate the utility and feasibility of the rational estimation method over cost accounting methods. The major purpose of the first study was to demonstrate that rational estimation methods can probably be just as accurate as cost accounting methods of estimating SD_y. The second study is an extension of the first: one of its purposes was to test the assumption that the dollar value of employee productivity is normally distributed.

The following hypotheses were tested in the studies:

Hypothesis 1: Supervisors feel confident about the accuracy of their estimates of SD_y.

Hypothesis 2: Supervisors who supervise larger numbers of employees make larger estimates of SD_y.

Hypothesis 3: Experienced supervisors make larger estimates of SD_y than inexperienced supervisors.

Hypothesis 4: Job performance is normally distributed, so that the difference between the performance of workers at the 85th percentile, and workers at the 50th percentile (defined as SD_{y1}) will not be significantly different from the difference between workers at the 50th percentile and workers at the 15th percentile (defined as SD_{y2}). The fourth hypothesis was tested in Study II only.

Study I

Sample

A list of 135 supervisors of budget analysts was compiled from various Federal agencies in Washington, D.C. A total of 394 incumbent budget analysts were listed as working under the 135 supervisors selected for the study. However, study materials were delivered to only 93 out of 135 supervisors because the rest had quit, transferred, or could not be reached. Shortly after the 93 were reached, six others quit, leaving a total of 87. These remaining supervisors had a total of 248 budget analysts working for them. Sixty of them returned questionnaires, for a return rate of 69%.

Procedure

The study focused on the budget analyst position at the GS 9-11 levels. The questionnaires that were mailed to supervisors were made up of three parts (see Appendix A). The first part consisted of two questions asking supervisors to estimate the percentage of budget analysts falling in the "break-even" and "positive-benefit" categories. The break-even category was defined as those employees who produced about the same in goods and services as it cost to keep them on the payroll. The assumption made was that employees rarely or never fall exactly at the break-even point. The break-even category was therefore deemed to be made up of 50% slightly positive-benefit and 50% slightly negative-employees.

The main reason for including the positive-benefit category was to determine the percentage of employees that supervisors felt were making a positive contribution to the agency. Thus, the final positive-benefit category was determined by adding half of the break-even category to the positive-benefit category. The remaining percentage plus half of the break-even category made up the final negative category. This category was included to find out the percentage of employees that supervisors felt were of negative value to the agency, thus, the negative category was not estimated directly.

The second part of the questionnaire was concerned with the yearly value of goods and services provided by the employee to the agency. Supervisors were asked to estimate this value in dollars for the average budget analyst (at the 50th percentile) and for the superior budget analyst (at the 85th percentile). These percentile points were arbitrarily selected. Three multiple-choice questions were also included to test three of the four hypotheses stated earlier in the text. The first question, on the perceived accuracy of estimation, was included to find out how much confidence supervisors felt in making their estimations. The second, the number of budget analysts supervised, was included to find out whether or not individual differences were more apparent to those who supervised more workers. The last question was concerned with the role experience plays in estimating SD_y. In other words, do the more experienced supervisors make larger estimates of SD_y than the less experienced?

Study II

Procedure

One hundred and forty-seven questionnaires were distributed to computer programmer supervisors. One hundred and five were returned, for a return rate of 71.4%.

Estimates of the standard deviation of productivity (SD_y) were provided by experienced supervisors of computer programmers in 10 Federal agencies. The procedure used in the budget analyst study was modified to include the poor performer (15th percentile) as well as the average (50th percentile), and superior performer (85th percentile), in order to allow for the testing of the normality assumption. It was hypothesized that if the dollar value of performance is normally distributed, the standard deviation of job performance, computed as the difference between the superior and average and computed as the difference between average and low, would not differ significantly in value. That is, the two estimates of SD_y should be similar in value.

Study I

Results

Table 1 presents the means and standard deviations for variables in Study I.

As shown in Table 1, a mean of 77.8% was found for the final

TABLE 1

Means, Standard Deviations and Standard Errors of Supervisors' Estimates, Study I (Budget Analyst) (N = 60)

	$\bar{x}$	SD	$Se_{\bar{x}}$
Percentage of Incumbents Showing Final Positive Benefit	77.8%	15.53	2.00
Estimated Productivity of Superior Budget Analyst	$34,848	$17,605	$2,272
Estimated Productivity of Average Budget Analyst	$23,378	$11,428	$1,475
SD_y (Superior less Average)	$11,466	$8,684	$1,121
Accuracy of Estimates (Response number)	3.80[a]	1.72	.22
Number of Employees Supervised (Response number)	2.25[a]	1.04	.29
Years of Experience as a Supervisor (Response number)	3.30[a]	.93	.42

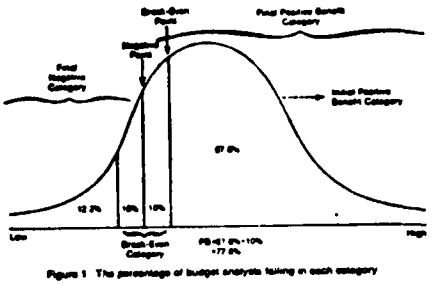
[a]Because these responses indicated ranges (e.g., "$2,000 or less, either way," "Three to five employees, three to four years"), statistics were calculated on the number of multiple-choice response alternative.

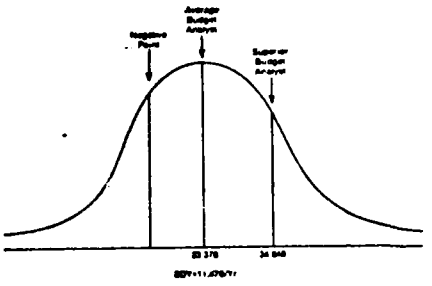
positive-benefit category. This indicates that, on the average, 77.8% of the budget analysts were estimated to be of positive benefit to the the agency. Since the assumption was made that employees rarely or never fall exactly at the break-even point, the final positive-benefit category was determined by adding 50% of the break-even to the positive-benefit category. The final negative category was determined by adding 50% of the break-even category to the negative category. This relationship is illustrated in Figures 1 and 2.

Figure 1 illustrates a normal distribution of the estimated percentages of budget analysts falling in each category. For example, 67.8% of the budget analysts were estimated to be in the initial positive-benefit category, and this percentage was added to half of the percentage in the break-even category, which resulted in a final estimate (final positive-benefit category) of 77.8%. The remaining 12.2% percent plus half of the break-even category resulted in an estimated 22.2% for the final negative category. Figure 2 illustrates the estimated mean dollar value for the superior ($34,848) and average ($23,378) budget analysts. The difference between these two values ($11,470) is the yearly [illegible] for this occupation.

The mean and standard deviation for perceived accuracy of estimation (Table 1) were 3.80 and 1.72, respectively. The results indicated that the average supervisor felt confident that his or her estimates were within 3,000 dollars or less in either direction.

For the number of budget analysts supervised, the mean was 2.25 and the standard deviation was 1.04. This finding indicated that the average number of budget analysts supervised was six to ten. Supervisory





experience was estimated to be three to five years.

Table 2 shows the relationship between SD_y and these variables. A positive correlation was found between the estimates for superior and average budget analysts (r = .91, p < .001), which indicates that both were estimated consistently. The estimates of SD_y and number of years experience correlated negatively (r = -.26, p < .05); the more experienced supervisors estimated a smaller SD_y than the less experienced supervisors, contrary to what was expected. A higher correlation was found between supervisors' estimates of the value of the superior performer and SD_y (r = .83, p < .001) than the correlations between their estimates of the value of the average performer and SD_y (r = .53, p < .001). These correlations are part-whole correlations, which indicate that the range of estimates for the superior performer was greater than the range of estimates for the average performer.

Study II

Results

Table 3 presents the means, standard deviations, and standard errors of the mean of the responses to the questionnaire addressed to supervisors of computer programmers.

The mean estimated worth of superior, average, and low-performing programmers was $36,582, $25,896 and $16,045, respectively. The mean estimated difference between the superior and average programmers (SD_{y1}) was $10,686, and the mean estimated difference between the average and low programmers (SD_{y2}) was $9,851. This difference of 834 dollars between SD_{y1} and SD_{y2} was not found to be statistically significant (p = .74). The standard error, however showed greater agreement among supervisors in estimating the difference between the average and low-performing programmers (SD_{y2}) than the error in estimating the difference between the superior and average programmers (SD_{y1}).

Figure 3 shows the estimated percentages of computer programmers falling in each category. The initial positive-benefit category included 62.5% of the computer programmers; this percentage was added to half of the percentage in the break-even category (15.1%), resulting in a final estimate of 77.6%. The remaining 7.3% plus 15.1% (half of the break-even category), made up the final negative category (22.4%). Figure 4 illustrates a normal distribution of the estimated dollar value of the superior, average, and low-performing programmers.

The correlations in Table 4 show the relationship of these variables with SD_{y1} and SD_{y2}.

Significant correlations were found between the estimated value of the superior and average (r = .92, p < .001), and superior and low (r = .70, p < .001) programmers, again indicating that supervisors estimated the worth

(640)

TABLE 2

Intercorrelations of Supervisor' Estimates, Study I, Budget Analyst (N = 60)

	(1)	(2)	(3)	(4)	(5)	(6)
(1) Superior Analyst	-	.91***	.83***	.13	.00	.15
(2) Average Analyst		-	.53***	.21	-.06	-.03
(3) SD_y			-	.00	.08	-.26*
(4) Accuracy				-	-.04	.01
(5) Number Supervised					-	.12
(6) Experience						-

*$p < .05$
**$p < .01$
***$p < .001$

TABLE 3

Means, Standard Deviations, and Standard Errors of Supervisors' Estimates, Study II, Computer Programmer

	$\bar{X}$	N	SD	$Se_{\bar{X}}$
Percentage of Incumbents Showing Final Positive Benefit	77.6%	105	26.4%	2.58%
Estimated Productivity, Superior Programmer	$36,582	105	$33,840	$3,302
Estimated Productivity, Average Programmer	$25,896	105	$20,585	$2,008
Estimated Productivity, Low Programmer	$16,045	104	$14,487	$1,420
SD_{y_1} (Superior less Average)	$10,686	105	$17,140	$1,672
SD_{y_2} (Average less Low)	$9,851	104	$10,243	$999
Accuracy	3.86[a]	105	2.0	.20
Number Supervised	2.62[a]	100	1.1	.11
Experience	3.95[a]	101	1.3	.13

[a]Because these responses indicated ranges (e.g., "$2000 or less either way," "Three to five employees," "Three to four years"), statistics were calculated on the number of multiple-choice response alternative.

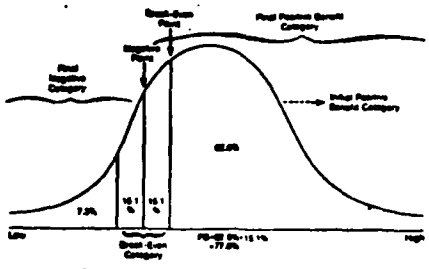


Figure 5. The percentage of computer programmers falling in each category

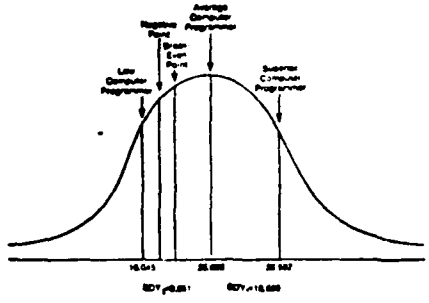


Figure 6. A normal distribution of dollar values for each computer programmer

of the superior, average, and low-performing programmer in a consistent manner. That is, supervisors who estimated high dollar values for the performance of superior employees also estimated high dollar values for the performance of average employees and for that of low-level performers.

The correlations between the supervisor's estimates of the superior, average, and low-performing programmers and accuracy (r = .56, p < .001; r = .46, p < .001, and r = .29, p < .01, respectively), indicate that the larger the estimates made, the less confident supervisors felt in estimating the value of their subordinates' performance. The same relationship held for estimates of SD_{y_1} (r = .54, p < .001) and SD_{y_2} (r = .52, p<.001). On the other hand, the mean in Table 3 indicates that supervisors were confident that these estimates, on the average, were accurate within 3,000 dollars.

The positive correlations between SD_{y_1} and SD_{y_2} with the number of programmers supervised (r = .25, p < .01 and r = .30, p < .01, respectively), indicate that supervisors who supervised more employees estimated a larger SD_y, contrary to the results in Study I, where r = .08. The more experienced supervisors were also found to supervise more programmers (r = .32, p < .001).

Discussion

The estimated percentages for the final positive-benefit category were almost identical in both studies (77.6% for the programmers and 77.8% for the budget analysts). It appears that supervisors generally felt that the majority of the employees they supervised were of positive benefit to the agency. That is, the amount of goods and services they produced exceeded the cost of keeping them on payroll. However, if goods and services did not exceed the employee's salary, the organization broke even, or the employee was of negative value to the organization. The break-even point was hypothetical--it was assumed that people rarely or never fall at this point. Thus, the break-even category, as explained earlier, consisted of 50% positive and 50% negative. The final positive and negative categories were formed by adding 50% of the break-even to each.

When half of the break-even category was added to the negative category, over 22% of the employees were considered to be of negative value to the agency, a relatively large percentage. However, it should be pointed out that only 7.3% of budget analysts and 12.2% of computer programmers fell in the pure negative category. The remaining percentages (10% and 15.1%) were part of the hypothetical break-even category.

The value of superior budget analysts was estimated to be $34,843, and the value of average budget

(642)

TABLE 4

Intercorrelations of Supervisors' Estimates of Productivity
Study II: Computer Programmers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) Estimate for Superior Performer	—	.92**	.70**	.88**	.84**	.56**	.28** (N = 100)	.07 (N = 101)
(2) Estimate for Average Performer		--	.89**	.61**	.76**	.46**	.26** (N = 100)	.07 (N = 101)
(3) Estimate for Low Programmer			--	.33**	.37**	.29*	.15 (N = 100)	.03 (N = 101)
(4) SD_{y_1}				—	.76**	.54**	.25* (N = 100)	.06 (N = 101)
(5) SD_{y_2}					--	.52**	.30* (N = 100)	.10 (N = 101)
(6) Estimate of Accuracy of Estimates						--	.12 (N = 100)	.00 (N = 101)
(7) Number Supervised							--	.32** (N = 100)
(8) Experience								--

Note. N = 105 except as indicated. Some supervisors did not respond to some questions.
*p < .01
**p < .001

analysts was $23,378. SD_y was computed by taking the difference between these two values, which is $11,470. Table 1 shows a smaller standard error for the mean estimate of the value of average budget analysts than for superior budget analysts. That is, there was less disagreement in estimating the contribution of the average budget analysts than in estimating the contribution of the superior budget analysts. A correlation of .91 was found between the estimate for superior and average budget analysts, which indicates that those supervisors who assigned a relatively high value to the average budget analyst also assigned a high value to the superior budget analyst. This correlation was of essentially the same magnitude in the computer programmer study (r = .92, p < .01). The effect of this high correlation indicated the similarity of estimates of SDy across supervisors even though there was considerable disagreement between supervisors on the value of the average and superior budget analysts (standard error of the mean = $2,272 vs. $1,475).

The estimated worth of the superior ($36,582) and average ($25,896) programmers was essentially the same as the values estimated for the superior ($34,843) and average ($23,378) budget analysts. However, the programmer study included an estimate of the low performer which was estimated to be $16,045. The inclusion of the low performer allowed for the testing of the normality assumption. It was hypothesized that if the distribution were normal, each estimate of SD_y would allow for the prediction of the value of the other. SD_{y_1} (the difference between the 85th and 50th percentile) was $10,686, and SD_{y_2} (the difference between the 50th and 15th percentile) was

$9,851. This difference of $834 between SD_{y_1} and SD_{y_2} was not found to be statistically significant. There was, however, greater disagreement on SD_{y_1} than SD_{y_2}.

Cascio and Sibley (1979) found standard deviations (SD_y) similar to those reported here in their study of the utility of assessment centers. The method they used to estimate SD_y was the same as the one used in the budget analyst and computer programmer studies and was suggested to them by Schmidt (Note 2). Third-level managers were asked to estimate the value of job performance for second-level managers at the 85th (superior) and 50th (average) percentile; SD_y was then computed as the difference between these two values. The results showed an estimated SD_y of $9,500 for the first year, similar to those found in the two studies being reported here.

A comparison of Tables 1 and 3 shows less variability (SD) in the estimates of superior, average, and SD_y for budget analysts than for programmers. It appears that budget analyst supervisors were better able to make these estimates, perhaps because their job duties consisted of the kinds of tasks that involve numerical estimation.

It was predicted that supervisors would feel confident about the accuracy of their estimates. Supervisors of both budget analysts and programmers felt confident that their estimates were within $3,000 of the true values. However, for supervisors of programmers, the larger the estimates they made, the less confident they felt.

It was also predicted that supervisors who supervised a larger number of employees would estimate a larger SD_y. This hypothesis was supported for supervisors of programmers but not for supervisors of budget analysts.

Experienced supervisors were also expected to estimate a larger SD_y than inexperienced supervisors. Support for this hypothesis was not found for supervisors' estimates of programmers. For budget analysts, experienced supervisors estimated a smaller SD_y than inexperienced supervisors, contrary to what was expected.

Conclusion

From the results of the computer programmer study, it appears that the assumption of normality is valid when estimating standard deviations of job performance (SD_y), at least in the case of the computer programmers. It also seems that supervisors of larger numbers of subordinates make larger estimates about the relative dollar value of subordinates than supervisors of smaller numbers of subordinates. There is also greater disagreement among supervisors when estimates are large.

The rational estimation method when compared to cost accounting methods is both time saving and cost saving. Brogden (1950) described at least six factors that can be assessed by cost accounting methods, for example; average value of production, quality of objects produced or services accomplished, overhead, errors, the cost of time consumed by other personnel, and social factors. To estimate just these factors alone would involve an extremely detailed and complex procedure. However, even if cost accounting methods are used, rational estimates will still have to be made as one moves up the occupational ladder. Also the fact that error is present in both the rational estimation and cost accounting methods in estimating SD_y suggests the need for liaison between cost accountants and personnel psychologists.

REFERENCE NOTES

1. Schmidt, F. L., Caplan, J. R., Bemis, S. E., Decuir, R., Dunn, L., & Antone, L. *The behavioral consistency method of unassembled examining* (TM-79-21),

Washington, D.C., U.S. Office of Personnel Management, Personnel Research and Development Center, November 1979.

2. Personal communication, January 1980.

REFERENCES

Brogden, H. E. On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology*, 1946, *37*, 65-76.

Brogden, H. E. When testing pays off. *Personnel Psychology*, 1949, *2*, 171-183.

Brogden, H. E., & Taylor, E. K. The dollar criterion: Applying the cost accounting concept to criterion construction. *Personnel Psychology*, 1950, *3*, 133-154.

Cascio, W. F., & Sibley, V. Utility of the assessment center as a selection device. *Journal of Applied Psychology*, 1979, *64*, 107-118.

Cronbach, L. J., & Gleser, G. C. *Psychological tests and personnel decisions*. Urbana, Illinois: University of Illinois Press, 1965.

Hull, C. L. *Aptitude testing*. Yonkers, NY: World Books, 1928.

Hunter, J. E., & Schmidt, F. L. Fitting people to jobs: The impact of personnel selection on national productivity. In *Human performance and productivity*, E. A. Fleishman, (ed.) 1980, in press.

Kelley, T. L. *Statistical method*. New York: MacMillan, 1923.

Roche, U. F. The Cronbach-Gleser utility function in fixed treatment employee selection. Unpublished doctoral dissertation, Southern Illinois University, 1961. Dissertation Abstracts International, 1961-62, *22*, 4413. (University Microfilms. No. 62-1570). Portions of Ph.D. dissertation reproduced in L. J. Cronbach, & G. C. Gleser, *Psychological tests and personnel decisions*, Urbana, Illinois, University Press, 1965, 254-266).

Schmidt, F. L., & Hunter, J. E. Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 1977, *62*, 529-540.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. The impact of a valid selection procedure on workforce productivity. *Journal of Applied Psychology*, 1979, *64*, 609-626.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. Statistical power in criterion-related validation studies. *Journal of Applied Psychology*, 1976, *61*, 473-485.

Taylor, H. C., & Russell, J. T. The relationship of validity coefficients to the practical effectiveness of tests in selection. *Journal of Applied Psychology*, 1939, *23*, 565-578.

APPENDIX A

The Questionnaire Sent to Budget Analyst Supervisors
Estimation of Selection Utility

Budget Analyst - (GS - 560)

Name _____ Dept._____ Agency _____

INSTRUCTIONS

The dollar utility estimates you are asked to make in this phase of the study are critical in estimating the relative dollar value to the government of different selection methods. In answering these questions, you will have to make some very [illegible] *ments*. We realize they are difficult and that they are judgments or estimates. You will have to ponder for some time before giving each estimate, and there is probably no way you can be absolutely certain your estimate is accurate when you do reach a decision. But keep in mind three things:

(1) The alternative to estimates of this kind is application of cost accounting procedures to the evaluation of job performance. Such applications are usually prohibitively expensive and in the end, produce, like our procedure, only imperfect estimates.

(2) Your estimates will be averaged in with those of other supervisors of Budget Analysts. Thus errors produced by too high and too low estimates will tend to be averaged out, providing more accurate final estimates.

(3) The decisions that must be made about selection methods do not require that all estimates be accurate down to the last dollar. Substantially accurate estimates will lead to the same decisions as perfectly accurate estimates.

The information you provide in this phase of the study, like all information provided in connection with this selection study, will not be available in identifiable form to anyone in your agency. It will go to the Civil Service Commission and will be used for research purposes only. Only averages across participating agencies will be released.

PART I

In every organization, there are some employees who just barely earn what they get from the organization in salary, fringe benefits, etc., and no more. These individuals can be referred to as "break-even" employees. On balance, the organization "breaks even" on these employees in terms of their value to the organization compared to the costs of having them on the payroll. Other employees are, on balance, of positive benefit to the organization. The value of their work is greater than the cost of having them on the payroll. Think about the budget analysts working for your agency that you have known and then fill in your answers to the two questions below:

1. Among budget analysts I have known in this agency, the percentage in the "break-even" category is

2. Among budget analysts I have known in this agency, the percentage in the "positive benefit" category is __ %.

PART II

Now, based on your experience with agency budget analysts, we would like for you to estimate the yearly value to your agency of the products and services produced by the average budget analyst. Consider the quality and quantity of output typical of the average budget analyst and the value of this output. In placing an overall dollar value on this output, it may help to consider what the cost would be of having an outside consulting firm provide these products and services.

Based on my experience, I estimate the value to my agency of the average budget analyst at _____ dollars per year.

We would now like you to consider the "superior" budget analyst. Let us define a "superior" performer as a budget analyst who is at the 85th percentile. That is, his performance is better than that of 85% of his fellow budget analysts, and only 15% of budget analysts turn in better performances. Consider the quality and quantitiy of the output typical of the "superior" budget analyst. Then estimate the value of these products and services. In placing an overall dollar value on this output, it may again help to consider what the cost would be of having an outside consulting firm provide these products and services.

Based on my experience, I estimate the value of a superior budget analyst to be _____ dollars.

PART III

A. How accurate, in general, do you feel your estimates are? (Check the most appropriate response.) I feel my estimates are generally accurate to within:

1. $100.00 or less either way.
2. $1000.00 or less either way.
3. $2000.00 or less either way.
4. $3000.00 or less either way.
5. $4000.00 or less either way.
6. $5000.00 or less either way.
7. $6000.00 or less either way.
8. $7000.00 or less either way.
9. More than $7000.00
10. Other (specify)

B. The number of Budget Analysts I supervise is:

1. One or two.

2. Three to five.

3. Five to ten.

4. Over ten.

C. I have been a supervisor of Budget Analysts for:

1. Less than one year.

2. One to three years.

3. Three to five years.

4. Over five years.

APPENDIX B

The Questionnaire Sent to Computer Programmer Supervisors Estimation of Selection Utility

Computer Programmers (GS-334)

Name _____ Dept. _____ Agency _____

INSTRUCTIONS

The dollar utility estimates we are asking you to make are critical in estimating the relative dollar value to the government of different selection methods. In answering these questions, you will have to make some very difficult judgments. We realize they are difficult and that they are judgments or estimates. You will have to ponder for some time before giving each estimate, and there is probably no way you can be absolutely certain your estimate is accurate when you do reach a decision. But keep in mind three things:

(1) The alternative to estimates of this kind is application of cost accounting procedures to the evaluation of job performance. Such applications are usually prohibitively expensive. And in the end, they produce only imperfect estimates, like this estimation procedure.

(2) Your estimates will be averaged in with those of other supervisors of computer programmers. Thus errors produced by too high and too low estimates will tend to be averaged out, providing more accurate final estimates.

(3) The decisions that must be made about selection methods do not require that all estimates be accurate down to the last dollar. Substantially accurate estimates will lead to the same decisions as perfectly accurate estimates.

The information you provide will not be available in identifiable form to anyone in your agency. It will go to the Civil Service Commission and will be used for research purposes only.

Note: The title Computer Programmer in the GS-334 series indicates positions in which the primary emphasis is on using programming methods and techniques. As you know, this series contains three titles in addition to Computer Programer. These other titles are Computer Systems Analyst, Computer Specialist, and Computer Equipment Analyst. In some cases, Computer Systems Analysts and Computer Specialists do a great deal of programming. You should consider the programming work done by such individuals in answering this questionnaire.

PART I

Based on your experience with agency programmers, we would like for you to estimate the yearly value to your agency of the products and services produced by the average GS 9-11 computer programmer. Consider the quality and quantitiy of output typical of the

average programmer and the value of this output. In placing an overall dollar value on this output, it may help to consider what the cost would be of having an outside firm provide these products and services.

Based on my experience, I estimate the value to my agency of the average GS 9-11 computer programmer at _____ dollars.

We would like for you to consider the *"superior" programmer*. Let us define a superior performer as a programmer who is at the 85th percentile. That is, his or her performance is better than that of 85% of his or her fellow GS 9-11 programers, and only 15% turn in better performances. Consider the quality and quantity of the output typical of the superior programmer. Then estimate the value of these products and services. In placing an overall dollar value on this output, it may again help to consider what the cost would be of having an outside firm provide these products and services.

Based on my experience, I estimate the value of a superior GS 9-11 computer programmer to be _____ dollars per year.

Finally, we would like you to consider the "low performing" computer programmer. Let us define a low performing programmer as one who is at the 15th percentile. That is, 85% of all GS 9-11 computer programmers turn in performances better than the low performing programmer, and only 15% turn in worse performances. Consider the quality and quantity of the output typical of the low performing programmer. Then estimate the value of these products and services. In placing an overall dollar value on this output, it may again help to consider what the cost would be of having an outside firm provide these products and services.

Based on my experience, I estimate the value to my agency of low performing GS 9-11 computer programmer at _____ dollars per year.

PART II

In every organization, there are some employees who just barely earn what they get from the organization in salary, fringe benefits, etc., and no more. These individuals can be referred to as "break-even" employees. On balance, the organization "breaks even" on these employees, in terms of their value to the organization compared to the costs of having them on the payroll. Other employees are, on balance, of positive benefit to the organization. The value of their work is greater than the cost of having them on the payroll. Think about the GS 9-11 computer programers working for your agency that you have known and then fill in your answers to the two questions below:

1. Among GS 9-11 computer programmers I have known in this agency, the percentage in the "positive benefit" category is ____%.

2. Among GS 9-11 computer programmers I have knows in this agency, the percentage in the "break-even" category is ____%.

Note: The two percentages above need not add up to 100%.

A. How accurate, in general, do you feel your dollar estimates are? (Check the most appropriate response.) I feel my estimates are generally accurate to within:

1. $100 or less either way.
2. $1000 or less either way.
3. $2000 or less either way.
4. $3000 or less either way.
5. $4000 or less either way.
6. $5000 or less either way.
7. $6000 or less either way.
8. $7000 or less either way.
9. More than $7000
10. Other (specify)

B. The number of computer programmers I supervise is:

1. One or two.
2. Three to five.
3. Six to ten.
4. Over ten.

C. I have been a supervisor of computer programmers for:

1. Less than one year.
2. One to two years.
3. Three to four years.
4. Five to six years.
5. Over six years.

MIRABELLA, Angelo., Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia, and WHEATON, George, American Institute for Research.

CRITERION DEFINITION FOR TACTICAL TRAINING OF COMBAT GROUND UNITS
(Thu P.M.)

The Army Training and Evaluation Programs (ARTEPS) provide reasonable models for the training and diagnostic evaluation of small combat units. More specifically, they outline tasks, conditions, and standards. However, criterion definition, i.e., standards, has been and continues to be a problem. The problem becomes increasingly complex as the tasks become more tactical, that is, as they involve more game playing, less proceduralized behaviors. The advent of MILES, Multiple Integrated Laser System, has increased the accuracy with which casualties can be assessed, but MILES in and of itself, does not solve the criterion problem. ARI, has been and continues to be involved in research aimed at finding more effective ways to define criteria for tactical training. This paper will review some of the measurement issues which are being addressed and the methods which are being explored to solve those issues. The use of three methods, in particualar will be discussed: Delphi, gaming, and CARMONETTE, a computer program which simulates two-sided tactical combat.

Criterion Definition For Tactical Training of Combat Ground Units

Angelo Mirabella
Army Research Institute
for the Behavioral and
Social Sciences

Eugene Johnson, III
George R. Wheaton
American Institutes
for Research

1. *Introduction: The Military Problem*

Narrowly defined, criterion definition refers to the establishment of one or more cut-off points on some scale of measurement in order to separate acceptable from unacceptable behavior, or in the case of multiple cut-off scores, to separate various degrees of acceptability of performance. But broadly defined, effective criterion definition refers to a performance evaluation system that begins with reliable, objective job and task analysis, and then goes on to specify measurement concepts, dimensions, scales, instruments, and *finally*, some meaningful cut-off score. For purposes of unit tactical evaluation, criterion definition, in either the narrow or the broad sense, is an extraordinarily complicated, even intimidating problem. It is hardly surprising, therefore, that the state of the art has not advanced beyond a rudimentary level.

Why should those involved with managing and conducting unit tactical training be concerned with criterion definition? What is the payoff in going beyond the current state of the art as defined by the Army Training and Evaluation Programs (ARTEPs)? The payoff could be a major increase in the efficiency with which training resources are allocated. We hear continual remonstrances from the Training and Doctrine Command (TRADOC) that training resources are shrinking and therefore more efficient management of resources is required. Major improvement in criterion definition can provide one mechanism for this improved efficiency. If, for example, a three-hour company-team exercise uncovered major leadership deficiencies, remedial training would be cheaper via the use of game boards or battle simulations, rather than through a rerun of the entire three-hour full-scale exercise as is now typical. But it would be very difficult for a training manager to make such a resource allocation decision unless he had confidence in his ability to measure the performance of various elements of the team as well as dependable criterion or cut-off scores for deciding where the training problems lie.

More specifically, an effective solution to the criterion definition problem could contribute to a number of developments currently evolving in the Army's training community. For example, once the National Training Center at Fort Irwin, California becomes fully operational, it will be producing data on casualties, position location, and communication. Software for displaying detailed video color replays of the exercises has already been developed, and prototype-tested for company team exercises. However, strategies and rules for interpreting these data are needed to effectively utilize the enormous data collection capability

655

that will eventually exist at NTC. Criterion definition is needed for judging the significance of casualties, as well as for determining the effectiveness of maneuver, communication, and leader decisions.

Similar criteria could also be useful when the Multiple Integrated Laser Engagement Simulation System (MILES)[1] becomes fully fielded at home stations. The ARTEPS being developed to accompany home station MILES exercises are an improvement over the existing ARTEPS for conventional training, but criterion definition is still rudimentary.

The state of the art of criterion definition for unit evaluation can be illustrated with a page from a Combined Arms Army Training and Evaluation Program (ARTEP 71-2). Table 1 shows portions of the Training and Evaluation Outlines for Movement to Contact and Hasty Attack, for a tank/mechanized-infantry company team. These training and evaluation outlines follow instructional system design (ISD) format in their statements of tasks, conditions, and standards. However, the criteria for success or failure are not very explicit. For example, notice the standard for the task described as "Prepare for the movement." It is left to evaluator judgment to determine what it means to "maximize the combat power and to minimize exposure and vulnerability to opposing force fire." This statement is very subjective, and open to a wide variation in interpretation. Newer versions of the Training and Evaluation Outlines for Movement to Contact and Hasty Attack (Table 2) have been developed for use with the MILES training program. These versions of ARTEP are improvements in several respects. The task analysis is considerably more detailed. The standards column has been improved with the use of a general standard backed up by observable events which can be taken into consideration in making a judgment of success or failure. The changes are partly reorganizations and reformattings of information contained in the older ARTEP, but some attempt has been made to reduce subjectivity.

In what ways can the existing state of the art be pushed? Consider first the narrower definition of a criterion as a cut-off point on some scale of measurement where quantitative scales already exist. The revised ARTEP has, in fact, established a quantitative cut-off for casualties. It provides a rule-of-thumb figure of 30 percent casualties as the limit for a remaining viable force. Such a rule of thumb may be satisfactory for a very limited set of circumstances and for a well-trained combat unit, but for the training manager, a *series* of such cut-off values for various stages of training and for a *variety of conditions* would provide a more accurate basis for evaluating a particular unit's progress. For example, a local commander may not want to use precisely the same force ratios that are described in the ARTEP. Or the particular terrain over which his exercises are run may be especially advantageous or especially disadvantageous to the attacking force being evaluated. What is needed is a more flexible and at the same time more comprehensive approach to defining criteria.

The approach discussed in this paper makes use of not one, but three somewhat related types or sets of criterion data. The first set is obtained somewhat the way horseraces are handicapped. This first set would consist of a family of

[1] Engagement Simulation is a method for conducting two-sided, realistic, free-play maneuver arms field exercises, with accurate casualty assessment, using either weapon-mounted telescopes or laser devices. When played with telescopes the method is called REALTRAIN. When played with laser guns, the method is called MILES.

TABLE 1
Training and Evaluation Outline

UNIT: Company Team **MISSION:** Movement to Contact/Hasty Attack

ID NO./TASK	CONDITIONS	TRAINING/EVALUATION STANDARDS	S	U
6-7-A Prepare for the movement.	Company team commander is provided task force warning and operation orders containing the information in the general conditions. He is given time to prepare for the movement (conduct a map reconnaissance, select formations, and issue his order).	Company team commander selects formations that maximize the combat power of the company and minimizes its exposure and vulnerability to opposing force fire through proper techniques, terrain, and fire support.		
6-7-B Conduct the movement.	Task force operation order provides the time at which the company team must begin to move.	a. Movement techniques appropriate for the terrain and expected degree of opposing force contact are used. Maximum use is made of covered and concealed routes. Advancing elements are supported from overwatch positions by other company team elements, and by organic indirect fires.		
(5 Additional Tasks)				
.	.	.		
.	.	.		
.	.	.		
.	.	.		
.	.	.		

TABLE 2

Training and Evaluation Outline: Test ARTEP 71-2 (for Use in MILES Implementation)

COMPANY TEAM — **MOVEMENT TO CONTACT/HASTY ATTACK**

Task	Condition	Standard
1. Team prepares for mission.	Team receives TF OPORD for movement to contact/hasty attack.	SEE MISSION 3-I-1, PLAN AND PREPARE OPERATIONS. (Use appropriate tasks throughout the mission.)
2. Team crosses LD.	OPFOR is not within effective range.	TEAM CROSSES THE LD ON TIME IN FORMATION SPECIFIED BY OPORD. Observable Events 1. Overwatch element is set. 2. Lead elements cross on time at the correct location. 3. Team is in specified formation and order of movement. 4. Team accurately reports crossing LD to TF.
3. Team conducts tactical movement.	Conditions above apply.	TEAM MOVES RAPIDLY (EVALUATOR-JUDGMENT BASED ON TERRAIN ANALYSIS) WITHOUT EXPOSING MORE THAN THE LEAD ELEMENT TO OPFOR OBSERVATION. Observable Events 1. Team makes maximum use of cover and concealment. 2. Team is in a formation that permits early detection and reaction to OPFOR with a minimum sized force forward.
(9 additional tasks)		
.	.	.
.	.	.
.	.	.

performance expectations for a unit, based upon various combinations of conditions under which evaluation of that unit takes place. How would one go about generating such expectations? If time and resources were unlimited, one could run field exercises under specified controlled conditions, thus providing an *empirical data base* for subsequent ARTEP evaluations. However, running a large number of standardized unit field exercises is not feasible because of cost. An alternative is to approximate field exercises through the use of low-cost simulations. There are at least three approaches to low-cost simulation that ARI/AIR are interested in researching. Included is the DELPHI technique, the use of board gaming, and the use of CARMONETTE, a computer simulation. The general purpose for such simulations is to generate large quantities of baseline data on an actuarial basis for a large number of field exercise conditions under which training occurs, and for multiple points in the training cycle.

A second criterion set, one which we believe critical to a comprehensive evaluation system, would be provided by having tactical experts generate *doctrine-based idealized solutions* for various exercise conditions. Gaming, DELPHI, and CARMONETTE are again potential tools for contributing to such solutions.

A third essential criterion set would be provided by combat system requirements, i.e., the *system performance requirements* generated from threat analysis. This third set would be especially important in telling the overall force developer/manager where training and doctrine have reached the limits of their contributions to battlefield success and where improvements in firepower, instrumentation, and/or mechanization are needed.

Where casualties are concerned, we can limit ourselves to the narrower criterion definition because casualties provide a fairly clear-cut quantitative scale. In this case, our only problem is to develop cut-off scores appropriate to a variety of tactical conditions, levels of training, doctrines, and system requirements. However, evaluation of tactical performance would be shallow and incomplete if we relied strictly on the measurement of casualties. For training diagnosis, we need to evaluate the tactical processes which eventually lead up to the infliction of casualties and to the taking of territory. In this case, we need to deal with the broader criterion definition, i.e., we do not yet have measurement concepts, scales, and instruments for tactical processes. We need these before we can face the problem of setting cut-off scores and before we can establish the relationship between process and outcome measures.

Pushing the state of the art in this case is an extremely challenging problem. We would like to develop scales for such concepts as "use of cover and concealment," "concentration of firepower," "use of communication," or "movement techniques," and then subsequently develop standard-setting procedures.

2. *The Scientific Problem*

Criterion development for tactical training is a systems analysis problem beginning with front-end analysis. It requires one or more simulation test beds (e.g., Engagement Simulation) that reflect training requirements identified through front-end analysis. It requires objective if not quantitative scales of measurement for both tactical processes and exercise outcomes. It also requires dependable baselines for interpreting performance data. And finally, it requires some way to partial out the effect of non-training variables, since true unit tactical

proficiency may be obscured by force ratio, terrain, and a host of other variables. We have tried to account for this complexity through the development of something we refer to as COTEAM, the Combat Training Effectiveness Analysis Model. COTEAM is outlined in Figure 1. Near the top of the figure is a list of non-training variables (i.e., such exercise conditions as weapon mix, terrain, etc.) that could have an impact on the outcome of any particular training exercise, and that therefore could influence conclusions about training effectiveness.

Under the current unit training model, these exercise conditions are alluded to in the ARTEP Manual and then further specified for any particular evaluation exercise. In the evaluation system we envision, a *criterion-value generating subsystem* would be used to generate a set of criterion values for attrition and for various tactical processes based upon these conditions and upon the evaluated unit's nominal stage of training. Included here would be values generated to supply the first two types of criterion data mentioned earlier. The first would consist of values for specific levels (i.e., stages) of training; the second would consist of values obtained from the play of the exercise by tactical experts who, rigorously following doctrine, would generate a school solution.

These same conditions would apply to Engagement Simulation (ES) evaluation exercises based upon the ARTEP Manual. Outcome values (e.g., casualties) obtained from the field exercise could then be compared with criterion values to derive conclusions about actual training level and combat readiness. These conclusions, in turn, could provide the basis for planning a finely tuned remedial training program.

What we've laid out here is a general scheme for defining and using performance criteria for evaluating unit training effectiveness. Coming up with the specific methods for making such a scheme work is another matter, one which is currently the subject of research at the Army Research Institute with contract support by the American Institutes for Research. The major objective of this research is to develop and validate the use of low-cost, low-fidelity simulations (such as board gaming) as tools for defining unit performance criteria.

We have already completed some pilot efforts to develop and validate a board game designed to generate baseline (i.e., criterion) data for ES exercises. Rules for this game, called SCUE (Small Combat Unit Evaluation) closely parallel those used to control ES exercises. As part of an ES (i.e., REALTRAIN) validation study at Fort Carson, Medlin (1979a)[2] demonstrated similar casualty outcomes for company-team field exercises and for game exercises. Illustrative results are shown in Table 3. Where data between the field and game exercises were discrepant it could be shown that there were significant differences in control procedures for the two types of simulation.

Medlin (1979b)[3] further reinforced confidence in SCUE by applying the Turing test to it (Turing, 1950).[4] This was a test of the ability of military officers

---

[2]Medlin, S. M. Behavioral Forecasting for REALTRAIN Combined Arms. U.S. Army Research Institute for the Behavioral and Social Sciences. Technical paper 365, Alexandria, VA, May 1979a.

[3]Medlin, S. M. A Partial Validation of Forecast Engagement Simulation Exercise Outcomes. U.S. Army Research Institute for the Behavioral and Social Sciences. Technical Paper 382, August 1979b.

[4]Turing, A. M. Computing Machinery and Intelligence. *Mind*, 1950, *59*, 433-460.

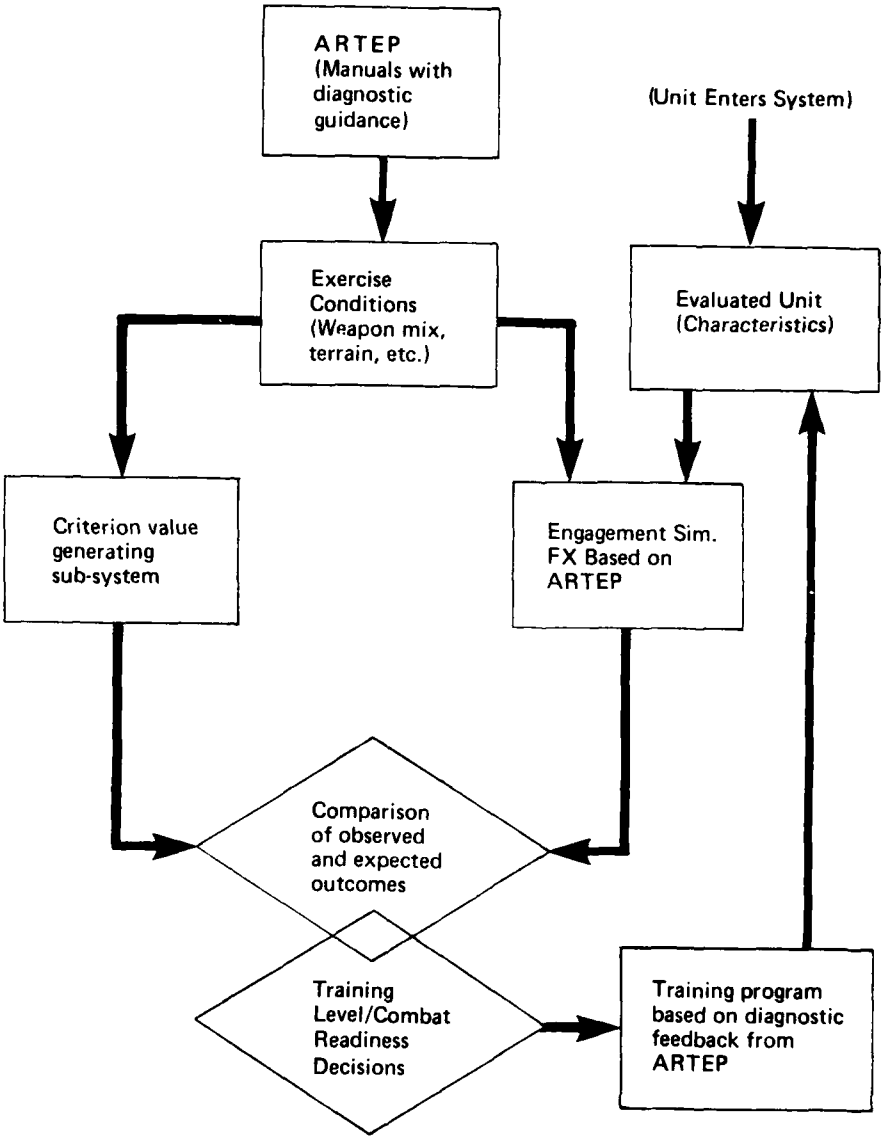


Figure 1. COTEAM EVALUATION SYSTEM

TABLE 3
Percent Casualties Across Eight Engagement Simulation Exercises

CASUALTIES INCURRED (Percent Across all Exercises)

	TOWs	TANKS	APCs	JEEPS	INF
Game	7	31	13	0	49
Field	8	36	7	3	47

CASUALTIES INFLICTED

	ARTY*	TOWs	TANKS**	APCs	LAWs	90s	INF
Game	14	12	31	1	5	7	29
Field	22	14	13	0	8	9	33

* Possible result of differences in control procedures.
** Difference reflects availability of Beehive to Board Game Players.

TABLE 4
Length of Bounds During Bounding Overwatch

EXPERIENCED SUBJECTS		INEXPERIENCED SUBJECTS			
1	700m	1	1000m	4	250
	1000		250		250
2	500	2	500		250
	700		600		600
	700	3	250		600
	500		250	5	250
3	600		500		600
	250		400		600
	700		200		250
4	250		250		250
	600		600		
Mean	591			Mean	414

to tell the difference between ES field data and SCUE data. Results showed that the two sets of data could not be distinguished.

Wheaton et al. (1980)[5] applied a variation of SCUE to the evaluation of tank platoon gunnery (i.e., Battlerun) performance. This study was of particular interest in demonstrating how a tactical process could be reduced to a quantitative scale and how in turn, a quantitative criterion might be defined for that process. Wheaton et al. measured the length of tank section bounds (during bounding overwatch) as units moved from a line of departure to assault positions. These data were collected from experienced and non-experienced armor officers. The pilot data are shown in Table 4. Note the lengths of the bounds are substantially and consistently longer for the more experienced group of officers. Such distributions of data as these could conceivably be used to define training or experience-referenced performance criteria i.e., performance expectations as functions of training level. Here we are talking about one of the three categories of criteria which were introduced earlier. We might also have developed the second type of criterion, i.e., the school solution, by having tactical experts carefully elaborate the appropriate doctrine, using a Delphi Technique, and then apply that doctrine to the simulated exercise. The first type of criterion data could assist in categorizing the training level of subsequent examinees. The second set of criterion data could assist in making combat readiness judgments.

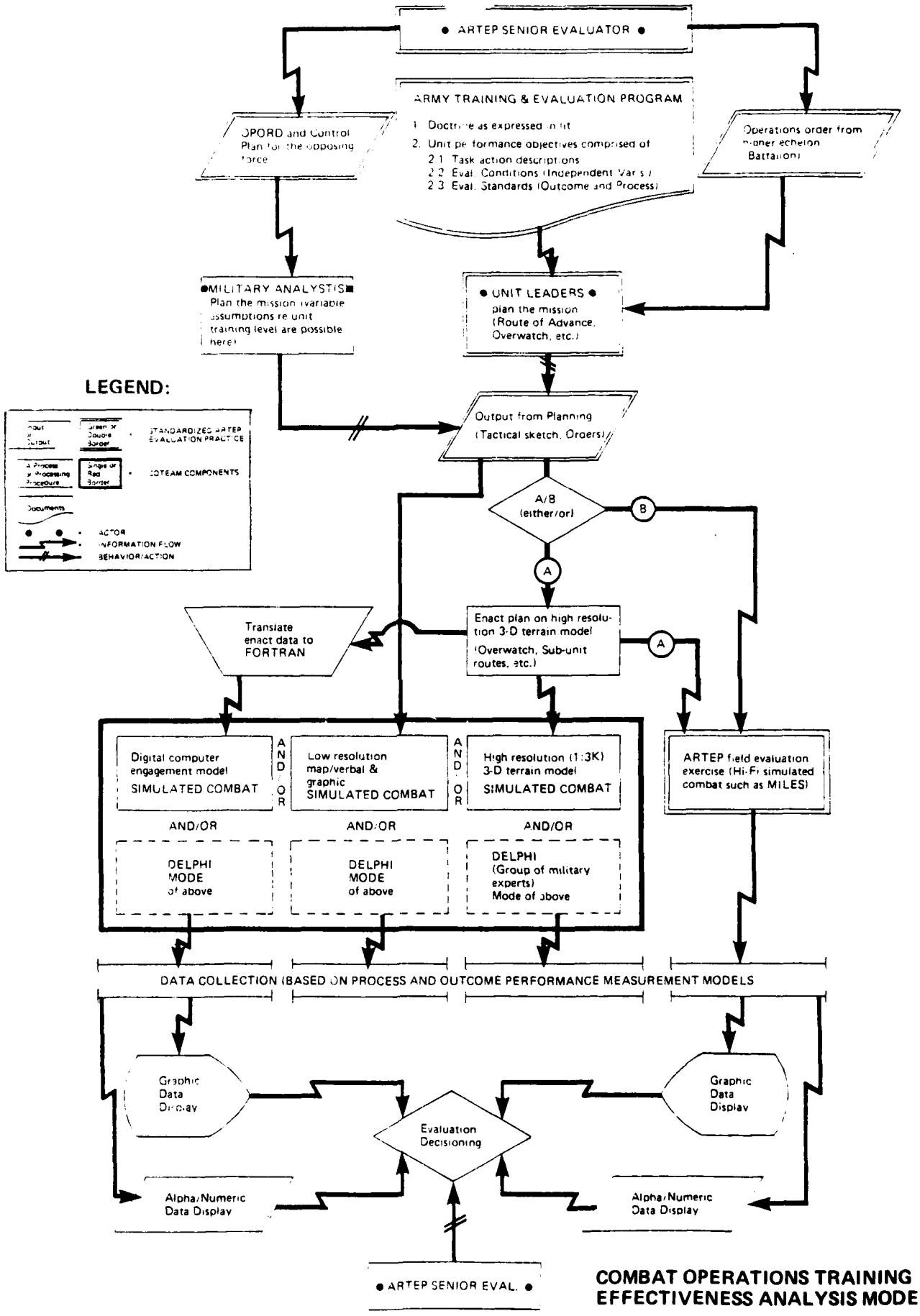
3. *Current Research Efforts*

Further efforts to develop these methods for defining and using unit performance criteria are currently under way along with efforts to detail the COTEAM evaluation model outlined earlier in Figure 1. The evaluation scheme which is flow charted in Figure 2 is experimentally-oriented. On the basis of research to be conducted over the next 6 to 12 months, this will evolve into an operational scheme which can be used both for home station and NTC evaluation. We will developmentally test the overall model, its specific methods for generating and using criterion data, and combinations of the methods.

The revised experimental model follows standard military procedures for ARTEP evaluations, except that additional evaluation-oriented processes are included. For example, doctrine as interpreted on the ground by a team of military experts would play an explicit role in criterion definition. Also, certain combat operations which are normally carried out informally would be formalized, so that criterion or evaluation data can be collected from these operations. For example, following mission planning by the unit leader(s), they would enact the plan on a high-resolution scale terrain model. Thus, data would be provided for plan evaluation.

The flow chart depicted in Figure 2 begins with operations that are part of the Army's current evaluation model. ARTEP and the Senior ARTEP Evaluator are points of departure. ARTEP defines the doctrine that is implicit in the ARTEP T&EO (Tasks, Conditions, and Standards). The Senior Evaluator disseminates OPORDs

[5] Wheaton, G.; Allen, T.; Johnson, E.; Drucker, E.; Ford, P.; Campbell, R.; and Boycan, G. U.S. Army Research Institute for the Behavioral and Social Sciences. Working paper, 1980.



COMBAT OPERATIONS TRAINING EFFECTIVENESS ANALYSIS MODE

from Battalion to the unit leaders, who in turn plan and carry out the mission. The remaining operations on this chart would be additions to the current model.

The first new addition would be the use of military analysts who receive OPORDs from the Senior Evaluator and plan the mission in parallel with the unit leaders. Outputs from this planning (i.e., tactical sketches, orders) by the military analysts and possibly unit leaders could then provide the basis for "sand tabling" the exercise, except that in this case high resolution 3-D terrain models would replace the traditional sand table. These outputs would in turn, support two subsequent activities: gaming on a high resolution 3-D terrain model or simulation of the mission on a digital computer via CARMONETTE. These latter two activities would provide two of three methods for generating criterion data. The third method makes use of low resolution (1:25K) maps and *graphic representation of tactics* and draws on the output from the initial planning sessions. Each of the three criterion methods could be used in combination with Delphi panel procedures. It is essential to the model that each of these criterion definition methods simulate the same ES field exercises. To that end the process and outcome measures that need to be developed will be common to all the methods.

The output of the model shown on the bottom of Figure 2 provides information for training diagnostic and/or combat readiness decision making. This decision making would follow from a comparison of ES-field exercise results with the output from the three criterion-generating methods or from just one of the methods.

Note finally, that the evaluation scheme ends as well as begins with the ARTEP Senior Evaluator. The Senior Evaluator, aided by his evaluation staff, is *the* evaluation decision-maker. He must integrate massive amounts of data to make both summative and diagnostic evaluative judgments. COTEAM is a decision-making aiding system designed to assist the Senior Evaluator in coping with the hyper-complex, multidimensional decision space that results from high fidelity, two-sided land combat engagement simulation.

Current Research

The research that we currently have underway is addressing three aspects of the experimental model depicted in Figure 2. These deal with: 1) methods for specifying the tactical processes and outcomes upon which to focus unit evaluation; 2) methods for generating performance criterion values, referenced to either a unit's stage of training or to military-science-based land combat theory; and 3) methods for interpreting unit performance observed during simulated combat in the field using model-generated standards and criteria.

The specification of tactical processes and outcomes traditionally has been based upon various types of task analysis. We are building on these efforts by using concepts such as, "maximize *offensive utility* and *defensive utility*," as second-order or intermediate performance objectives. Relating unit tasks and substrate behaviors to such dimensions should permit tactical processes to be prioritized for measurement and evaluation. We are presently focusing on the task information provided in ARTEP Level 1 Mission statements.

Development of a methodology for generating performance criteria for evaluation standards is being approached experimentally. We are exploring different criteria-generating methods using movement-to-contact/hasty attack missions at Fort Irwin as the test bed.

A decision-aiding system to support the Senior Evaluator and staff in carrying out their difficult information-processing task is yet to be developed. The third major aspect of COTEAM research is to provide a well human-engineered, psychometrically sound methodology for comparing standard level criterion performance with observed performance ensuing from hi-fidelity simulated combat.

ARI/AIR is exploring multi-variate statistical comparison processes, e.g., profile analysis, and adaptations of psychophysical judgment methods such as behaviorally anchored magnitude estimation. We are also developing non-parametric comparison techniques for tactical graphics data such as that describing tank section route selection or fields of observation and fields of fire. A third approach for comparison of standard and observed performances uses a digital computerized land combat model to replay observed ARTEP data, so that both criterion (combat model-generated standard) and observed (high fidelity field) performance data are in a common language and enactment mode. Then both statistical and human judgment-based comparison processes using either alpha-numeric or graphic data from selected "slices of battle" can be readily researched.

Additional ideas from the military testing community pertaining to any aspect of COTEAM research and implementation would be welcomed by ARI/AIR.

MITCHELL, Lt. Col. J.L., KEETH J., and WIEKHORST, Lt. L., USAF
Occupational Measurement Center, Randolph AFB, Texas.

TRENDS IN REENLISTMENT INTENT: GENERAL VERSUS OCCUPATION-SPECIFIC ANALYSIS (Thu A.M.)

The USAFOMC has collected responses to a few standard job interest and reenlistment items in all studies conducted over a number of years. Typically, such data are summarized at the end of each year to serve as a comparative sample for the studies conducted the following year. Analysis of the general trend in reenlistment intent for the years 1975 through 1980 suggests a general decline among first enlistment personnel (54% in 1976 - 34% in 1979). However, sorting data into six major types of occupations (aircrew specialities, mission equipment operations, mission equipment maintenance, command support personnel, direct support personnel, and medical specialties) reveals several major differences. When these data are plotted by major occupational type across time, it was found that the trends vary markedly. For command support and medical specialties, there has been an increase in first enlistment intent to reenlist across time. For direct support and mission equipment specialties, there is a general decrease across time, and for other groupings the trends are mixed. This analysis suggests that general trend analysis in the areas of job satisfaction or reenlistment intent could be very deceptive and we should be reluctant to release such data for use in management decision making. The use of major occupational groupings is to be preferred since it permits more speicific analysis and prediction of potentioal problems for recruiting, training, and management. Finally, even here data must be used with caution depending on which occupations are studied in any given year. When only one or two occupations in an occupational-specific to reveal any general trend.

TRENDS IN REENLISTMENT INTENT: GENERAL VERSUS OCCUPATION-SPECIFIC ANALYSIS[1]

Lt Col Jimmy L. Mitchell
Mr Jim Keeth
1Lt Linda Wiekhorst

Occupational Analysis Program
USAF Occupational Measurement Center (ATC)
Randolph AFB, Texas 78148

INTRODUCTION

For a number of years the USAFOMC has been collecting occupational information involving about 51 enlisted specialties per year. At the end of each year, data on standard job attitude questions and reenlistment intent are summarized for use as a comparative data base for the following year (Olivo and Weber, 1978). Such standard questions include items dealing with job interest, how the job uses the individuals's talents and training, and the individual's reenlistment intent. Within a study of a specialty, these types of information can be contrasted among job groups as one way to identify potential problem areas. This kind of within specialty variance among job groups is a key strength of the occupational analysis methodology and permits the in-depth study of personnel utilization, classification structure, and the appropriateness of training programs (Mitchell and Driskill, 1979). Typically, when we find a job group of junior airmen performing relatively few tasks where their talents and training are not well utilized, we also frequently find that their job interest is also low and quite often there are few who plan to reenlist.

This is by no means a perfect relationship. Alley and Gould (1975:9) found correlations of .31 between job interest and intent to reenlist and .27 between use of talents and training and career intent. These authors also noted changes in intent to reenlist for different year groups and different specialties. They concluded that statements of career intent had different functions and validities for various specialties but that some grouping of specialties led to more realistic predictions (Ibid:24). This conclusion implies that a functional grouping of specialties is appropriate when making comparisons of stated reenlistment intention data.

Guinn, et al. (1977) attempted to develop a general prediction equation to predict actual reenlistment. Their Reenlistment Potential Index (RPI) made use of a variety of background and attitudinal variables to predict reenlistment. A multiple correlation (R=.51) was obtained which was statistically significant but which did not account for the majority of the variance in reenlistments. These authors concluded, as had Alley and Gould (1975), that a specialty-specific prediction of reenlistment was probably more appropriate (Guinn, et al. 1977:14).

OPERATIONAL PROGRAM

In operational practice, reenlistment intent and job interest data are most useful in characterizing job groups within an occupation or in making comparisons across related specialties in the more complex multispecialty studies. For example,

1. The opinions expressed in this article reflect the thoughts of the authors and are not official policy of the USAF Occupational Measurement Center or the U. S. Air Force.

in a study of Air Force computer operators and programmers, it was noted that 55 percent of the operators said they would reenlist (yes or probably yes) whereas only 47 percent of the programmers had such an intent (Wiekhorst, 1980:Table 21). These data are for the specialty as a whole; a more meaningful comparison is to contrast the reenlistment intent for those individuals in their first enlistment (1-48 months TAFMS), which is the group where reenlistment is critical.

For the first term computer operators and programmers, 38 percent and 25 percent respectively say yes or probably yes in the two specialties (Ibid:Table 29). This is the same pattern as for the total groups but the levels of intended reenlistment are much lower.

Since these specialties have been studied more than once in the Air Force Occupational Survey Program, it is also possible to contrast these data over time to see what trends emerge between 1977 when they were last studied and 1980 when the present report was completed. When this kind of contrast is made, the difference becomes more interesting. Where computer operators are about the same (39 percent in 1977 and 38 percent in 1980), the data for programmers is different (32 percent in 1977 versus 25 percent in 1980). Since job interest, perceived use of talents and perceived use of training are all higher in 1980 than in 1977 for the programmers, some other variable must account for this drop. A review of the same kind of data for second term personnel (49-96 months TAFMS) and career programmers (97+ months TAFMS) also show some drop in reenlistment intent (55 to 43 percent for second term, 67 to 61 percent for career). However, among these groups, there is a constant decline in job interest, perceived use of talent, and perceived use of training (Ibid:Tables 30 and 31). We suspect that military programmers are more in demand today in the civilian labor market at very attractive salaries. However, it is probably much more complex an issue.

There is no simple explanation for these trends across time - there are a number of factors which may have been responsible for such changes. To put some perspective on the larger Air Force context, in the occupational analysis program we make a contrast with a sample of comparable specialties which were studied the previous year. This comparison gives us a broader picture by which to assess current results. In this case, the comparative sample would be other direct support specialties as outlined in AFM 26-3 (see Olivo and Weber, 1978:p74), which include weather, vehicle maintenance, civil engineering fields, transportation, food services, supply, education and training, and security police. When the direct support specialties surveyed in 1979 were summarized, it was found that 47 percent of the 7,141 direct support personnel said they would reenlist, versus 55 percent for computer operators and 47 percent of the programmers. These data let us see that the programmers are average in terms of reenlistment intent where computer operators are above average (Wiekhorst, 1980:Table 21). This finding certainly presents a much more positive picture than was developed looking just at trends within the specialties across time.

GENERAL TRENDS

The data we've discussed up to this point suggests that we need to be conscious of general trends in reenlistment across time, but that we need to be cautious about the generality of the groups which are to be used as a basis for comparison. This caution becomes a very serious matter when we examine our general data for all those individuals who have been surveyed over the last few years in the Air Force occupational survey program.

When we look at the overall reenlistment intent of those surveyed each year, we can see that there is a general decline in reenlistment intent (see Figure 1) over time. We've not tested to see if these are statistically significant differences; however, with the sample sizes involved they most probably are. In any case, in an era of the all volunteer force, we can see that there is a problem of major importance when reenlistment intent is declining, and the question of statistical significance becomes trivial.

When we focus on the first enlistment population, the pattern of decreasing percentages remains, although the absolute level is much lower. It is this first enlistment group which is our major concern when we look at potential career manning, utilization patterns, and training programs.

The picture presented by these data are pretty depressing and if we project this trend to future years, we could get panicked quite easily into making sweeping generalizations about the failure of the All Volunteer Force concept. However, recalling Alley and Gould's conclusion that there is differential validity by career fields, perhaps we should withhold any conclusion until we examine more specific groupings of specialties to see if there are any differential trends.

Using the groupings of specialties outlined in AFM 26-3 (see Olivo and Weber, 1978:p74), we can break the overall data into six major areas: aircrew; mission equipment operations; mission equipment maintenance; command support; direct support; and medical. Figure 2 displays first termer reenlistment intent for 1976 studies grouped by these specialty clusters. As you can see, there is considerable variance by occupational group with aircrew being exceptionally high versus other occupational areas. In Figure 3 for 1977 studies, reenlistment intent ranged from a high of 56 percent among aircrew (flight engineer, loadmasters, aerial refuelers, etc.) to a low of 38 percent for medical specialties.

For 1978, the data show some changes. Figure 4 reflects that no occupational surveys were conducted for specialties in the mission equipment operations area. Again, however, for those areas studied in 1978, we can see sizeable differences in the percentages who expect to reenlist. Medical specialties are about the same as in 1977 but the mission equipment maintenance area has dropped and is now the lowest at 36 percent.

For 1979 (see Figure 5), we can see some additional changes. Mission equipment maintenance is still lower at 34 percent but direct support is even lower at 32 percent. However, command support area is up to 47 percent and the medical specialties are up to 41 percent. Thus, we start to see that there are, in fact, differential trends across time for the various occupational groupings.

Another way of reviewing this data is by occupational group across time. Figure 6 displays reenlistment intent for medical areas across time. Obviously, the overall trend here is stable or slightly rising across time where the overall trend had been downward. This clearly suggests that reenlistment in the medical specialties may not be as great a problem as for some other occupational areas. Please note that here we have added in the data for those studies we have done in 1980; these data became available as we were developing this paper.

Figure 7 gives some rather incomplete data for the mission equipment operations area. For those years where occupational studies were accomplished in this area, the trend is a very sharp decline which has leveled out in 1980. This would suggest that this area is one where we may have major problems - we need to look further in these studies as to the causes of dissatisfaction and lack of reenlistment intent.

Figure 8 displays reenlistment intent among those working in mission equipment maintenance. Here again we have a decline from 49 percent in 1977 to 34 percent in 1979 and 33 percent in 1980; this roughly parallels the decline we saw earlier for the mission equipment operations area. This is indeed a troublesome trend, since both these areas are critical to the major Air Force mission and are high investment areas in terms of technical training.

Figure 9 summarizes the trends from 1976 through 1980 for the direct support areas. We see a decline across time although the function is not as steep as for mission equipment and the trend seems to reverse itself in 1980. In 1979 at our low point, our data includes the security police career area and these data may have a major impact on this chart simply by weight of numbers. These results suggest that we need to be conscious of the studies which were done each year and their relative size in terms of how one or two specialties may impact on historical trends (see Table 1).

Figure 10 displays reenlistment intent among command support personnel; this grouping includes such specialties as first sergeant, safety, procurement, logistics, administration, personnel, and public affairs. Here we see a mixed trend upwards across time.

Finally, the aircrew specialties are shown in Figure 11. The percipitious drop from 89 percent in 1976 to 44 percent in 1978 is at first alarming. However, only one specialty was surveyed in 1976, the flight engineers (AFS 113X0); two in 1977 including loadmaster (AFS 114X0) and pararescue (AFS 115X0); and only two in 1978; defensive aerial gunner (AFS 111X0) and inflight refueling (AFS 112X0). In all cases, intent to reenlist was well above average; in the case of flight engineers, an 89 percent is a rather fantastic reenlistment intent suggesting that flight engineers are exceptionally well motivated in their jobs. No aircrew studies were done in 1979 or 1980.

What we are seeing here are specific specialty differences rather than historical trends; when only one or two studies are done in a year, this does not provide a sufficient sampling of the general occupational area to wash out the specific specialty. There are so few of the aircrew specialties that it is impossible to have a general comparative group and any contrasts need to be made against other aircrew specialties individually.

CONCLUSIONS

From the data displayed here it is clearly evident that the use of a single general statistic covering all Air Force specialties tends to disguise the trends visible for generic occupational areas. Such an overall figure presents an erroneous picture of declining reenlistment intent and to the conclusion that the All Volunteer Force is a failure. When generic occupational areas are examined in more detail, a different picture emerges. There are some key problem areas, particularly among mission equipment operations and maintenance; however, there are other areas where expressed intent to reenlist is higher than in previous years. The medical and command support areas are among those areas where reenlistment intent has improved; the implication (but not proof) is that management in these areas may be doing a good job at motivating their personnel to an Air Force career.

These data also suggest that future Air Force internal recruiting efforts need to be targeted specifically toward key occupational areas, such as mission equipment operators and maintainers. Air Force management is well aware of the problems and is working hard to retain such personnel. Recently the decision was made to upgrade our internal recruiting program by retaining the career adviser (AFS 732X4) as a distinct specialty and to provide such career adviser with the manning and resources needed to conduct an increased effort to motivate first termers to reenlist. With this kind of effort, by targeting specific occupational areas, and through the improving compensation and benefits now being enacted by Congress, the USAF hopes to solve our future personnel retention problems.

In Figure 12, we have again reproduced the total trend and have added in the 1980 data. This overall picture is a hopeful one which suggests that our internal recruiting efforts are having an impact. An alternative explanation might be that jobs are hard to find on the outside. Again, however, the data presented earlier was convincing that there are key occupational areas, such as mission equipment operations and support, where we still have problems. You have to look at the more specific occupational groupings to see where the problems are.

Finally, the data displayed here tend to confirm the conclusions of Alley and Gould (1975) and Guinn, et al. (1977) that a general prediction equation for reenlistment of Air Force personnel is probably not practical. It appears to be much more realistic to deal with a few generic occupational areas or with individual specialties rather than to make general policies which are based on overly generalized data.

References

Alley, William E. and R. Bruce Gould
1975 Feasibility of estimating personnel turnover from survey data - a longitudinal study. AFHRL-TR-54, Air Force Human Resources Laboratory, Lackland AFB, Texas 78236

Guinn, Nancy, George Berberich and Bart M. Vitola
1977 Reenlistee/Non-reenlistee profiles and prediction of reenlistment potential. AFHRL-TR-77-24, Personnel Research Division, Air Force Human Resources Laboratory, Lackland AFB, Texas 78236

Mitchell, J. L. and Walter Driskill
1979 Variance within occupational fields: job analysis versus occupational analysis. Proceedings of the 21st Annual Conference of the Military Testing Association, San Diego, CA.

Olivo, John X. and Elena J. Weber
1978 The use of job satisfaction data in the occupational survey program. Proceedings of the 20th Annual Conference of the Military Testing Association, Oklahoma City, OK:65-74.

Wiekhorst, Linda
1980 Occupational Survey Report of the Computer and Programming Specialties AFS 511X0 and 511X1. AFPT 90-511-413, USAF Occupational Measurement Center, Randolph AFB, Texas 78148.

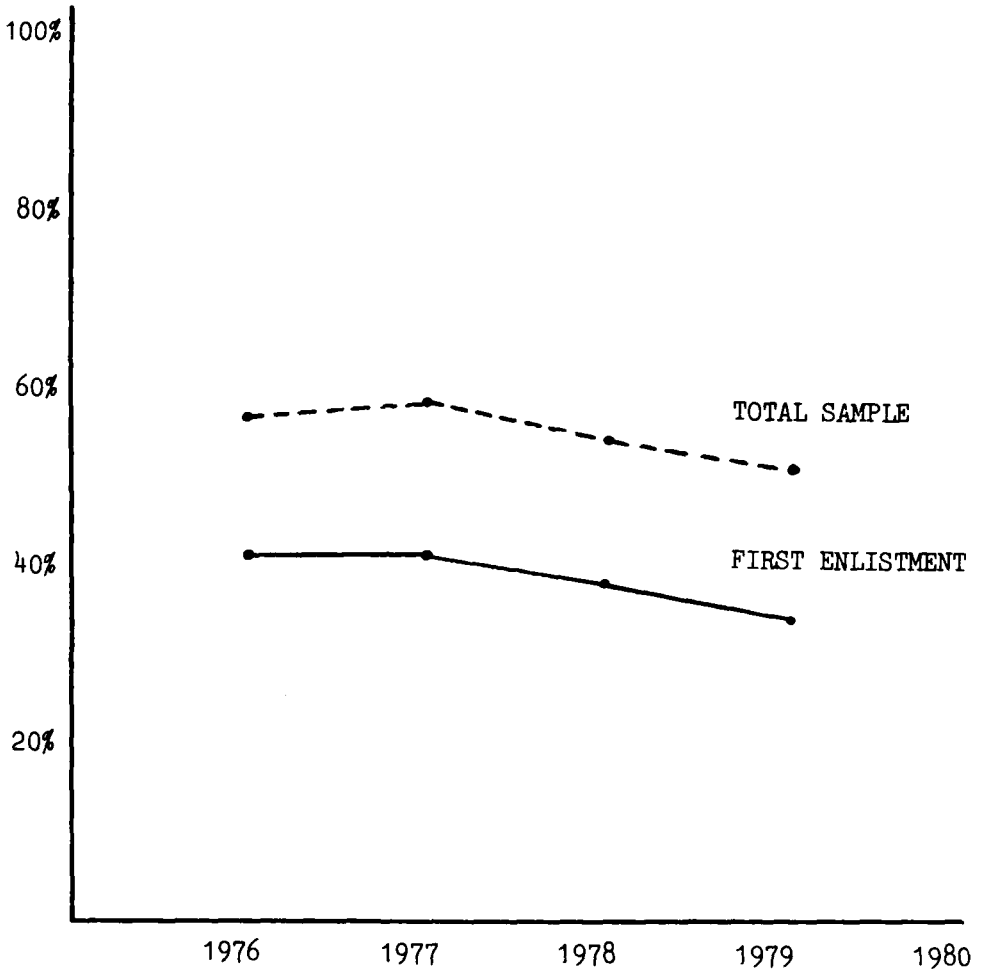


FIGURE 1. OVERALL TRENDS IN REENLISTMENT INTENTIONS FOR FIRST ENLISTMENT AND TOTAL SAMPLE

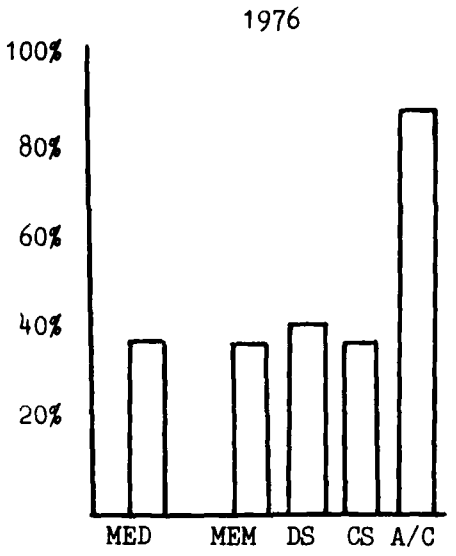


FIG 2

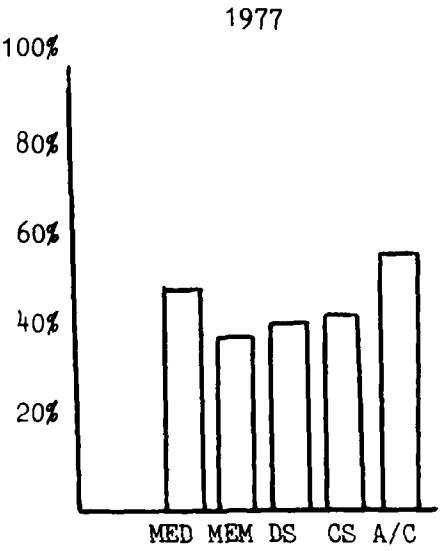


FIG 3

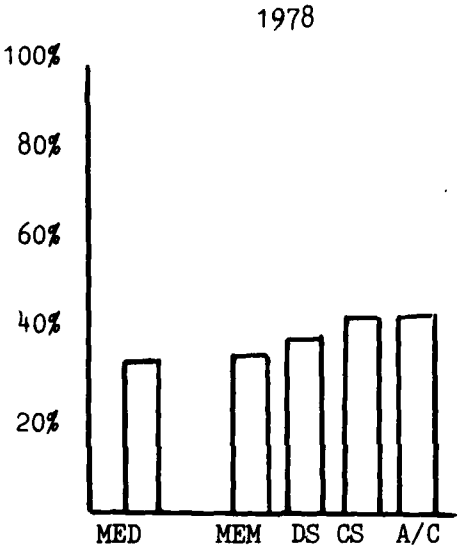


FIG 4

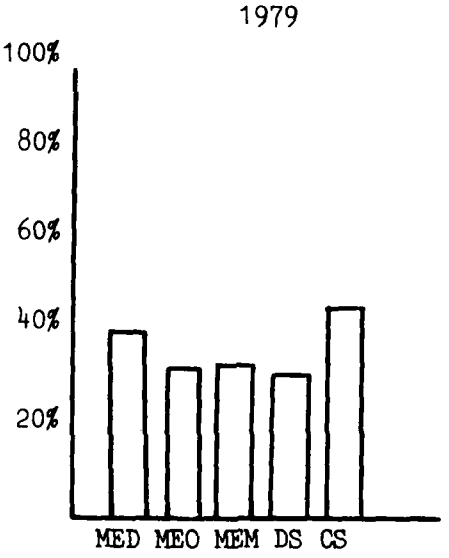


FIG 5

FIGURES 2-5. FIRST TERM REENLISTMENT INTENT BY MAJOR OCCUPATIONAL GROUPINGS. MED=MEDICAL: MEO=MISSION EQUIPMENT OPERATIONS: MEM=MISSION EQUIPMENT MAINTENANCE: DS=DIRECT SUPPORT: CS=COMMAND SUPPORT: A/C=AIRCREW

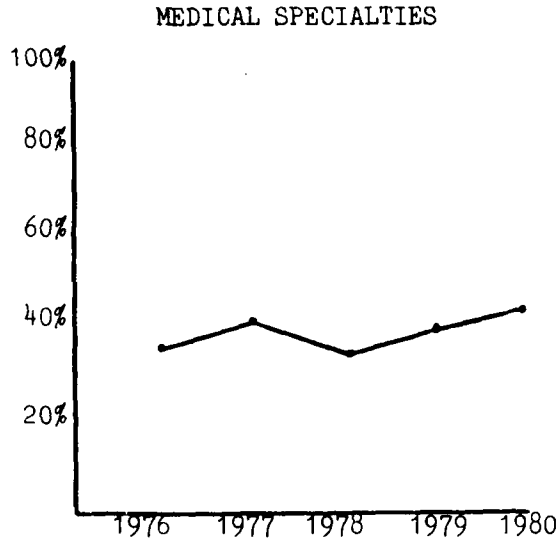


FIG 6

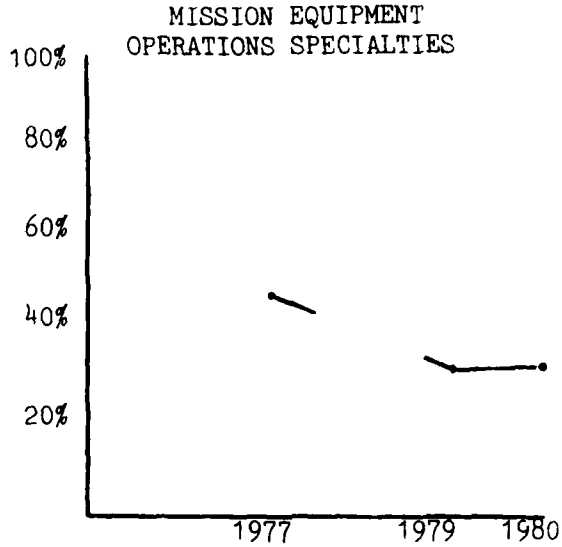


FIG 7

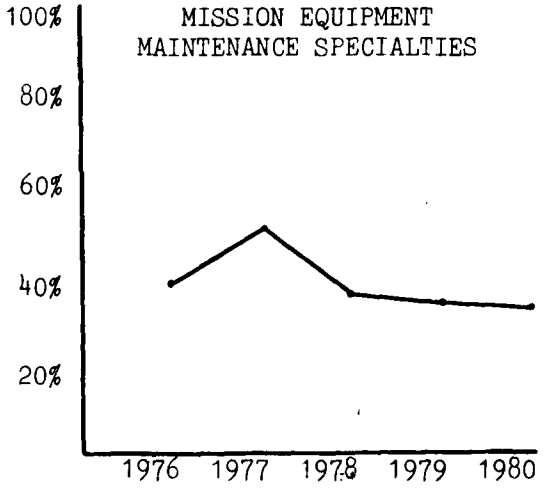


FIG 8

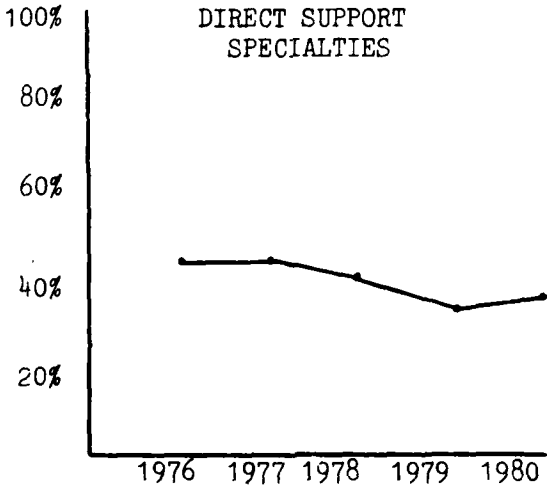


FIG 9

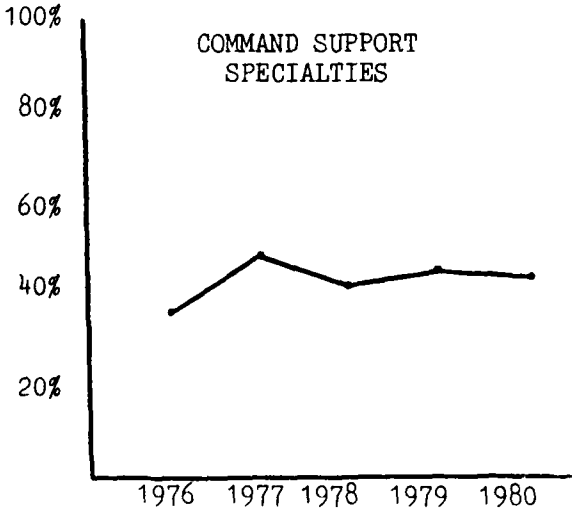


FIG 10

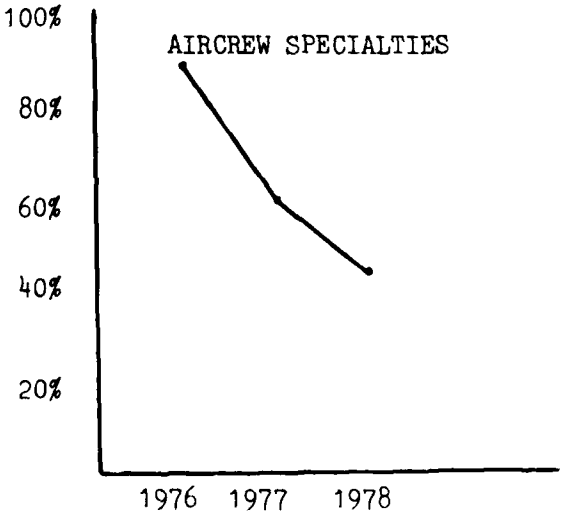


FIG 11

FIGURES 6-11. FIRST TERM REENLISTMENT INTENTIONS ACROSS TIME FOR EACH MAJOR OCCUPATIONAL GROUP. NOTE, DURING SOME YEARS, STUDIES WERE NOT CONDUCTED IN SOME OCCUPATIONAL AREAS

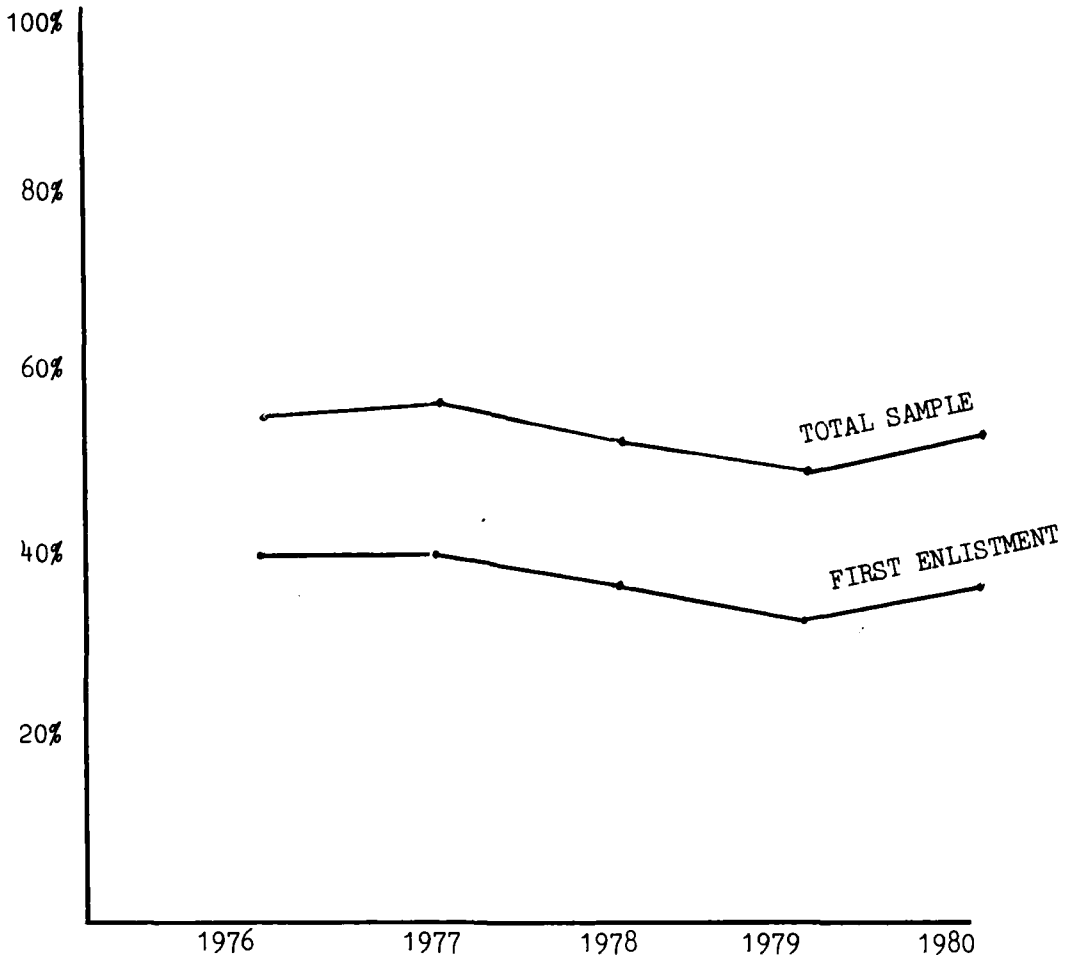


FIGURE 12. OVERALL TREND IN REENLISTMENT INTENTIONS FOR FIRST ENLISTMENT AND TOTAL SAMPLE, INCLUDING 1980

TABLE 1

SAMPLE SIZES FOR USAF OCCUPATIONAL SURVEYS

	TOTAL NUMBER PEOPLE SURVEYED	TOTAL NUMBER FIRST ENLISTMENT PERSONNEL SURVEYED
1976	28,103	11,820
1977	36,769	15,345
1978	31,534	13,580
1979	32,783	13,362
1980	23,592	8,760
	152,781	62,867

* Not all studies were completed for 1980.

(678)

MOTOWIDLO, Stephan J., State University of New York, Binghamton, New York, and LAWTON, George W., U.S. Army Research Institute, Alexandria, Virginia, and DUNNETTE, Marvin D., Personnel Decisions Research Institute, Minneapolis, Minnesota.

FACTORS IN REENLISTMENT DECISIONS OF FIRST TERM ENLISTED PERSONNEL
(Thu A.M.)

To assess the importance of factors entering into reenlistment decisions by first term enlisted personnel, Personnel Decisions Research Institute, and the Army Research Institute, developed the Army Reenlistment Opinion Questionnaire. This paper describes the development of this questionnaire, and some of the results which have been obtained with it. Cross sectional analyses of data from 4300 enlisted soldiers in the US and Europe showed that positive reenlistment intentions were dramatically lower for soldiers assigned to their first permanent duty station than for those still in training. Soldiers decided against an Army career very early in their first tour of duty. Large monetary incentives, and combinations of monetary incentives with choice of location or duty assignment, would make a substantial difference in their reenlistment intentions. Responses to questions about Army work indicated feelings of boredom, lack of challenge, and meaningless work were significant correlates of negative reenlistment intentions and discriminated between combat and noncombat soldiers. However, the hardship of combat training itself did not show up as a significant negative influence. Changes which give soldiers more control over their personal lives, improved housing, on-post facilities, assignment practices, and career counseling should improve retention of first term soldiers.

REENLISTMENT FACTORS FOR FIRST TERM ENLISTED PERSONNEL

Stephan J. Motowidlo
Marvin D. Dunnette
Personnel Decisions Research Institute
Minneapolis, minnesota

George W. Lawton
U. S. Army Research Institute
Alexandria, Virginia

ARI has been conducting research in support of the Army's reenlistment program. A substantial portion of this research was conducted by Personnel Decisions Research Institute, Minneapolis, Minnesota, under contract to ARI.[1]

ARI's approach to reenlistment research was to identify variables that determine reenlistment and which have direct implication for Army management. We (ARI and PDRI) decided to consider reenlistment in a framework provided by current theories of work motivation and career decision making. On the basis of a review of the literature on military retention and civilian personnel turn-over, interviews with soldiers considering reenlistment, and discussions with Army reenlistment policy makers, a number of variables which were potential determinants of the reenlistment decision were identified. These variables are shown in Figure 1, arranged according to a general expectancy-value model.

Expectancy-value models and decision theory have proven useful in understanding both work motivation (3) and personnel turnover (2). The conception was that soldiers compare the overall attraction of civilian life and work with that of an Army career. The overall attraction of the Army is determined by the soldiers expectations about Army life combined with the values of the things the Army is expected to provide. The overall attraction of civilian life and work are similarly determined by expectations about those things that civilian life can provide. Precisely how these variables are combined in decision making remains an object of continuing research.

Certain demographic/personal variables, which have been shown to play a very important role in predicting reenlistment, were not included in our basic attempts to understand reenlistment because of their limited management utility. And the model of the decision process outlined above was built for soldiers in general, rather than for each specific MOS. We hypothesize that the decision process is the same for each soldier. Soldiers in different MOSs simply have different values and different expectations about their Army careers.

This paper was prepared for presentation to the Military Testing Association, Toronto, Canada, 1980. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the United States Army, or of the U. S. Department of Defense.

[1]Data described in this paper were gathered under contract DAHC 19-78-0020 to Personnel Decisions Research Institute, with M. D. Dunnette as Principal Investigator, and S. J. Motowidlo as Project Director.

Based on this conception of the reenlistment process, the Army Reenlistment Opinion Questionnaire (AROQ) was designed to measure the variables shown in Figure 1, through self report. This questionnaire was constructed to test a number of specific hypotheses about reenlistment, to provide information about the reenlistment decision process outlined above, and to evaluate current and proposed reenlistment management policies. The AROQ was constructed to include the following scales (4):

1. Strength of positive preservice expectations about the Army;
2. Strength of current positive perceptions about the Army;
3. Confirmations - the extent to which expectations and perceptions about the Army are the same;
4. Reenlistment expectations;
5. Civilian expectations;
6. Overall satisfaction;
7. Job Diagnostic Survey Scales - these are described below;
8. Social desirability and consistency scales.

Scales 4 and 5 measure the subjective probability of obtaining certain valued outcomes in the Army, and in civilian life, respectively.

These scales were developed to permit examination of the importance of realistic expectations in subsequent satisfaction, the relative importance of Army and civilian expectations in reenlistment decisions, and to control for social desirability and careless responding. Other scales are largely self explanatory.

Subsets of items were also based on the variables currently included in the Army Reenlistment Program. The Army's reenlistment program is controlled by Army Regulation 601-280, "Personnel Procurement: Army Reenlistment Program." This regulation prescribes roles for the personnel involved in management of Army reenlistment, details the qualifications for reenlistment, and lists the reenlistment options for which soldiers may be eligible.

In the section on reenlistment personnel, roles are most carefully spelled out for reenlistment noncommissioned officers (NCOs) and unit commanders. Unit commanders are in charge of their unit's reenlistment program and have responsibility for performing reenlistment interviews. Reenlistment NCOs are specially trained to perform career counseling, and to sell soldiers on reenlistment. Reenlistment qualifications determine who is and is not eligible to reenlist. Reenlistment options are combinations of retraining opportunities, choice of location, unit, and job assignment. In addition to these options, soldiers in certain MOSs are eligible for monetary incentives. The amount of these incentives depends on the soldier's MOS, and the amount of time for which he or she reenlists.

In order to examine the influences of unit commanders and reenlistment NCOs, they were included in a list of potential social influences. For each category of person on the list, soldiers were asked which ones they discussed reenlistment with, and how the discussion made them feel about reenlisting. In order to assess the impact of reenlistment options, soldiers were asked for which options they were currently eligible, and how they would feel about reenlisting if they could get other options.

Before discussing the results, we would like to insert the following caveat. The results are strictly correlational, and are based almost solely on self report data. We regard these data primarily as a source of hypotheses about the causes of reenlistment, some of which will be tested experimentally in other research.

RESULTS

In Figure 2 are shown the results of cross sectional analyses of reenlistment intentions and other variables for groups of soldiers at different points in their Army career. New recruits and soldiers in training appear to have an open mind concerning reenlistment. Early in their first tour, however, these intentions change dramatically. The group of soldiers who are in their fourth through their seventh month of service show substantially different intentions from those soldiers early in their careers. The difference is in the reversal of the order of "no" and "don't know" responses. There is a substantial increase in the number of people at this point who say they do not intend to reenlist, and a decrease in the number of people who say they "don't know".

These rather abrupt changes occur at the point in the first tour at which many of the soldiers are arriving at their first permanent duty station. Our hypothesis is that the change from the regimental organization, and full schedule, in basic and advanced training, to the reality of the soldier's first permanent duty assignment, results in substantial change in expectations concerning the Army career. If so, judicious changes in appropriate military experience variables could counter the downtrend and increase reenlistment intent.

We constructed scales to measure the strength of preservice positive expectations about the Army, and the strength of current positive perceptions of the Army. The differences between scores on these two scales provided a measure of the extent to which these expectations and perceptions agree. This is shown in Figure 3 as a confirmation score. As shown in Figure 3, the extent to which preservice expectations agree with present perceptions decreases with time in service. This is due largely to a decline in the strength of soldiers' present positive perceptions about the Army. The longer the soldier stays in the Army, the less he or she believes positive things about the Army. The strength of of positive preservice expectations about the Army, as recalled by these soldiers, changes very little with time in service. Again, the most dramatic changes occur in the group of soldiers who have been in the service between four and seven months.

We also measured soldier's expectations about receiving given valued outcomes after reenlisting, and after returning to civilian life. From Figure 4 one can see that expectations about civilian career change little at all, while expectations about the Army change substantially. These changes in reenlistment intentions are also paralleled by a growing tendency to express general dissatisfaction with the Army, changes in the specific tendencies to express boredom with the Army, and to say that the Army job is not challenging. These two factors: boredom and lack of challenge, stand out as strong correlates of negative reenlistment intentions.

The Job Diagnostic Survey, or JDS (1) was also included in our questionnaire to measure soldier's attitudes about their jobs. The JDS is a standardized instrument which has been used in a variety of work environments. The scales of the JDS are: skill variety, task identity, task significance, autonomy, feedback from the job, feedback from agents, and dealing with others. While our sample did not allow thorough analysis of reenlistment intent by MOS, many comparisons of clusters of MOSs showed that men in combat MOSs are more likely to have negative perceptions about their job than noncombat soldiers. When scores on the JDS scales are ranked by MOS, the two highest scoring MOSs on every scale were noncombat, although not the same MOS on every scale. Soldiers in combat MOSs also express more negative reenlistment intentions, are less satisfied with specified aspects of Army life, and less satisfied with the Army in general. However, the hardship of combat training itself did not show up as a significant negative influence.

In responding to questions about social influences in the reenlistment process, soldiers identified the people or types of people with whom they had discussions concerning reenlistment. Soldiers were also asked how the discussion with each person made them feel about reenlistment. The fact that a soldier did or did not have reenlistment discussions with a person was not strongly related to reenlistment intention. There were moderate positive correlations between how the discussion made the soldier feel and their reenlistment intentions. The highest correlations were for spouse, soldiers who have reenlisted, commanding officers, and reenlistment career counselors.

Soldiers' beliefs about their eligibility for reenlistment incentives, like monetary bonuses, retraining, choice of MOS, and choice of unit, were not related to their reenlistment intentions. When asked how they would feel about reenlisting for a variety of reenlistment incentives, soldiers who had said they would not reenlist, said they would be "more likely" or "much more likely" to reenlist for the reenlistment outcomes shown in Figure 5.

SUMMARY AND CONCLUSIONS

One interpretation of these data is that the majority of soldiers who leave the Army at the end of their first tour, do so because they are dissatisfied with the Army rather than because they have a better civilian opportunity. There are probably two reasons for this. In some cases, there is simply a mismatch between the person and the organization. Some people are not cut out to be soldiers, or at least not to make a career out of soldiering. But in other cases, good potential career soldiers are leaving the Army because of their dissatisfaction.

Our results indicate that the most attractive reenlistment incentive would be a large monetary bonus. While it appears to be true that many soldiers reenlist for a bonus, and that even more would reenlist if bonuses were larger or more universally available, it also appears that simply increasing bonuses, without changing other aspects of Army work and Army life which may be causes of or related to dissatisfaction with the Army, would have the effect of getting soldiers to reenlist while they remain dissatisfied with the Army.

The strongest correlates of negative reenlistment intentions in our data, were reports that the Army job is boring, that Army life is boring, and that Army work is not challenging. The Army Research Institute is currently beginning research into the reaons for these reports, and into possible remedies for the problem. We suspect that some changes in the way soldiers' time is organized, and in the way they are utilized, may make their lives less boring, and their jobs more challenging and interesting.

REFERENCES

1. Hackman, J. R. & Oldham, G. R. The Job Diagnostic Survey: An Instrument for the diagnosis of jobs and the evaluation of job redesign projects. Technical Report No. 4, Department of Administrative Sciences, Yale University, May, 1974.

2. Hom, P. W., Katerberg, R., & Hulin, C. L. Comparative examination of three approaches to the prediction of turnover. Journal of Applied Psychology, 1979, 64, 280-290.

3. Lawler, E. E. Motivation in work organizations. Belmont, CA: Brooks/ Cole, 1973.

4. Motowidlo, S. J., Dunnette, M. D., & Rosse, R. L. Reenlistment motivations of first term enlisted men and women. Technical Report, Personnel Decisions Research Institute, Minneapolis, MN, February, 1980.

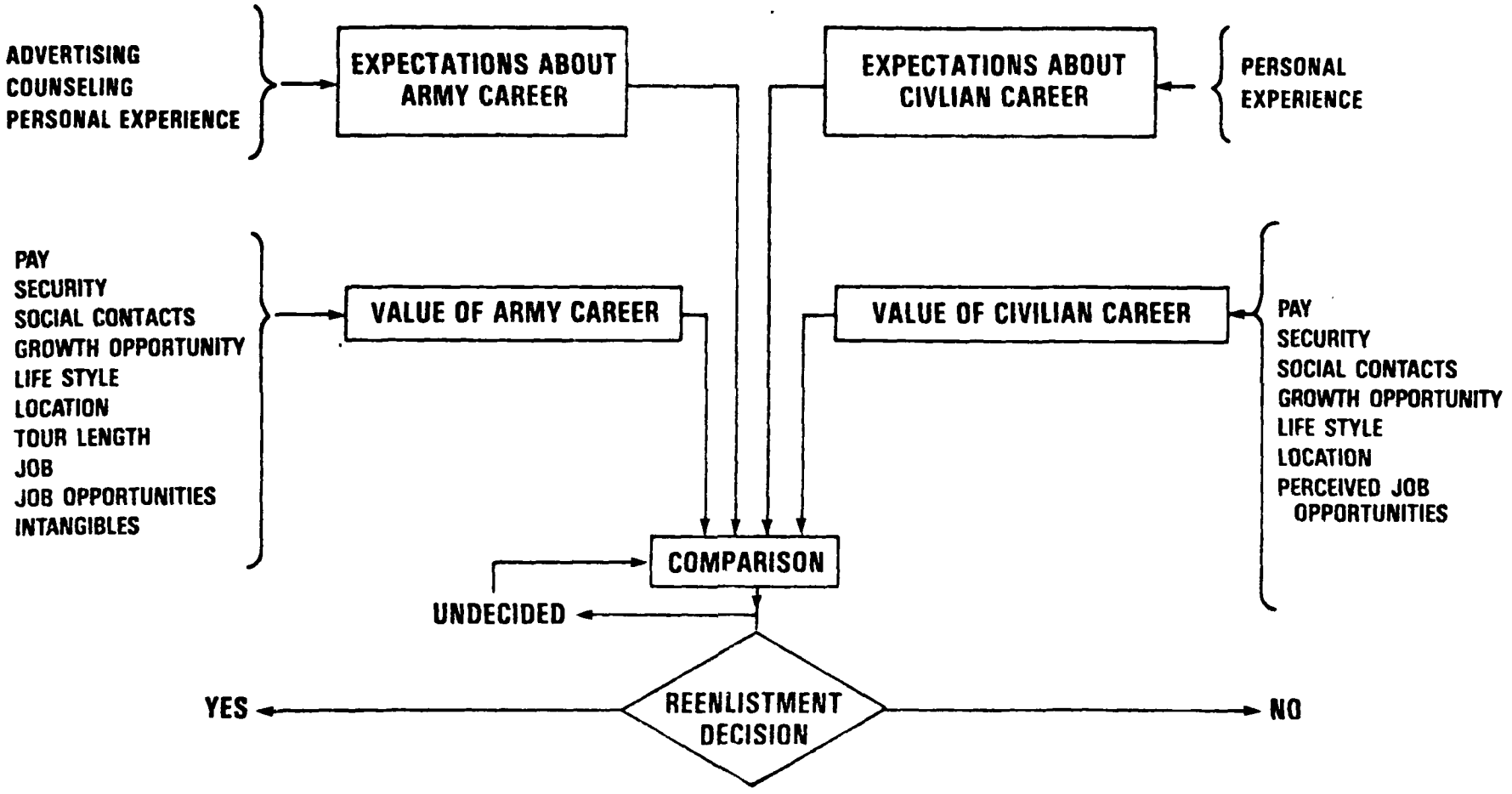


Figure 1. Expectancy--value framework for the reenlistment decision process

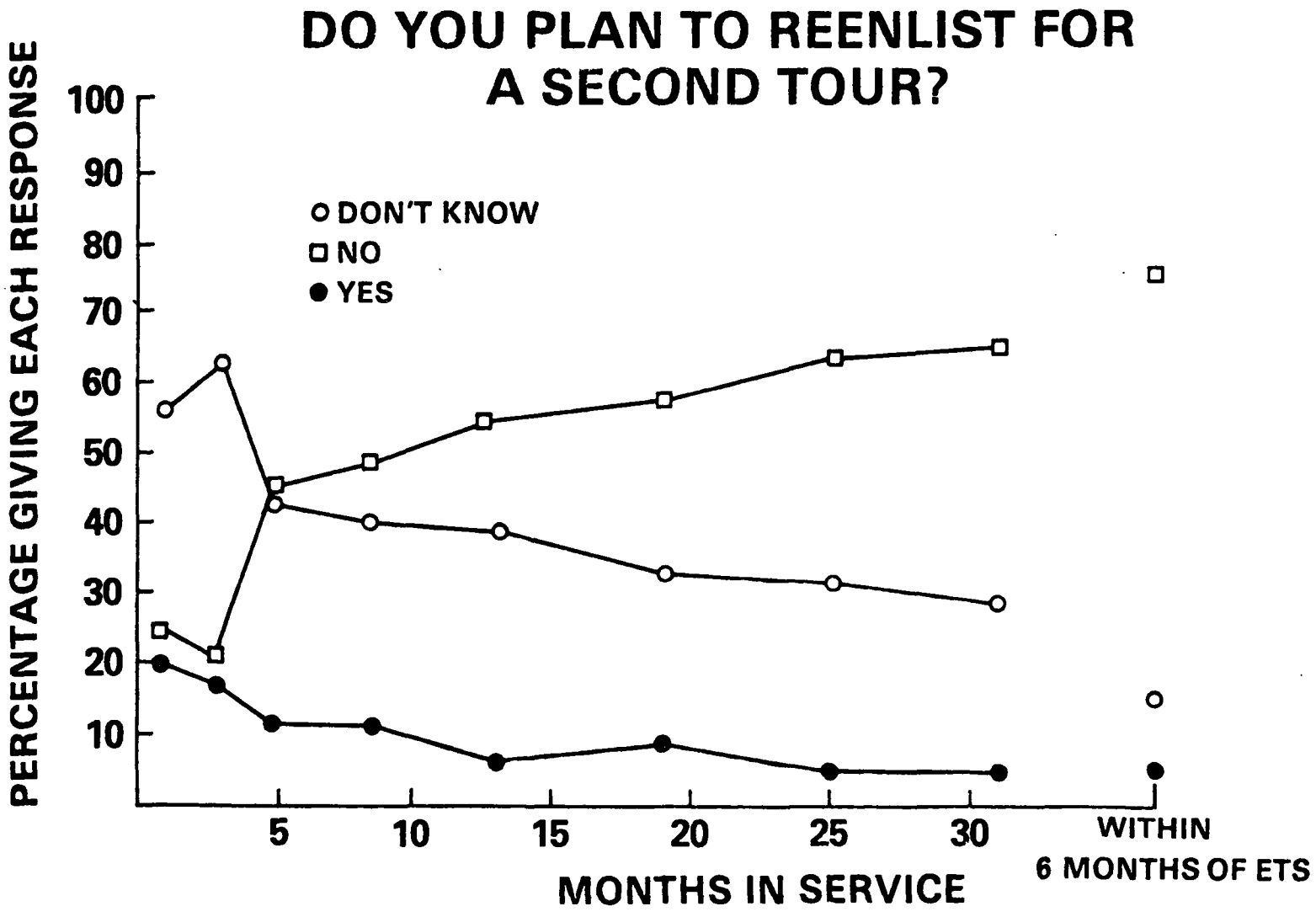


Figure 2. Cross-sectional analysis of reenlistment intention by time in service.

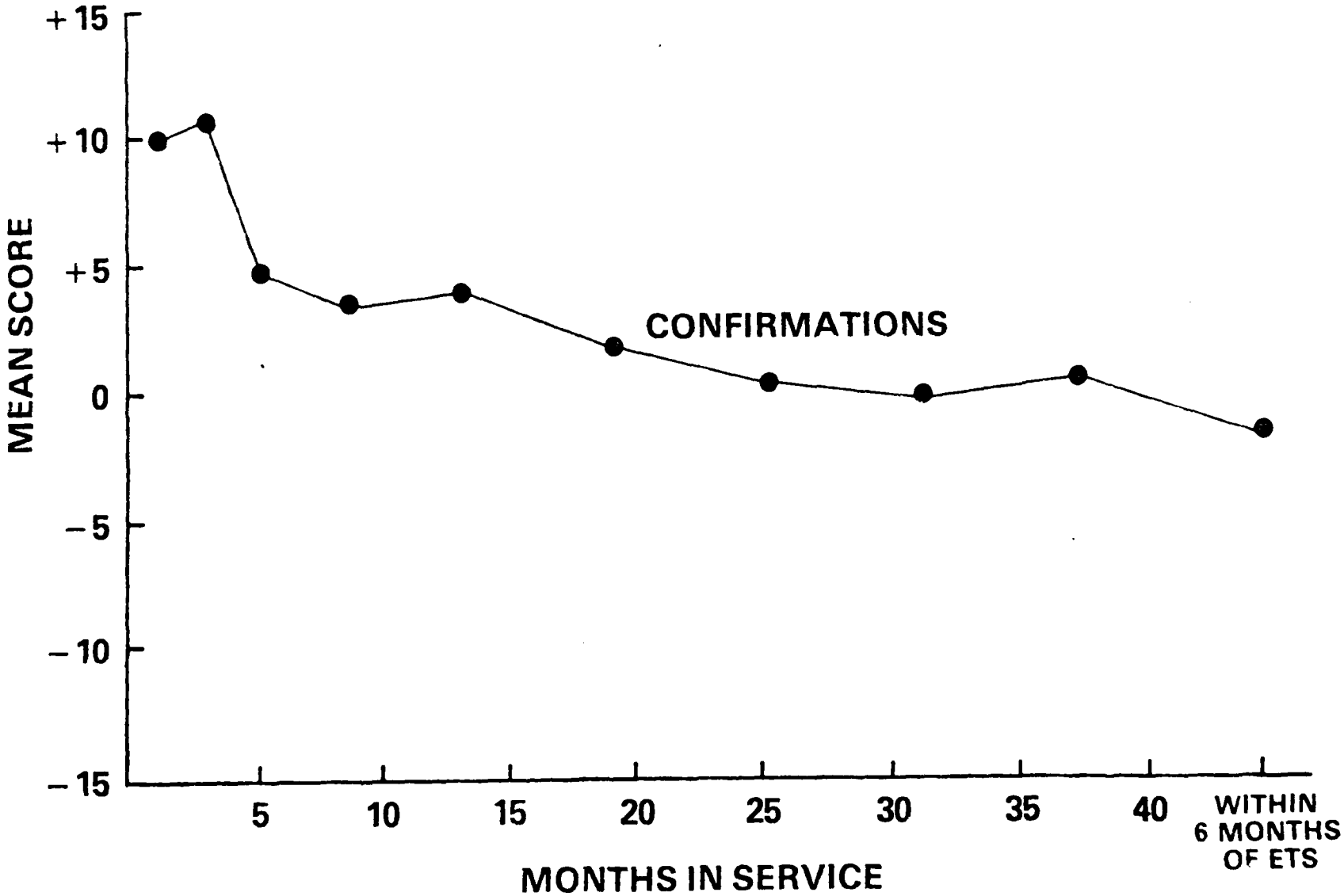


Figure 3. Cross sectional analysis of the extent to which preservice expectations agree with current beliefs about the Army as a function of time in service.

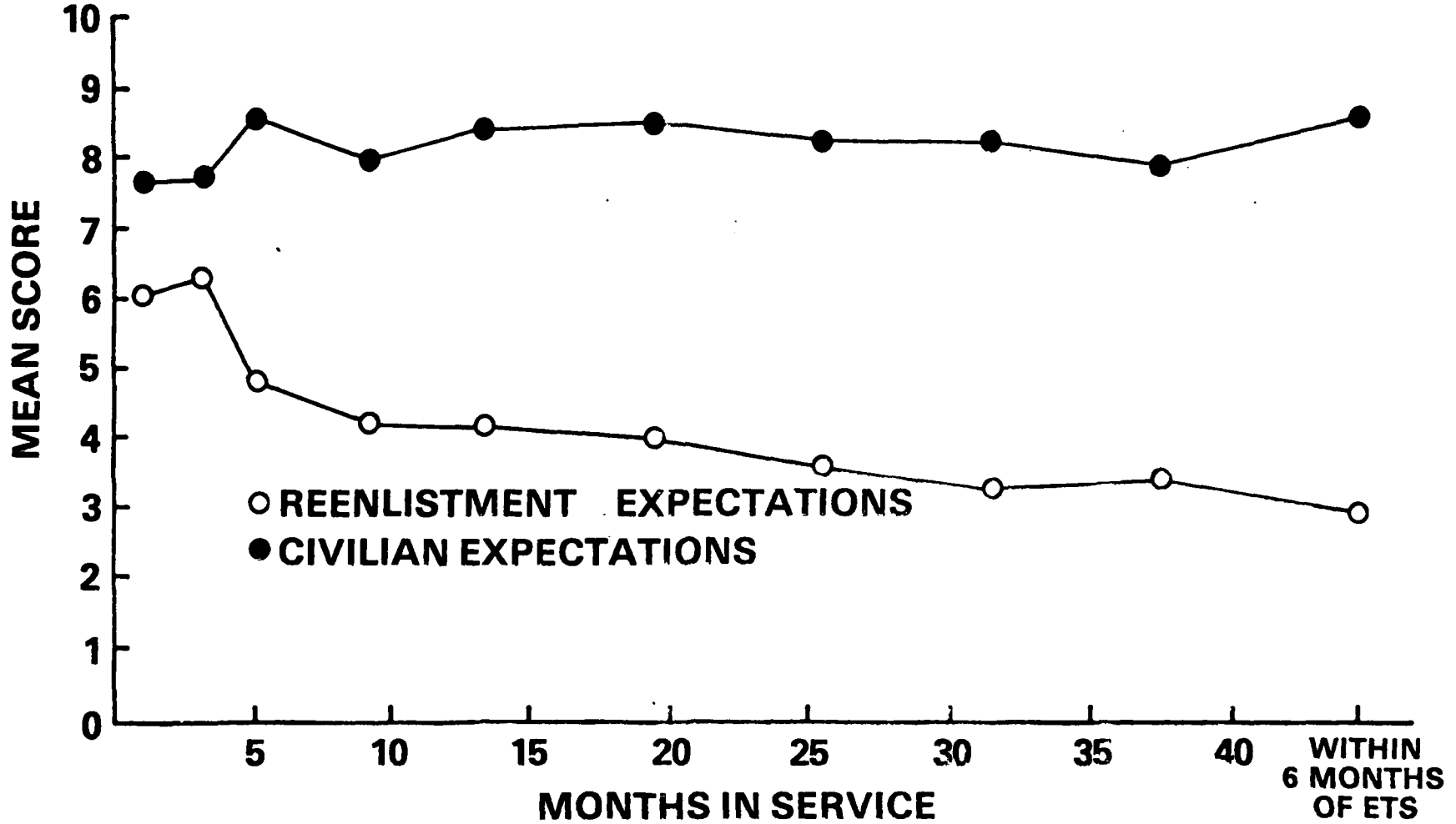


Figure 4. Cross sectional analysis of Reenlistment and Civilian Expectations according to time in service.

INCENTIVE	PERCENT SAYING THE INCENTIVE WOULD MAKE THEM MORE LIKELY TO REENLIST FOR 3 YEARS
Bonus of $9,000 and choice of location	73
Bonus of $9,000	72
Bonus of $9,000 and choice of MOS	71
Bonus of $6,000	64
Time off to attend civilian classes	61
Afford to live in own private residence	60
Stay in location of choice for entire reenlistment period	59
Choice of location and MOS for first assignment	56
Bonus of $3,000 and choice of MOS	56
Bonus of $3,000 and choice of location	54

INCENTIVE	PERCENT SAYING THE INCENTIVE WOULD MAKE THEM MORE LIKELY TO REENLIST FOR 6 YEARS
Stay in location of choice for entire reenlistment period	39
Bonus of $9,000 and choice of location	37
Bonus of $9,000 and choice of MOS	30
Work in MOS of choice for entire reenlistment period	30
Choice of location and MOS for first assignment	28
Afford to live in own private residence	28
Time off to attend civilian classes	26
Bonus of $9,000	21
Promotion points for completing civilian training courses related to the MOS	20
Assignment to an elite unit	20

Figure 5. Percent of soldiers saying they would be more likely to reenlist for 3 or 6 years for various reenlistment incentives

690

O'LEARY, Brian S., U.S. Office of Personnel Management, Washington, D.C.

COLLEGE GPA AND JOB PERFORMANCE: APPLICATION OF META-ANALYSIS
(Thu P.M.)

Last year, Hunter (1979), in an invited address to the Division of Industrial/Organizational psychology of the American Psychological Association, recommended the use of meta-analysis to integrate the results of previous research studies in industrial-organizational psychology. Glass (1977) has defined meta-analysis as the statistical analysis of a large collection of results from individual studies for the purpose of integrating the findings. Following Hunter's suggestion, this paper reports on a study that applied the meta-analysis procedure to one potentially useful selection device: college grade-point average (GPA).

Previous reviewers of the GPA-job performance relationship (Hoyt, 1965; Calhoon & Reddy, 1965) have based their conclusions on simple counts of statistically significant findings. Hunter (1979) indicated that such tabulations of significant results can lead to completely erroneous conclusions because all studies are treated equally regardless of sample size differences.

The research literature published since 1964 on the relationship of GPA to job performance was reviewed. Furthermore, personnel research units in the Army, Navy, and Air Force were contacted to obtain any additional findings relevant to the GPA-performance relationship. Nineteen studies which reported both correlation coefficients and sample sizes were located for the meta-analysis. An estimate of the population correlation was obtained by computing the weighted average of sample correlations across studies. In this procedure, each sample correlation is weighted by the sample size of the study. Thus, greater importance is placed on the results of studies which have the least sampling error (i.e., greater sample sizes). Then the variance of the distribution of sample correlations were computed and corrected for sampling errors. Finally, a confidence interval around the estimated population correlation was determined. The average correlation between GPA and job performance, weighted by sample sizes, was .17. The standard deviation of this distribution of sample correlations was .120. After correcting the standard deviation for sampling errors, the 90% confidence interval becomes + .01 to + .33. This result indicates a small positive relationship between GPA and job performance. Given the many factors which might have attenuated this relationship, both on the predictor and criterion sides of the equation, this result is not surprising. Implications of this meta-analysis approach for integrating and summarizing the results of multiple studies of any selection procedure are explored.

COLLEGE GRADE POINT AVERAGE (GPA) AND JOB PERFORMANCE:

APPLICATION OF META-ANALYSIS

Brian S. O'Leary*
U.S. Office of Personnel Management
Personnel Research and Development Center
Washington, D.C. 20415

My talk today has two major purposes: (1) to review the concept of meta-analysis and its importance for applied research, and (2) to illustrate the concept of meta-analysis with an example from our work at the Office of Personnel Management (OPM) on the search for alternative selection techniques (i.e. alternatives to written tests), namely a review of the literature relating Grade Point Average (GPA) to occupational success.

Most of what I say today about the concept of meta-analysis can be found in the work of Glass and his associates (c.f., Glass, 1976, 1978; Smith & Glass, 1977), the works of Schmidt and Hunter and their associates, (c.f., Schmidt & Hunter, 1977; Pearlman, Schmidt & Hunter, 1980; Schmidt, Hunter, & Pearlman, 1980), and the work of Rosenthal and his associates (c.f., Rosenthal, 1978, 1979; Cooper & Rosenthal, 1980). What I will attempt to do is put these three independent but complementary research efforts in some perspective.

It is a well known fact that little can be proven from a single study. As Hunter (Note 1) has indicated, the elimination of alternative hypotheses requires the work of many people in many studies. Moreover, although the vast majority of articles published in journals are reports of individual research studies, Cooper and Rosenthal (1980) indicate that it is probably the literature review, which involves a synthesis of individual studies, which reaches the widest audience.

Even though the literature review has such great impact, it is interesting to note that the methodology used in literature reviews is so much less rigorous than that required for primary data handling. As Glass (1976) states, the accumulated findings of dozens or even hundreds of studies should be regarded as complex data points, no more comprehensible without the full use of statistical analysis than the hundreds of data points in a single study could be so casually understood. And there is evidence that even researchers with a relatively high level of statistical sophistication cannot accurately integrate data points without the aid of statistical analyses. Bobko and Karren (1979) presented correlation scatterplots to a sample of members of Division 5 of the American Psychological Association (Measurement and Evaluation) and asked them to estimate the size of the correlation. The size of the correlation was consistently underestimated, with the most pronounced disparity in the "real world" range of observed correlations (.2 - .6). Moreover, a recent study comparing statistical vs. traditional methods of summarizing research findings found that traditional review procedures tend to underestimate the magnitude of the effect being reviewed (Cooper & Rosenthal, 1980).

Cooper (1979) has suggested that traditional literature reviews lack analytic precision in at least three ways:

*All statements expressed in this paper are those of the author and do not necessarily represent the official policies or opinions of the U.S. Office of Personnel Management.

(1) They are susceptible to the idiosyncracies of a particular reviewer's perspective.

(2) They usually ignore the issue of relationship strength by failing to assess the size of the effect under study.

(3) The typical reviewer imprecisely weights conclusions with respect to the volume of available evidence.

Meta-analysis is one answer to this lack of rigor in literature reviews. Glass (1976) has categorized data analysis into three separate activities--primary, secondary, and meta-analysis:

1. Primary Analysis - data of a given study are analyzed for the first time.

2. Secondary Analysis - data of a given study are reanalyzed (i.e., new questions are asked with an old data set such as Project Talent data).

3. Meta-analysis - results of independent studies are combined for the purpose of integrating the findings.

Meta-analytic studies are now becoming quite popular. Perhaps the most quoted meta-analytic study is Glass and Smith's (1976) integration of the literature concerning psychotherapy outcome. Based on approximately 375 studies, they found that psychotherapy is effective. A typical client can be expected to move from the 50th to the 75th percentile of the untreated population. Moreover, therapy of any type is effective (e.g., behavioral therapies exhibit only a trivial advantage over nonbehavioral therapies).

The measure which Glass and his associates use to aggregate findings across studies is an "effect size" defined as the difference between the means of the psychotherapy and control group, divided by the control group standard deviation (Glass & Smith, 1976). Cumulating the individual study "effect size" across studies obtains an average effect size. For example, in their study on psychotherapy outcome, the average effect size was .68. This says, in effect, that the average person receiving some form of psychotherapy was about 2/3 of a standard deviation more improved on an outcome measure than the average control group member.

Similar studies have been conducted by Glass and his associates on the relation between class size and school achievement (Glass & Smith, 1979), and the relationship between socioeconomic status and I.Q. (White, Note 2).

Rosenthal (1978) working independently of Glass and his associates has also been conducting meta-analytic studies. Rosenthal has focused on the methods used for combining the probabilities obtained from independent studies. That is, how does one obtain a "combined p" level for all studies reviewed? Rosenthal notes, however, that even if a "combined p" level is determined, this tells nothing about the typical size of the

effect. Thus, Rosenthal also recommends computing a "size effect" statistic.

Rosenthal and his associates have used meta-analytic procedures to review the literature on the effects of the interpersonal self-fulfilling prophecy (Rosenthal & Rubin, 1978) as well as the research on the issue of whether females conform more than males (Cooper, 1979).

One of the most intriguing aspects of Rosenthal's work is what he has termed the "file drawer problem" (Rosenthal, 1979). No literature review will uncover every relevant study addressing a particular issue. One cannot tell how many studies have been conducted but never reported. As Rosenthal (1979) indicates, the extreme view of this "file drawer" problem is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show nonsignificant results.

Rosenthal's procedure answers the question of how many new, filed or unretrieved studies with essentially null results would be required before the results could be attributed to sampling bias. For a small number of studies which are not very significant, even when their combined "p" is significant, it takes only a few nonsignificant studies "filed away" to change a combined significant result to a nonsignificant one. On the other hand, when the number of studies is large, or the effect size is large, it takes a very large number of unretrieved or unpublished studies with essentially null results to bring the combined "p" to a nonsignificant level.

Quite independent of Glass and his associates and Rosenthal and his associates, Schmidt and Hunter (1977) have been using meta-analytic techniques in their work on validity generalization.

Although Schmidt and Hunter's work is a form of meta-analysis, their work has had a somewhat different focus than that of either Glass or Rosenthal. Glass and Rosenthal have focused mainly on the mean of statistics across studies (e.g., mean effect size). Schmidt and Hunter, in their validity generalization work, have focused on the variance of results across studies. Their concern has been with the false variation produced by artifacts of research design such as small sample sizes.

Ghiselli (1966) found that observed validity coefficients vary considerably from study to study even when jobs and tests appear to be quite similar (Ghiselli, 1966). This led to the belief that empirical validation is required in each situation (i.e., situational specificity) and that validity generalization was impossible. Schmidt and Hunter and their associates, through the use of meta-analysis, have demonstrated that much of this variation in test validities across studies is due to statistical artifacts.

Conceptually, their method is really quite simple. One first gathers a data-base of validity coefficients for a given test-job combination. The variance of this distribution of validity coefficients is then computed. Variance due to various sources of error (e.g., sampling error, differences between studies in test and criterion reliability) is then subtracted from the total variance. Schmidt and his associates have found that much of the variation in validity coefficients that is observed from study to study is due to these artifacts. In fact, sampling error due to small sample sizes accounts for approximately one-half the variability in validities across studies.

The mean validity coefficient obtained by Schmidt and his associates is analogous to Glass' mean size effect. However, Schmidt and his associates correct the average observed validity for attenuation due to criterion unreliability and range restriction. Using the standard deviation of the validity distribution corrected for statistical artifacts, they establish a 90% confidence interval. Using this confidence interval one can be 90% confident that the true validity would be at or above this mean level in a new situation involving the same test-type and job without carrying out a validity study of any kind.

These meta-analytic procedures have wide applicability in applied research. In addition to the application of meta-analysis in validity generalization, we are now using the procedure in our work on the search for alternatives to written tests.

With the increased commitment of this administration to obtaining a work force that is representative of society, as well as the Uniform Guidelines requirement to make a reasonable search for alternatives to selection devices which have substantial adverse impact (U.S. Equal Employment Opportunity Commission, 1978), a Task Force was established within the Personnel Research and Development Center at OPM with the specific objective of developing alternatives. One project initiated under this Task Force was a review of the literature relating college grade point average (GPA) to job success (O'Leary, Note 3).

Evidence of scholastic achievement is currently used by many organizations to select employees. In the federal government, an outstanding scholar provision, by which college graduates with outstanding academic achievement records could be hired without taking a written ability test competitively, was instituted in the mid-sixties as part of the entry level selection process for administrative and professional occupations.

Concern about the job-relatedness and validity of the program led to a modification of the program. At present, all applicants must take the written test. Outstanding academic achievement is now rewarded in that it is averaged with an applicant's written test score to obtain a final rating for the applicant. Since the original outstanding scholar

provision had been a major avenue of entry into the federal service for minority group members, a decision was made to reexamine the validity evidence concerning GPA.

Previous reviewers of the relationship between GPA and occupational success based their conclusions on simple counts of statistically significant findings. Last year, Hunter (Note 1), in a invited address to Division 14, indicated that such tabulations of significant results can be quite misleading because all studies are treated alike regardless of differences in sample size. Following Hunter's (Note 1) recommendation, a meta-analytic procedure was used to integrate the results of previous research relating GPA to occupational success.

A literature review, consisting of a search of the Psychological Abstracts for the years 1964 through June 1979, was conducted. In addition, the personnel research units of the Army, Navy, and Air Force were contacted concerning the relationship between GPA and officer performance. Twenty-eight studies were found which related academic achievement to occupational success. Only studies in which the primary source could be obtained were included in the review. Six of these 28 studies were concerned with grades in graduate programs, such as these leading to a Master of Business Administration degree, and three were concerned with community college or nursing school grades. Thus, only 19 studies actually dealt with undergraduate grades.

The fact that this investigation focused only on the empirical relationship between GPA and occupational success should not be construed as an indication that issues in the measurement of scholastic achievement (e.g., different meaning of grades from different institutions, and even across departments within the same institution, grade inflation, etc.) are no longer of concern. Such factors tend to mask true differences in academic proficiency and limit the usefulness of grades as predictors of job performance. The question addressed in this study was very pragmatic: That is, despite these measurement problems, is there still a relationship between GPA and job success?

In combining results across studies, a weighted average of sample correlations was used. If we assume that the correlation between GPA and job performance is the same across studies, except for errors, then the best estimate of the correlation is not its simple mean across studies, but the weighted average where each correlation is weighted by the number of persons in the study. In using this weighting procedure, greater importance is placed on the results of studies which have the least sampling error (i.e., those studies with greater sample sizes). It is not unreasonable to assume that the validity of GPA should be similar across occupations studied. Recent studies seem to indicate that task differences have little effect on validity. For example, Schmidt, Hunter, and Pearlman (in press) recently found that the moderating effects of tasks on aptitude test validity is negligible even when jobs differ grossly in task makeup.

In calculating the weighted average, if grades in individual courses were presented as well as overall GPA, only the overall GPA was used in determining the average correlation. In studies where several independent measures of job success were reported (e.g. performance ratings and salary level), each was considered to be an independent study.

The average correlation across all studies located relating GPA to job performance was .17 with a standard deviation of .12. Thirty-one correlations entered into this average with a total sample size of 6782.

In their work on validity generalization, Schmidt and Hunter (1980) have demonstrated that much of the variation in employment test validity coefficients across studies is due to statistical artifacts. In fact, they have found that about 70% of the variation in correlations across studies can be accounted for by four artifacts: Differences between studies in criterion reliability, differences between studies in test reliability, differences between studies in range restriction and sampling error (i.e., variation due to $N < \infty$).

By correcting the variance in the present study for only one artifact, sampling error, the standard deviation becomes .10. The 90% confidence interval for the uncorrected correlations is +.01 to +.33. Thus, one can be 90% confident that the population correlation between GPA and job performance is greater than zero.

The average correlation of .17 is, of course, an attenuated validity coefficient. As the new Division 14 Principles (American Psychological Association, Division of Industrial-Organizational Psychology, Note 4) indicate, the adjusted coefficient is generally the best point estimate one can make of the relationship. Information on range restriction and criterion reliability was not reported in most of the studies of GPA reviewed. If we use the estimates of these values that Schmidt and his colleagues use (Pearlman et al, 1980) in their validity generalization work, the validity of GPA becomes r = .22 corrected for unreliability and r = .38 corrected for both criterion unreliability and range restriction.1

A comment should be made on how applicable these estimates of criterion unreliability and range restriction are to the present study. Obtaining such estimates is somewhat more difficult when studying the predictability of GPA as compared to a traditional test validation study. In the typical test validation study one has a common test and in most cases the same criterion for all research participants. This is not the case with studies of GPA. In fact, studies of GPA can be broken down into two major categories: (1) studies of an entire graduating class and (2) studies of all college graduates employed by a single organization. In the first case, where an entire graduating class is studied, there is no restriction

---

1The expected values used by Schmidt and his colleagues are .60 for criterion reliability and a restricted standard deviation of 6.0 (from 10.0).

in range on the predictor, but since these individuals are employed in different organizations and different occupations, the criterion reliability is probably less than in the typical test validation study. In the second case, where a given organization studies all college graduates employed, there is probably the normal amount of restriction in range and criterion unreliability. Thus, the Schmidt, et al. estimate of criterion unreliabiity is probably an underestimate in this case, while the range restriction estimate is probably somewhat of an overestimate.

Five studies were located which were most closely associated with training success rather than job performance. The average correlation for these five studies was .20 with a standard deviation of .25. Correcting for sampling error, the confidence interval becomes .-20 to +.60. Only seven correlations entered into this average with a total sample size of 1043.

While there is no theoretical basis for imposing a lower limit on the number of coefficients required for such an analysis from a Bayesian viewpoint, Pearlman, Schmidt, and Hunter (1980) have limited their analyses to distributions with at least 10 coefficients to partially control for the accuracy of the priors. Given this decision rule and the fact that the obtained results with training criteria are somewhat at variance with the commonly observed findings that validities for training criteria are more than .10 correlation points higher than for job proficiency criteria (c.f. Ghiselli, 1966; Pearlman, Schmidt, and Hunter, 1980) any conclusions based on training criteria would seem premature until more data are available for analysis.

In summary, the results of this study indicate there is a small relationship between GPA and job performance. This finding is not surprising given the many factors in the measurement of GPA which tend to mask true differences in academic proficiency.

The data base upon which these conclusions were based is relatively small. We plan to update and expand the data base as more information becomes available. We are also using a similar paradigm to cumulate information on the validity and adverse impact of other alternative selection techniques. We would welcome receiving any validity data you might have on alternative techniques for inclusion in our data base.

REFERENCE NOTES

1. Hunter, J. E. Cumulating results across studies: correction for sampling error, a proposed moratorium on the significance test, and a critique of current multivariate reporting practice. Unpublished paper based on an invited address to the American Psychological Association, September, 1979. Department of Psychology, Michigan State University, December 10, 1979.

2. White, K. R. The relationship between socio-economic status and academic achievement. Ph.D. Dissertation, University of Colorado, 1976.

3. O'Leary, B. S. College grade point average as an indicator of occupational success: An update (PRR-80-23). Washington, D.C.: Office of Personnel Management, Personnel Research and Development Center, August 1980.

4. American Psychological Association, Division of Industrial-Organizational Psychology. Principles for the validation and use of personnel selection procedures. (Second edition), Berkeley, CA: Author, 1980.

REFERENCES

Bobko, P., & Karren, R. The perception of the Pearson product-moment correlations from bivariate scatterplots. Personnel Psychology, 1979, 32, 313-325.

Cooper, H. M. Statistically combining independent studies: A meta-analysis of sex differences in conformity research. Journal of Personality and Social Psychology, 1979, 37, 131-146.

Cooper, H. M., & Rosenthal, R. Statistical versus traditional procedures for summarizing research findings. Psychological Bulletin, 1980, 87, 442-449.

Ghiselli, E. E. The validity of occupational aptitude tests. New York: Wiley, 1966.

Glass, G. V. Primary, secondary, and meta-analysis of research. Educational Researcher, 1976, 5, 3-8.

700

Glass, G. V. Integrating findings: The meta-analysis of research. In L. Shulman (Ed.), Review of Research in Education. Volume V, Itasac, IL: Peacock, 1978.

Glass, G. V., & Smith, M. L. Meta-analysis of research on the relationship of class-size and achievement. Evaluation and Policy Analysis, 1979, 1, 2-15.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 1980, 65, 373-406.

Rosenthal, R. Combining results of independent studies. Psychological Bulletin, 1978, 85, 185-193.

Rosenthal, R. The "file drawer problem" and tolerance for null results. Psychological Bulletin, 1979, 86, 638-641.

Rosenthal, R., & Rubin, D. B. Interpersonal expectancy effects: the first 345 studies. The Behavioral and Brain Sciences, 1978, 3, 377-415.

Schmidt, F. L., & Hunter, J. E. Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 1977, 62, 529-540.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. Task differences as moderator of aptitude test validity in selection: A red herring. Journal of Applied Psychology, in press.

Smith, M. L., & Glass, G. V. Meta-analysis of psychotherapy outcome studies. American Psychologist, 1977, 32, 752-760.

U.S. Equal Employment Opportunity Commission, U.S. Civil Service Commission, U.S. Department of Labor, U.S. Department of Justice. Uniform Guidelines on Employee Selection Procedures. Federal Register, 1978, 43, (166), 38290-38315.

OREND, Richard. J., Human Resources Research Organization, Mannheim, West Germany.

ADAPTATION TO USAREUR (Thu P.M.)

Substantial effort has been directed toward obtaining an understanding of factors which influence attrition and performance in the military. The product of this effort has been the identification of a large number of factors which are related to individual behavior while on active military duty. In broad groups these factors include reward mechanisms, expectations, work environment, SES and demographic characteristics, individual psychological characteristics, and social and physical environments, among others. While we think that these factors influence behavior in varying degrees, depending on the individual, we have little information about the process involved in translating a set of attitudes, expectations, desires, various individual characteristics, and environmental conditions into responsive behaviors. That is, how does one adapt in the real world.

In an attempt to address this issue, a longitudinal study of USAREUR first-term enlisted personnel was developed and fielded during the period from November 1979 through September 1980. This study tapped the attitude/expectation/characteristic set of about 590 individuals as they arrived in USAREUR and followed changes in those factors plus behavior responses to the new environment in three susbequent interviews. The paper reports on some of the results of that study. The specific focus is on changes in attitudes and mechanisms used to adapt to perceived hostile environments.

ADAPTATION TO USAREUR:

The First Six Months

Introduction

In the Fall of 1979 HumRRO, under sponsorship of the Army Research Institute, undertook an indepth study of the process by which first term enlisted personnel adapt to the USAREUR environment. This report presents interim selected results from the first six months of data collection and, more importantly, provides a brief overview of the conceptual and methodological approach being used. The data collection is currently in the fourth of six planned phases and will continue until at least May 1981. Subsequent more thorough reports will be completed in February, 1981 (on the results of the first year in USAREUR) and approximately August, 1981 (on the results of the first 18 months).

The overall goals of the research program are to provide a better understanding of the factor which contribute performance differences (including attrition, on the job performance, and reenlistment) and to identify more effective techniques for intervening to bring about performance improvements.

Objective

The specific objective is to identify any relationship between patterns of adaptation to USAREUR and various types of behavior relevant to the Army. Behaviors of particular interest are attrition, performance on the job, reenlistment, and other social behaviors like drug utilization and disruptive behavior, which may impact on the first three general behaviors.

Approach

The construct of adaptation is useful because it facilitates the inclusion of a variety of factors which have been examined individually in previous DOD and civilian literature, e.g., expectation, socialization, personal background, peer influence, attitudes toward the military, psychological facits, etc., and allows the inclusion of variables and processes that have not received a great deal of previous attention. Of particular importance in the latter area are: (1) the impact current experience; (2) information processing - including source, type, extent, and integration; and (3) problem solving style or reaction type. The specification of the above factors suggests that we view adaptation as a process in which the individual brings a set of experiences, perspectives, propensities, and cognitive processes to a situation and these factors interact with the events encountered in the new situation to produce a particular behavior pattern. In a military environment, such a pattern could include attrition before completing the initial assignment, poor performance, reenlistment, drug usage, or a variety of other disruption or supportive behaviors.
All of these behaviors represent a kind of adaptation on the part of the individual service member. Some of these adaptations may be beneficial to the military, some only to the individual, some to both and others to no one. The important point is that, as with any complex social behavior, there are likely to be a multiplicity of causal factors operating in a variety of ways to produce what social scientists often reduce to simple behavioral dimensions. Primary examples are attrition and reenlistment. Our approach has been to try to identify as many of the possible causally related factors as possible, to identify patterns or combination of behavior which typify individuals or groups of individuals using these patterns, and then to determine if these patterns are related to the criterion behaviors of interest. This approach is distinguished from trying to show the impact of each variable on some criterion or producing regression where the individual and collective impact of a string of such variables across the entire population is demonstrated.
We feel the approach will provide more useful results because (1) it permits the specification of multiple patterns which may lead to the same relevant behavior; and (2) in accomplishing (1), it makes development of appropriate pinpointed intervention strategies more feasible by identifying relevant variables and the process in which they operate. Thus, we should be able to determine not only that some SM's want more money, but when and how to provide it so that it will have the greatest positive impact on relevant behaviors.

Methodology

Sampling and Returns: To address the issue of whether such a complex approach could be effectively researched in a single project a small study was developed and applied to a sample of 611 first term enlisted Army personnel coming to

USAREUR. This sample size represented a compromise between our efforts to do an intensive study and the sponsors interest in designing a study which produced "generalizable" results. The study is longitudinal with the initial interview (self administered questionnaire) taking place the day each subject entered the division and subsequent administrators taking place after 6 weeks, about 6 months, and approximately each 4-5 months thereafter. The study is currently scheduled to last for 18 months. Data processing is complete for the first three administrations. Data were collected for the fourth in the September/ October timeframe and are currently being processed. The data collection was begun in November, 1980 using the incoming personnel of two USAREUR combat divisions.

The questionnaires are administered by Army personnel to groups of respondents at Battalion level. Each Battalion in the study has provided a project action officer who is responsible for administering the questionnaire to from 3 to 30 personnel and for distributing and collecting evaluation instruments (two for each subject) which are circulated each time the survey is given. Thus, we have responses from the individuals and their raters and endorsers covering the entire study duration. This procedure has proven very effective through the first two follow-up surveys. The initial sample included 611 respondents which represents about 95% of those asked to participate. The first follow-up had a response rate of over 90% on the survey with most misses accounted for by logistical failures rather than refunds. Over 85% were completed for the second, six month, follow-up. We feel these are excellent return rates and represent a very good return for the level of effort. Evaluation return rates lag somewhat behind, running at 75 to 80%. It seems particularly difficult to obtain endorser support which is 5 to 10 percentage points behind rater returns.

Questionnaire Development: An initial questionnaire was developed using a large number of open-ended questions and as many previously developed indexes and scales as possible. In the questionnaire we attempted to include as many of the basic issues found in previous literature to be related to relevant Army behavior as possible. We also included a number of questions on decision-making processes and previous leisure behavior patterns in order to complete the model used to develop the study. The entire package, about 1½ hours worth, was pretested on a sample of 50 SM's, as was the initial version of the first follow-up instrument. (This permitted the testing of administration procedures as well.) Results of these tests permitted the categorization of many of the open-ended questions and the entire procedure reduced the initial questionnaire to an average test time of from 45 minutes to 1 hour. The follow-up survey, which is substantially the same at each administration generally takes 45 minutes or less.

In lieu of presenting the entire questionnaire we shall briefly list the major issue areas covered. Arriving subjects were asked questions in the following areas:

1. attitudes toward being in Germany, Germany as a place and the German people;

2. attitudes toward the Army and, particularly, the work environment;

3. exrectations about living in Germany;

4. expectations about working and being in the Army;

5. leisure activity patterns before entering the Army;

6. behavior patterns vis-a-vis authority figures prior to entering the Army;

7. reasons for entering the Army and some of the informtion gathering and social influence patterns associated with the enlistment decision;

8. five personality characteristics (authoritarianism, locus of control, self esteem, anomy, and rigidity);

9. family and social relationships; and

10. sociodemographics.

Follow-up questionnaires covered the same issues with the exception of the historical and demographic information which would remain constant. The questions were rephrased to reflect current behavior. Thus, questions on leisure activities were put in the present tense and questions on expectations were worded to reflect current experience. In this way it will be possible to establish a behavioral/expectation/experience base against which change may be charted. Additional items on process and reaction were included to expand data on how individuals reacted to both adverse and positive experiences since arriving in USAREUR. This combination of historical and current experience information provides the data base necessary to analytically pursue the conceptual issues raised in the first part of this report.

Results

The initial results pertain to the first three survey administrations and focus on describing expectations about Germany and the Army, changes in behavior, the degree to which expectations are met, and some preliminary examination of the relationship between expectations and attitudes, on the one hand, and attrition, reenlistment intentions, and performance, on the other. In this paper we will focus a brief discussion on changes in experience and the degree to which expectations are met.

Table 1 as it appears in the written text of this paper shows the proportions of respondents who indicated some change in their leisure behavior patterns at six weeks and six months after arrival in USAREUR. Perhaps the most noticeable aspect of this table is the high change rate apparently occurring in all major leisure activity areas. All show at least 50% changes. Even allowing for reliability error in responses, this level of change is substantial. The level of disruption in SM's leisure lives suggested in this table is perhaps a major factor in producing adjustment difficulties. A second overall characteristic of the change pattern is that with the exception of dating, the pattern of changes is well established in six weeks and changes little from the original pattern thereafter. The anomaly on dating does not have an immediate obvious explanation since the pattern seems to show even less dating after 6 months than after six weeks. In a more substantive view, music related activities and dating are the popular activities which show the greatest decline. In both instances, this may reflect the absence of accessible facilities. Attending movies, on the other hand, shows the greatest increase, although it is not one of the most popularly pursued activities in civilian life. Again, access is probably a major factor.

Table 2 carries the analysis of leisure behavior to a more general level in talking about "styles" of leisure behavior, primarily in terms of where and with whom leisure activities were pursued. After six weeks respondents were more likely to be spending time in the barracks area, indoors, alone, and close to home. Six months in Germany brings a greater balance in the amount of time in the barracks, but even spring (the 6 month survey covered March and April as well as February) does not produce a substantial change in the shift toward indoor activities. Nor does the extended time significantly change the number of people more likely to do things alone or to having reduced travel activities.

All of these results suggest substantially altered leisure habits, whose impact will be examined in greater detail in subsequent analyses. They also suggest that the direction of that change, for a substantial number of respondents, is in the direction of less action (indoor) more isolated leisure activities. The behavioral implications for these individuals will be of particular interest in subsequent analyses.

In table 3 we look at problems SM's expected to encounter in Germany.* Again, a change from expectation to experience is likely in 50% to 60% of the respondents. After 6 weeks getting time off seemed to be the expectation with the greatest amount of error, although it was in the negative direction, i.e., for most people it was less of a problem than expected. Finding familiar things to do was the area with greatest discrepancy between expectations and experience. The unfamiliarity of the German culture does not seem to wear off for many people, although talking to Germans, the area that was initially perceived as a serious problem by most people, is relatively not as threatening once respondents have been in the country for a while. Interestingly,

---

*This list of "problems" was developed from open-ended questions used on the pretest. A three-part side was used; serious, minor and no problem.

the absolute proportions change little from six weeks to six months. Further analyses will examine those individuals who continue to find major problem areas and those who continue to identify serious problems to determine if these perceptions have implications for other behavior as well.

In light of these more specific changes in leisure behavior and expectations about Germany, we also elicited opinions about overall satisfaction with being in Germany using an expectation mode. Table 4 shows that more than half of the respondents found Germany worse than they had expected.* This was true even after only 6 weeks in the country. This suggests that many of those individuals changing leisure patterns and being (negatively) surprised by various conditions in Germany had adverse reactions which could impinge on performance.

We also examined the impact of other characteristics which have probably been found to be problem areas in USAREUR. The most prominent is race. Results presented in table 5 suggest two conditions. First, very few people anticipated service racial problems. Only 5.27% and 4.87% expected and found a serious negative effect of race on their ability to "get along" in Germany. Second, what they experienced was likely to be worse than what they expected for about 70% of the respondents. The percentages were virtually identical for 6 weeks and 6 months. They suggest that racial tensions are still high. A similar finding occurred on the question of gender as well.

The final three tables presented in this report present some preliminary results using performance (criterion) variables. In this instance, attrition and re-enlistment intention were used. Table 6 shows the relationship between expectations about serious problems and attrition. The reader should be careful in generalizing these results because of the small number of attrites (25) who had appeared in our sample as of this six-month administration. Ignoring our warning, the results depicted on this table, if they hold up in subsequent analyses, provide an interesting picture of expectation differences. Attriters were significantly more likely to see "finding interesting places to go" and "talking to the German people" as serious problems and significantly less likely to be concerned about getting time off. If there is a pattern here, it is that Germany was likely to be a bigger problem for them than the limitations of the Army.

Table 7 provides a basis for arguing why this should be the case. With the exception of trying to find information for themselves and getting information from their families (a seemingly unlikely source), attriters lag behind non-attriters in all of the information source categories. In other words, it would seem that they are not as well informed. In another cross-tab, not presented here, it was shown that individuals who were likely to have more information sources were more apt to like being in Germany. On this flimsy evidence, one could begin to construct a model of the attriter as being more likely to be information poor. Fortunately, we will be able to test this proposition with more substantial data as our analysis progresses.

---

*Measurement was in a 5 point independent scale at each administration.

The final table 8 shows an interesting lack of relationship between two of the "criterion" variables, attrition and intent to re-enlist. At the time of entry to USAREUR attriters show no less a propensity toward re-enlistment than non-attriters. This suggests that experiences in Germany or just plain lack of ability (social or job) are more likely causes of problems than a developed dislike of the Army. Again, subsequent analyses will be able to draw these issues out.

Based on the results presented here, few conclusions are possible, but they are suggestive of areas which will be fruitful in further analysis. Our main objection was to describe the study and its methodology because we feel that later results will be exceptionally helpful in understanding the problems of performance among first term enlisted personnel. The results presented are some of the trends which show up in the data and some suggestions of areas of useful analysis.

GROSS CHANGE IN SPARE TIME ACTIVITIES

ACTIVITIES	6 WEEKS LESS	6 WEEKS MORE	6 WEEKS N	6 MONTHS LESS	6 MONTHS MORE	6 MONTHS N
Playing or watching sports	42.6%	19.2%	470	41.1%	18.4%	423
Watching TV	21.2	29.0	464	22.3	28.4	410
Going out on dates	38.0	33.2	464	60.2	12.1	422
Doing my hobbies, e.g., collections crafts, etc.	54.8	11.5	444	48.6	18.0	406
Just hanging around with my friends	33.0	29.4	466	30.3	30.6	422
Going to the movies	22.5	46.0	448	18.2	44.8	403
Going to museums, galleries and concerts or other "cultural' activities	17.1	33.4	437	31.3	25.5	392
Playing, listening to or dancing to music	34.6	17.9	463	32.2	20.9	410
Reading	29.1	31.1	433	31.9	20.4	386
Travel	48.8	19.4	430	39.6	26.9	384

TABLE 2

GROSS CHANGE IN SPARE TIME ACTIVITY STYLES

		LESS	SAME SOME	SAME NONE	MORE	N
1.	Where Spare Time was Spent					
	6-weeks away from home (barracks)	39.2%	35.6%	0%	25.3	533
	6-months away from home (barracks)	32.0	36.6	.4	30.3	480
2.	Time Spent Indoors or Outdoors					
	6-weeks indoors	8.9	32.6	-	58.6	528
	6-months indoors	12.5	40.1	-	47.3	476
3.	Who Participated With					
	6-weeks doing things alone	4.2	6.1	-	15.4	527
	6-months doing things alone	6.4	3.9	-	13.2	466
4.	Traveling More Than 25 miles from Home					
	6-weeks - Yes	31.9	36.9	18.8	11.7	520
	6-months - Yes	29.5	40.1	17.3	13.0	451

TABLE 3

EXPECTATIONS AND EXPERIENCE:

PROBLEMS IN GERMANY

	CHANGE-MORE SERIOUS	SAME SERIOUS	SAME NOT SERIOUS	CHANGE-LESS SERIOUS	N
Getting Time Off (6 wk)	20.1%	10.1%	30.5%	39.3%	503
Getting Time Off (6 mo)	29.4	14.2	29.6	26.8	438
Finding Interesting Places (6 wk)	30.4	9.3	38.5	21.8	507
Finding Interesting Places (6 mo)	29.7	5.0	39.2	26.2	439
Finding Things that are in the States (6 wk)	40.0	21.7	23.1	15.4	502
Finding Things that are in the States (6 mo)	39.4	18.3	22.3	19.9	436
Talking to German People (6 wk)	19.1	31.2	20.3	28.9	507
Talking to German People (6 mo	18.7	28.3	20.3	32.7	440
Members of Opposite Sex (6 wk)	32.6	10.3	39.7	17.4	494
Members of Opposite Sex (6 mo	29.7	7.0	40.3	23.3	429
Able to Afford (6 wk	29.3	26.1	19.3	25.3	507
Able to Afford (6 mo	29.8	27.2	22.7	20.2	430

TABLE 4

EXPECTATION AND EXPERIENCE:
GENERAL SATISFACTION WITH BEING IN GERMANY

	SAME AS EXPECTED					
	BETTER THAN EXPECTED	GOOD	NEUTRAL	BAD	WORSE THAN EXPECTED	N
6- week	15.8%	11.3%	11.8%	9.4%	51.2%	468
6- month	17.8	10.1	11.3	8.0	55.1	516

TABLE 5

THE IMPACT OF RACE ON ARMY LIFE:
EXPECTATIONS AND EXPERIENCE

		SAME			
	Change Negative	Positive & Neutral	Negative	Positive Change	N
6 Weeks	70.6	13.9	5.2	10.1	494
6 Months	69.9	14.4	4.8	11.0	443

TABLE 6

PROPORTIONS WHO EXPECTED SERIOUS
PROBLEMS FOR ATTRITES AND NON-ATTRITES

Problem Areas	Attrites	Non-Attrites
Getting time off to go all the places I want to go	13.0%	26.9%*
Finding interesting places to go	30.4	18.1*
Finding things to do that I like to do in the States	30.4	30.5
Talking to the German people	75.0	55.6*
Getting to know people of the opposite sex to date	17.4	18.3
Being able to afford the things I want to do	43.5	42.7
	N=23-24	531-540

*Differences significant at the .01 level using a difference in proportions test.

TABLE 7

INFORMATION SEEKING AND SOURCES
PRIOR TO COMING TO GERMANY:
ATTRITES VS. NON-ATTRITES

Information Related Activity	Proportion Who Engaged in the Activity: Attrites	Non-Attrites
Received Briefings	33.3%	44.2%
Received other literature	13.3	37.3
Tried to find material yourself	54.2	46.2
Found Some Information In:		
Army Pulbications	66.7	74.2
Army NCO's & Officers	81.3	85.3
Army Orientation Program	44.5	79.6
Army Friends	70.6	84.4
Civilian Friends	63.6	74.7
My Family	84.6	77.6
Books & Magazines	75.0	79.4
Newspapers	62.5	69.3

TABLE 8

REENLISTMENT AND CAREER EXPECTATIONS
FOR ATTRITES AND NON-ATTRITES

Reenlistment Intentions	% Attrites	% Non-Attrites
Definitely Reenlist	8.3	7.2
Probably Reenlist	16.7	13.7
Undecided	37.5	44.5
Probably Won't Reenlist	16.7	12.2
Definitely Won't Reenlist	20.3	22.4
	N = 24	N = 531

(714)

PIGEON, R., and KELLETT, Capt R.G., National Defence Headquarters, Ottawa, Canada.

EFFECT OF PARTICIPATION IN THE OFFICER PROFESSIONAL DEVELOPMENT PROGRAM ON PROFESSIONAL DEVELOPMENT (Wed P.M.)

The Officer Professional Development Program (OPDP) is a major component of the Canadian Officer Professional Development System. As a first step in the global evaluation of the program empirical evidence regarding the effect of participation in the OPDP on the professional development of officers was sought. Thus, an experiment was conducted to ascertain the contribution of OPDP to professional development, to measure the difference in performance attributable to certain officer characteristics, and to compare the knowledge level of officers in each course prior to yearly program participation.

This paper describes the various steps of this evaluative research. The discussion will emphasize the experimental design used (Solomon Four-Group), and the statistical treatment of data (ANOVA and ANCOVA). Conclusions and recommendations of the research will also be discussed and should enhance comprehension of the concept of military officers' professional development.

STUDY OF THE EFFECT OF PARTICIPATION IN THE OFFICER PROFESSIONAL DEVELOPMENT PROGRAM ON THE PROFESSIONAL DEVELOPMENT OF OFFICERS

INTRODUCTION

Data on examination performance in the Officer Professional Development Program (OPDP) has been accumulated since the inception of the program in 1975. Descriptive analysis of the examination data has indicated that certain well known officer characteristics such as language group, military occupational classification (MOC), service experience, rank, education level, and command affiliation have apparently influenced the OPDP achievement of officers. Yearly examination and student feedback statistics have also supported the basic program assumption that a period of study is generally required to obtain a pass standard in each OPDP subject.

Descriptive analysis of accumulated statistical data has been revealing, but precise conclusions could not have been made from the analyzed data because the representativeness and generalizability of the findings have always been in question. For example, by just looking at OPDP examination results one could not conclude definitively that study is directly and positively related to achievement. Data such as basic knowledge levels in each OPDP course prior to course commencement are essential in appraising the effect of program participation (study) on examination results.

Such data are usually gathered in the needs assessment stage of program development since it is considered (in the behavioural sciences) an essential prerequisite to program design. In the case of OPDP, the assessment was judgemental rather than experimental. Of course, this is not unusual when a need is strongly felt by competent authority. The approach to program development taken in the OPDP has merit but does not make eventual program evaluation any easier. Queries about basic knowledge level (before study knowledge level) in each course and the influence of known officer characteristics (MOC, education and so on) on OPDP performance have naturally been raised since program inception because those questions were not addressed in the needs assessment stage.

The first attempt to relate OPDP performance and study was an experiment conducted in 1977. In that experiment the difference in performance between a group of candidates who studied (experimental group) and a group who did not study (control) was investigated. The experimentation was conducted in NDHQ for two OPDP courses and yielded significant results as to the effect of study on examination performance. However, the experiment suffered three shortcomings: limited sample size, lack of representativeness in the sample, and data gathered for only two of the six OPDP courses. The findings were sufficient to justify attempting a more sophisticated research project which would test more scientifically the relationship between OPDP participation (study) and OPDP achievement.

BACKGROUND INFORMATION

The OPDP is a major component of the Officer Professional Development System. The program is concerned with the ability of the CF Officer to relate his branch and specialty to other military activities.

As stated in the administrative order governing the program, the aim of the OPDP is "to broaden and deepen the Canadian Forces Officer's knowledge and understanding of the military profession beyond the specific technical expertise of classification training, and to contribute to the foundation of knowledge upon which further professional development will be built".

Canadian Forces officers in the ranks of second lieutenant to major who were commissioned in 1971 or later are required to successfully complete six courses deemed to be fundamental to their professional development. The six courses are:

a. OPDP 2 - General Service Knowledge;

b. OPDP 3 - Personnel Administration;

c. OPDP 4 - Military Law;

d. OPDP 5 - Financial Administration and Supply;

e. OPDP 6 - National and International Studies; and

f. OPDP 7 - War and the Military Profession.

For officers commissioned in 1971 or later, completion of the six courses is a selection prerequisite for further professional development training, namely, attendance on the Command and Staff Course.

The OPDP is a self-study program with successful completion of each course determined by measurement of performance on an objective (usually multiple choice) examination. Regular analysis each year conducted by the OPDP examination analysis cell confirms the assumption that in each course the confirmatory examinations are a good measure of course content. Because the six courses were developed from statements of knowledge and abilities, it can be further assumed that achievement in the OPDP is one measure of professional development of Forces officers. The latter point is reinforced when it is recognized that officers' OPDP participation and achievement are reported annually in the Personnel Evaluation Report (PER), which in turn is the primary device used to determine the extent of officers' professional development and their potential for further development.

(7?)

PURPOSE OF THE STUDY

The Officer Professional Development Program is entering its sixth year of operation. Close to 20,000 examinations have been written and over 800 officers have graduated, i.e. completed the six required courses. There has been general agreement at all levels within the Forces concerning the necessity of the program and the suitability of the subject matter as a means of achieving the OPDP aim, i.e. broadening and deepening officers' careers and encouraging further professional development. However, empirical evidence regarding the necessity of participation, or study, to reach a pass standard in each OPDP course has been lacking.

The central purpose and first objective of this study was TO ASCERTAIN THE CONTRIBUTION OF PARTICIPATION IN THE OPD PROGRAM TO THE PROFESSIONAL DEVELOPMENT OF OFFICERS. Actually, for each of the six OPDP courses, differences in professional development between participating and non-participating officers were examined. The regular OPDP examinations, considered to be measures of different aspects of officer professional development, were used to generate data for analysis.

Additional objectives of the study were:

a. SECOND OBJECTIVE: TO MEASURE THE DIFFERENCE IN PERFORMANCE ATTRIBUTABLE TO CERTAIN OFFICER CHARACTERISTICS (RANK, OCCUPATION, EDUCATION, ENTRY PLAN, COMMAND, LANGUAGE) IN OPDP COURSES;

b. THIRD OBJECTIVE: TO COMPARE THE GENERAL KNOWLEDGE LEVEL OF CANDIDATES IN EACH OPDP COURSE PRIOR TO YEARLY PROGRAM PARTICIPATION.

METHOD

Experimental Design

As a forerunner to the present study, an exploratory investigation, designed to measure the effect of participation in OPDP on examination results, was carried out. In that investigation significant differences were observed between "with study" and "no study" groups in terms of annual confirmatory examination performance. The scope of the study was limited to OPDP 3, Personnel Administration, and OPDP 7, War and the Military Profession, and the sample was drawn from NDHQ participating officers only.

In order to achieve the main purpose of the present study and to meet the additional goals stated above, the Solomon Four-Group approach was used. This design provides greater control in a study

than a simple "after the fact" experimental-control design. It has been asserted that this design "is the most desirable of all the really basic experimental designs, both quasi and true". The design involves four groups of which two are pretested, one of them receiving a treatment (participation or study in OPDP) and the other not, and of which two are unpretested, one of them receiving a treatment and the other not. (see figure below).

Solomon Four-Group Design

R	O_1	X	O_2
R	O_3		O_4
R		X	O_5
R			O_6

where R, randomized assignment of subjects
O, observations of performance in OPDP
X, participation in OPDP

Even though it is not widely used in educational research due to cost, this design has many advantageous features. It has been said: "The Solomon-Four-Group Design is considered to be a highly prestigious experimental design. Not only does this design control for all the threats to internal validity (history, maturation, etc.) but it also allows the researcher to ascertain whether there is a pretest-treatment interaction".

Sample

The OPDP population consists of some 5000 officers in the ranks of second lieutenant to major commissioned on or after 1 Jan 71. Officers register in the fall of a study year, and then confirm their success or failure in the course or courses undertaken by writing a confirmatory examination the following spring. To obtain a sample of candidates (N) for groups O_1 and O_3, eleven large CF representative bases were visited in the fall of 1978. In order to assure minimum disruption to the regular OPDP registration cycle, individuals assigned to experimental groups, O_1, were chosed at random from the list of candidates already registered for given courses. Individuals assigned to control groups, O_3, were chosen also at random from the list of non-registered, non credited, "exam never written" candidates in given courses.

Group O_5 for each course was drawn from the regular population writing examinations in April or June 1979. Samples represented about 5% of the examination population. Group O_6 were individuals chosen in the same fashion as Group O_3, but from different large CF bases.

Instruments

Instruments used to gather data on professional development knowledge level of participating officers were the six OPDP examinations already constructed for such purpose. Each of the six examinations measures an aspect of professional development and corresponds to each of the six OPDP courses. The 1978 bilingual primary examination form was used in the pretest (O_1 and O_3). Content analysis and expert opinion permit us to assume that the curricula of the six OPDP courses represented defined levels of knowledge in the professional development of officers, and that attainment of those knowledge levels was adequately measured by each of the six OPDP examinations. Reliability coefficients of the examinations, calculated from populations varying in number between 487 and 1044 officers, yielded KR-20 values between .82 and .91. Instruments used in posttest (O_2, O_4, O_5, O_6) were the regular 1979 OPDP course examinations, a parallel form of the 1978 versions. Internal consistency of the 1979 examinations also ranged between the same KR-20 values as the 1978 versions.

OPDP examinations 2, 3, 5, 6, 7 each consist of 75 mulitple-choice items covering the study material. OPDP 4 has 60 multiple-choice items and 20 items of the type True-False with the obligation for the candidate to write down a reference number justifying his answer. This latter part of the OPDP 4 examination was not considered in the present study. All examinations administered were closed-book.

Statistical Treatment of Data

In order to meet the first objective of this study, i.e. to ascertain the contribution of participation in the OPD program, analysis of variance techniques were used. Following Campbell and Stanley's recommendation, a 2 X 2 factorial ANOVA was applied to posttest scores (groups O_2, O_4, O_5, and O_6), with scores on pretest being disregarded. This ANOVA yielded for each OPDP course one F-ratio for the "treatment" main effect, one for the "pretest" main effect, and one for the interaction between the two main effects.

An analysis of covariance using pretest scores as a covariate was then made. This was done in accordance with a Campbell and Stanley recommendation: "if the main and interactive effects of pretesting are negligible, it may be desirable to perform an analysis of covariance of O_4 versus O_2, pretest scores being the covariates". This practice of using pretest data as a covariate is advocated by Linn and Slinde with caution given to the basic assumptions of this technique, namely homogeneity of regression and measurement of the covariate without error.

The second objective of measuring the difference in performance between groups of certain variables was met by breaking down, for the six courses, posttest results (O_2) into categories representing the different groups of variables. ANOVA was used to test for significant differences between weighted means of these groups while pretest scores served as covariates when justifiable. The third objective, which was to measure the pre-study knowledge level of candidates, was met by calculating descriptive statistics for group O_1 and testing for significant differences between means of given variable groups. Calculations were performed by the programs SPSS (Statistical Package For Social Sciences) and FRULM (Full Rank Univaviate Linear Model).

RESULTS AND DISCUSSION

Selection of the "Solomon Four-Group Design" as the experimental plan for the OPDP Research Project was based on the known effectiveness of the design in controlling the influence of intervening variables. Use of the Solomon design permitted the quantitative assessment of the effect of participation, or study, in OPDP on examination results, and thus on officers' professional development, exclusive of the influence of intervening variables related to officers' backgrounds or of the impact of pretesting on posttesting. Assumptions underlying use of the design were not violated, except that complete random assignment of subjects to groups was not possible. As mentioned earlier, subjects were assigned to pretest experimental or control groups based on their course registrations for the 1978-79 study year.

OPDP Contribution to Officer's Professional Development

In order to perform the analysis of variance on the four posttested groups the three sources of variance identified were, as mentioned earlier: "pretest", "treatment", and "interaction". "Pretest" effect involves two groups of subjects (O_2 and O_4) to whom the 1978 examinations were administered, and two groups (O_5 and O_6) who did not write these examinations. The "treatment" effect also involves two categories: two groups (O_2 and O_5) consisting of regular participants in a given OPDP course for the academic year 1978-79, and two groups (O_4 and O_6) who were not participants that year. The treatment effect is best described as a process of registering in a given subject, drawing study materials, maintaining registration, updating study material, studying from zero to over seventy hours, and writing confirmatory examinations. "Interaction" effect is the combined effect of pretest and treatment.

Results clearly showed that treatment was significant for every OPDP course with $P < 0.001$. Thus, in and of itself treatment truly influences posttest performance. On the other hand, the analysis of variance (ANOVA) showed that the effect of pretest on posttest, and the interaction of pretest and treatment were, as expected, not significant for each of the six OPDP courses. This means that pretesting candidates as part of the experimental procedure did not alter, or contaminate, posttest scores.

At this state it was informative to assess the strength of association represented by significant treatment. This may be done by rough estimates of ω^2. These values were estimated according to Hays (1966). The percent of variance accounted for by study or treatment effect was as follows: OPDP 2: 0.23; OPDP 3: 0.22; OPDP 4: 0.35; OPDP 5: 0.41; OPDP 6: 0.12; OPDP 7: 0.22 These estimations of associations suggest that study was moderately correlated to achievement in OPDP 4 and 5, correlated in OPDP 2, 3, and 7 and lowly correlated in OPDP 6. The nature of courses explains these results. Military Law (4) and Financial Administration (5) are for most officers brand new subjects and study has to be the major contributor to explained variance in results. Success in General Service Knowledge (2), Personnel Administration (3), and War and Military Profession (7), is probably a mixture of study and other factors (experience on the job, entry plan, personal interest, etc.). Success in National and International Studies (6) is least affected by study mainly because it is a subject of general interest to all officers regardless of whether or not they participate in the OPDP.

As a second step in analysis, the effect of treatment on posttest results with before study knowledge level removed was determined by carrying out an analysis of covariance (ANCOVA) with pretest scores serving as the covariate. In this case only posttest scores for pretested groups (experimental and control) were examined. Again, treatment was found to be significant. Use of pretest scores as a covariate was originally justified on the basis of high correlation between pretest results and posttest results. Correlations were significant at $P < 0.001$ for all OPDP courses except one. It was not significant for OPDP course 5 (Financial Administration and Supply), so pretest scores were not used as a covariate in the analysis of data for that course.

Comparison of Performance Between Groups of Certain Officer Characteristics.

Historically, differences in OPDP performance between officers with specific background characteristics had been observed and seemed to be significant. Having no measures of before study knowledge in each course it was impossible to determine accurately whether differences in final examination performance were attributable to officers' backgrounds, or to before study knowledge differences. As part of the OPDP Research Project, a measure of difference in performance between groups of six identifiable officers characteristics (MOC, education level, command, entry plan, rank, and language) was obtained using pretest scores (pre-study knowledge level) as covariates. In other words, if officers' performance in OPDP had been significantly influenced by one or more of the identified characteristics, with the effect of before study knowledge removed or discounted, the analysis of covariance would have revealed it.

To test for significant differences between groups of officer characteristics in each OPDP course, posttest scores (group O_2) were submitted to analysis of variance, and pretest scores (group O_1) served as the covariate (except for OPDP 5 where pretest as covariate was not significant: P = 0.49). Sources of variance for each ANOVA were the covariate, i.e. the pretest score, the officer characteristic, and the error. F's for covariates and officer characteristics were tested at α = 0.05.

Overall, the analysis showed that the influence of officer characteristics in determining final examination performance was minimal. There were only seven F values which were significant out of thirty-six possibilities (six OPDP courses by six groups of officer characteristics). Post-hoc procedures were employed in seven instances, the procedure in each case being the application of the Scheffé test, which is a particularly useful means of locating a significant difference when the n's are very different. In fact, two out of the seven significant F values had to be discounted due to small sample sizes. In two other cases the F values over the entire groups were significant, but it was impossible to locate where the differences lay by making simple contrasts between pairs of means.

The three cases where the F value was significant and sample size large enough to enable simple contrasts were:

a. Course 5. There was a significant difference in performance between those whose entry plan was ROTP, and those characterized by entry plan OCTP. ROTP entrants scored consistently higher than OCTP entrants. The mean score for ROTP subjects was 74.82 (n=21), and the mean score for OCTP subjects was 70.00 (n=6).

b. Course 2. Anglophones performed significantly better in this course than francophones. Mean score for anglophones was 73.71 (n=46) while for francophones the mean score was 69.22 (n=20).

c. Course 6. As was the case in course 2, anglophone scores were superior to francophone scores. Mean score for anglophones was 73.34 (n=59) compared to 67.82 for francophones (n=20).

It is surmised that subjects whose entry plan was ROTP performed significantly better than those with entry plan OCTP in course 5 because a military college or civilian university academic background in addition to a long apprenticeship period as officer cadets had given those officers an experiental framework which enabled them to make maximum use of the course study package. Superiority in performance of anglophones over francophones in courses 2 and 6 can be attributed largely to the nature of the subject matter in both cases. More than any of the other OPDP courses, courses 2 and 6 consist of terminology and concepts which exist predominately in the English language and culture. Simple translation of study material does not assure conceptual acceptability of the content by another linguistic or cultural group. As well, there is little doubt that anglophones enjoy an abundance of background reading material for the current events portion of course 6, an advantage not shared by francophones.

Comparison of Candidates' Pre-Study Professional Knowledge Level.

The third objective of the present study was to identify and compare the pre-study knowledge levels of candidates in each of the six OPDP courses. Used as samples for each OPDP course were the original 0_1 and 0_3 groups having written the pretest. For each OPDP course descriptive statistics were computed.

In order to compare pre-study knowledge level of candidates between groups of a given characteristic, analysis of variance was performed on each characteristic for each OPDP course. Whenever p-values were equal to or smaller than 0.05, Scheffé procedure was carried out in order to identify pairs of significantly different means. Again the use of the Scheffé test was especially appropriate because of very different group sizes.

In terms of pretest results (before study knowledge level) in relation to the six officer characteristics mentioned earlier, significant effects were observed in four of the six courses. There was no significant difference in before-study knowledge level between officers of different characteristic backgrounds for courses 5 and 7. In courses 2, 3, and 4, anglophones had a higher level of knowledge. However, that superiority was not accentuated by participation, or study, in courses 3 and 4. As noted above, anglophones profited more from study in course 2 than did francophones. In courses 3 and 6 university graduates had significantly higher pretest scores than non-university subjects. It is evident that a university education at least provides officers with a greater understanding of personnel administration, training systems, governmental and world affairs than would otherwise be the case. Finally, in course 6 captains obtained consistently higher pretest scores than lieutenants. It is speculated that captains are better able to keep up with national and international affairs than are lieutenants because they normally are no longer undertaking intensive, basic job training and so have the time

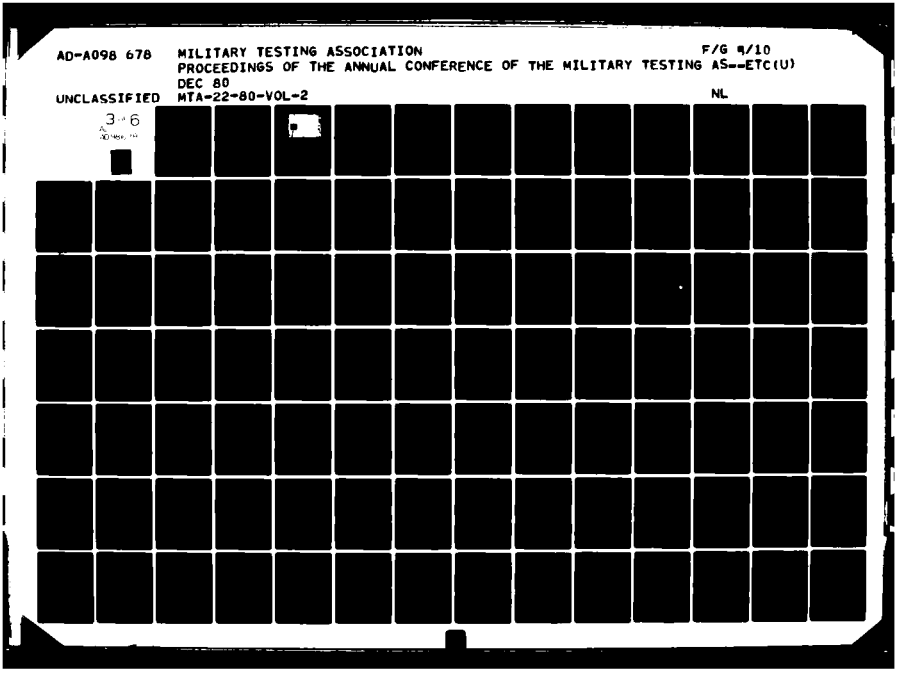
to read more broadly and have, by that point in their professional development, acquired some profitable reading habits.

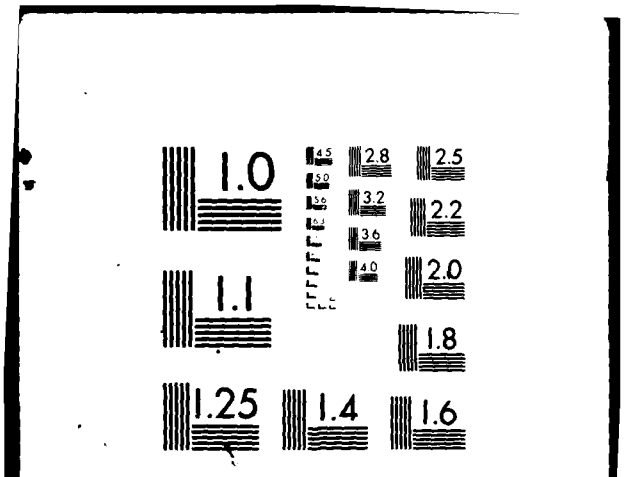
CONCLUSIONS AND RECOMMENDATIONS

The major finding of this study is that participation in the OPD program contributes significantly to the professional development of officers. OPDP study provides officers with professional knowledge that they do not ordinarily acquire in the course of career development. Even though officers may commence study at unequal knowledge or experience levels, the OPD program does not in general give advantage to a particular group or groups of officers distinguished by MOC, entry plan, years of commissioned service, education level, or primary language. The notable exceptions to this conclusion were courses 2 and 6, where renewed effort to equalize the presentation of study materials to the linguistic groups needs to be made. It is further recommended that in course 5 OCTP officers be formally advised to devote more time to study than the number of hours allotted in the Study Guide.

726

AD-A098 678 MILITARY TESTING ASSOCIATION F/G 9/10
PROCEEDINGS OF THE ANNUAL CONFERENCE OF THE MILITARY TESTING AS--ETC(U)
DEC 80
UNCLASSIFIED MTA-22-80-VOL-2 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
7.1
8.0
9.0
10.0
11.2
12.5
14.0
16.0
18.0
20.0
22.4
25.0
28.0
32.0
36.0
40.0
45.0
50.0
56.0
63.0
71.0
80.0
90.0
100.0
112.0
125.0
140.0
160.0
180.0
200.0
224.0
250.0
280.0
320.0
360.0
400.0
450.0
500.0
560.0
630.0
710.0
800.0
900.0
1000.0
1120.0
1250.0
1400.0
1600.0
1800.0
2000.0
2240.0
2500.0
2800.0
3200.0
3600.0
4000.0
4500.0
5000.0
5600.0
6300.0
7100.0
8000.0
9000.0
10000.0
11200.0
12500.0
14000.0
16000.0
18000.0
20000.0
22400.0
25000.0
28000.0
32000.0
36000.0
40000.0
45000.0
50000.0
56000.0
63000.0
71000.0
80000.0
90000.0
100000.0
112000.0
125000.0
140000.0
160000.0
180000.0
200000.0
224000.0
250000.0
280000.0
320000.0
360000.0
400000.0
450000.0
500000.0
560000.0
630000.0
710000.0
800000.0
900000.0
1000000.0
1120000.0
1250000.0
1400000.0
1600000.0
1800000.0
2000000.0
2240000.0
2500000.0
2800000.0
3200000.0
3600000.0
4000000.0
4500000.0
5000000.0
5600000.0
6300000.0
7100000.0
8000000.0
9000000.0
10000000.0
11200000.0
12500000.0
14000000.0
16000000.0
18000000.0
20000000.0
22400000.0
25000000.0
28000000.0
32000000.0
36000000.0
40000000.0
45000000.0
50000000.0
56000000.0
63000000.0
71000000.0
80000000.0
90000000.0
100000000.0
112000000.0
125000000.0
140000000.0
160000000.0
180000000.0
200000000.0
224000000.0
250000000.0
280000000.0
320000000.0
360000000.0
400000000.0
450000000.0
500000000.0
560000000.0
630000000.0
710000000.0
800000000.0
900000000.0
1000000000.0
1120000000.0
1250000000.0
1400000000.0
1600000000.0
1800000000.0
2000000000.0
2240000000.0
2500000000.0
2800000000.0
3200000000.0
3600000000.0
4000000000.0
4500000000.0
5000000000.0
5600000000.0
6300000000.0
7100000000.0
8000000000.0
9000000000.0
10000000000.0
11200000000.0
12500000000.0
14000000000.0
16000000000.0
18000000000.0
20000000000.0
22400000000.0
25000000000.0
28000000000.0
32000000000.0
36000000000.0
40000000000.0
45000000000.0
50000000000.0
56000000000.0
63000000000.0
71000000000.0
80000000000.0
90000000000.0
100000000000.0
112000000000.0
125000000000.0
140000000000.0
160000000000.0
180000000000.0
200000000000.0
224000000000.0
250000000000.0
280000000000.0
320000000000.0
360000000000.0
400000000000.0
450000000000.0
500000000000.0
560000000000.0
630000000000.0
710000000000.0
800000000000.0
900000000000.0
1000000000000.0
1120000000000.0
1250000000000.0
1400000000000.0
1600000000000.0
1800000000000.0
2000000000000.0
2240000000000.0
2500000000000.0
2800000000000.0
3200000000000.0
3600000000000.0
4000000000000.0
4500000000000.0
5000000000000.0
5600000000000.0
6300000000000.0
7100000000000.0
8000000000000.0
9000000000000.0
10000000000000.0
11200000000000.0
12500000000000.0
14000000000000.0
16000000000000.0
18000000000000.0
20000000000000.0
22400000000000.0
25000000000000.0
28000000000000.0
32000000000000.0
36000000000000.0
40000000000000.0
45000000000000.0
50000000000000.0
56000000000000.0
63000000000000.0
71000000000000.0
80000000000000.0
90000000000000.0
100000000000000.0
112000000000000.0
125000000000000.0
140000000000000.0
160000000000000.0
180000000000000.0
200000000000000.0
224000000000000.0
250000000000000.0
280000000000000.0
320000000000000.0
360000000000000.0
400000000000000.0
450000000000000.0
500000000000000.0
560000000000000.0
630000000000000.0
710000000000000.0
800000000000000.0
900000000000000.0
1000000000000000.0
1120000000000000.0
1250000000000000.0
1400000000000000.0
1600000000000000.0
1800000000000000.0
2000000000000000.0
2240000000000000.0
2500000000000000.0
2800000000000000.0
3200000000000000.0
3600000000000000.0
4000000000000000.0
4500000000000000.0
5000000000000000.0
5600000000000000.0
6300000000000000.0
7100000000000000.0
8000000000000000.0
9000000000000000.0
10000000000000000.0
11200000000000000.0
12500000000000000.0
14000000000000000.0
16000000000000000.0
18000000000000000.0
20000000000000000.0
22400000000000000.0
25000000000000000.0
28000000000000000.0
32000000000000000.0
36000000000000000.0
40000000000000000.0
45000000000000000.0
50000000000000000.0
56000000000000000.0
63000000000000000.0
71000000000000000.0
80000000000000000.0
90000000000000000.0
100000000000000000.0
112000000000000000.0
125000000000000000.0
140000000000000000.0
160000000000000000.0
180000000000000000.0
200000000000000000.0
224000000000000000.0
250000000000000000.0
280000000000000000.0
320000000000000000.0
360000000000000000.0
400000000000000000.0
450000000000000000.0
500000000000000000.0
560000000000000000.0
630000000000000000.0
710000000000000000.0
800000000000000000.0
900000000000000000.0
1000000000000000000.0
1120000000000000000.0
1250000000000000000.0
1400000000000000000.0
1600000000000000000.0
1800000000000000000.0
2000000000000000000.0
2240000000000000000.0
2500000000000000000.0
2800000000000000000.0
3200000000000000000.0
3600000000000000000.0
4000000000000000000.0
4500000000000000000.0
5000000000000000000.0
5600000000000000000.0
6300000000000000000.0
7100000000000000000.0
8000000000000000000.0
9000000000000000000.0
10000000000000000000.0
11200000000000000000.0
12500000000000000000.0
14000000000000000000.0
16000000000000000000.0
18000000000000000000.0
20000000000000000000.0
22400000000000000000.0
25000000000000000000.0
28000000000000000000.0
32000000000000000000.0
36000000000000000000.0
40000000000000000000.0
45000000000000000000.0
50000000000000000000.0
56000000000000000000.0
63000000000000000000.0
71000000000000000000.0
80000000000000000000.0
90000000000000000000.0
100000000000000000000.0
112000000000000000000.0
125000000000000000000.0
140000000000000000000.0
160000000000000000000.0
180000000000000000000.0
200000000000000000000.0
224000000000000000000.0
250000000000000000000.0
280000000000000000000.0
320000000000000000000.0
360000000000000000000.0
400000000000000000000.0
450000000000000000000.0
500000000000000000000.0
560000000000000000000.0
630000000000000000000.0
710000000000000000000.0
800000000000000000000.0
900000000000000000000.0
1000000000000000000000.0
1120000000000000000000.0
1250000000000000000000.0
1400000000000000000000.0
1600000000000000000000.0
1800000000000000000000.0
2000000000000000000000.0
2240000000000000000000.0
2500000000000000000000.0
2800000000000000000000.0
3200000000000000000000.0
3600000000000000000000.0
4000000000000000000000.0
4500000000000000000000.0
5000000000000000000000.0
5600000000000000000000.0
6300000000000000000000.0
7100000000000000000000.0
8000000000000000000000.0
9000000000000000000000.0
10000000000000000000000.0
11200000000000000000000.0
12500000000000000000000.0
14000000000000000000000.0
16000000000000000000000.0
18000000000000000000000.0
20000000000000000000000.0
22400000000000000000000.0
25000000000000000000000.0
28000000000000000000000.0
32000000000000000000000.0
36000000000000000000000.0
40000000000000000000000.0
45000000000000000000000.0
50000000000000000000000.0
56000000000000000000000.0
63000000000000000000000.0
71000000000000000000000.0
80000000000000000000000.0
90000000000000000000000.0
100000000000000000000000.0
112000000000000000000000.0
125000000000000000000000.0
140000000000000000000000.0
160000000000000000000000.0
180000000000000000000000.0
200000000000000000000000.0
224000000000000000000000.0
250000000000000000000000.0
280000000000000000000000.0
320000000000000000000000.0
360000000000000000000000.0
400000000000000000000000.0
450000000000000000000000.0
500000000000000000000000.0
560000000000000000000000.0
630000000000000000000000.0
710000000000000000000000.0
800000000000000000000000.0
900000000000000000000000.0
1000000000000000000000000.0
1120000000000000000000000.0
1250000000000000000000000.0
1400000000000000000000000.0
1600000000000000000000000.0
1800000000000000000000000.0
2000000000000000000000000.0
2240000000000000000000000.0
2500000000000000000000000.0
2800000000000000000000000.0
3200000000000000000000000.0
3600000000000000000000000.0
4000000000000000000000000.0
4500000000000000000000000.0
5000000000000000000000000.0
5600000000000000000000000.0
6300000000000000000000000.0
7100000000000000000000000.0
8000000000000000000000000.0
9000000000000000000000000.0
10000000000000000000000000.0
11200000000000000000000000.0
12500000000000000000000000.0
14000000000000000000000000.0
16000000000000000000000000.0
18000000000000000000000000.0
20000000000000000000000000.0
22400000000000000000000000.0
25000000000000000000000000.0
28000000000000000000000000.0
32000000000000000000000000.0
36000000000000000000000000.0
40000000000000000000000000.0
45000000000000000000000000.0
50000000000000000000000000.0
56000000000000000000000000.0
63000000000000000000000000.0
71000000000000000000000000.0
80000000000000000000000000.0
90000000000000000000000000.0
100000000000000000000000000.0
112000000000000000000000000.0
125000000000000000000000000.0
140000000000000000000000000.0
160000000000000000000000000.0
180000000000000000000000000.0
200000000000000000000000000.0
224000000000000000000000000.0
250000000000000000000000000.0
280000000000000000000000000.0
320000000000000000000000000.0
360000000000000000000000000.0
400000000000000000000000000.0
450000000000000000000000000.0
500000000000000000000000000.0
560000000000000000000000000.0
630000000000000000000000000.0
710000000000000000000000000.0
800000000000000000000000000.0
900000000000000000000000000.0
1000000000000000000000000000.0
1120000000000000000000000000.0
1250000000000000000000000000.0
1400000000000000000000000000.0
1600000000000000000000000000.0
1800000000000000000000000000.0
2000000000000000000000000000.0
2240000000000000000000000000.0
2500000000000000000000000000.0
2800000000000000000000000000.0
3200000000000000000000000000.0
3600000000000000000000000000.0
4000000000000000000000000000.0
4500000000000000000000000000.0
5000000000000000000000000000.0
5600000000000000000000000000.0
6300000000000000000000000000.0
7100000000000000000000000000.0
8000000000000000000000000000.0
9000000000000000000000000000.0
10000000000000000000000000000.0
11200000000000000000000000000.0
12500000000000000000000000000.0
14000000000000000000000000000.0
16000000000000000000000000000.0
18000000000000000000000000000.0
20000000000000000000000000000.0
22400000000000000000000000000.0
25000000000000000000000000000.0
28000000000000000000000000000.0
32000000000000000000000000000.0
36000000000000000000000000000.0
40000000000000000000000000000.0
45000000000000000000000000000.0
50000000000000000000000000000.0
56000000000000000000000000000.0
63000000000000000000000000000.0
71000000000000000000000000000.0
80000000000000000000000000000.0
90000000000000000000000000000.0
100000000000000000000000000000.0
112000000000000000000000000000.0
125000000000000000000000000000.0
140000000000000000000000000000.0
160000000000000000000000000000.0
180000000000000000000000000000.0
200000000000000000000000000000.0
224000000000000000000000000000.0
250000000000000000000000000000.0
280000000000000000000000000000.0
320000000000000000000000000000.0
360000000000000000000000000000.0
400000000000000000000000000000.0
450000000000000000000000000000.0
500000000000000000000000000000.0
560000000000000000000000000000.0
630000000000000000000000000000.0
710000000000000000000000000000.0
800000000000000000000000000000.0
900000000000000000000000000000.0
1000000000000000000000000000000.0
1120000000000000000000000000000.0
1250000000000000000000000000000.0
1400000000000000000000000000000.0
1600000000000000000000000000000.0
1800000000000000000000000000000.0
2000000000000000000000000000000.0
2240000000000000000000000000000.0
2500000000000000000000000000000.0
2800000000000000000000000000000.0
3200000000000000000000000000000.0
3600000000000000000000000000000.0
4000000000000000000000000000000.0
4500000000000000000000000000000.0
5000000000000000000000000000000.0
5600000000000000000000000000000.0
6300000000000000000000000000000.0
7100000000000000000000000000000.0
8000000000000000000000000000000.0
9000000000000000000000000000000.0
10000000000000000000000000000000.0
11200000000000000000000000000000.0
12500000000000000000000000000000.0
14000000000000000000000000000000.0
16000000000000000000000000000000.0
18000000000000000000000000000000.0
20000000000000000000000000000000.0
22400000000000000000000000000000.0
25000000000000000000000000000000.0
28000000000000000000000000000000.0
32000000000000000000000000000000.0
36000000000000000000000000000000.0
40000000000000000000000000000000.0
45000000000000000000000000000000.0
50000000000000000000000000000000.0
56000000000000000000000000000000.0
63000000000000000000000000000000.0
71000000000000000000000000000000.0
80000000000000000000000000000000.0
90000000000000000000000000000000.0
100000000000000000000000000000000.0
112000000000000000000000000000000.0
125000000000000000000000000000000.0
140000000000000000000000000000000.0
160000000000000000000000000000000.0
180000000000000000000000000000000.0
200000000000000000000000000000000.0
224000000000000000000000000000000.0
250000000000000000000000000000000.0
280000000000000000000000000000000.0
320000000000000000000000000000000.0
360000000000000000000000000000000.0
400000000000000000000000000000000.0
450000000000000000000000000000000.0
500000000000000000000000000000000.0
560000000000000000000000000000000.0
630000000000000000000000000000000.0
710000000000000000000000000000000.0
800000000000000000000000000000000.0
900000000000000000000000000000000.0
1000000000000000000000000000000000.0
1120000000000000000000000000000000.0
1250000000000000000000000000000000.0
1400000000000000000000000000000000.0
1600000000000000000000000000000000.0
1800000000000000000000000000000000.0
2000000000000000000000000000000000.0
2240000000000000000000000000000000.0
2500000000000000000000000000000000.0
2800000000000000000000000000000000.0
3200000000000000000000000000000000.0
3600000000000000000000000000000000.0
4000000000000000000000000000000000.0
4500000000000000000000000000000000.0
5000000000000000000000000000000000.0
5600000000000000000000000000000000.0
6300000000000000000000000000000000.0
7100000000000000000000000000000000.0
8000000000000000000000000000000000.0
9000000000000000000000000000000000.0
10000000000000000000000000000000000.0
11200000000000000000000000000000000.0
12500000000000000000000000000000000.0
14000000000000000000000000000000000.0
16000000000000000000000000000000000.0
18000000000000000000000000000000000.0
20000000000000000000000000000000000.0
22400000000000000000000000000000000.0
25000000000000000000000000000000000.0
28000000000000000000000000000000000.0
32000000000000000000000000000000000.0
36000000000000000000000000000000000.0
40000000000000000000000000000000000.0
45000000000000000000000000000000000.0
50000000000000000000000000000000000.0
56000000000000000000000000000000000.0
63000000000000000000000000000000000.0
71000000000000000000000000000000000.0
80000000000000000000000000000000000.0
90000000000000000000000000000000000.0
100000000000000000000000000000000000.0
112000000000000000000000000000000000.0
125000000000000000000000000000000000.0
140000000000000000000000000000000000.0
160000000000000000000000000000000000.0
180000000000000000000000000000000000.0
200000000000000000000000000000000000.0
224000000000000000000000000000000000.0
250000000000000000000000000000000000.0
280000000000000000000000000000000000.0
320000000000000000000000000000000000.0
360000000000000000000000000000000000.0
400000000000000000000000000000000000.0
450000000000000000000000000000000000.0
500000000000000000000000000000000000.0
560000000000000000000000000000000000.0
630000000000000000000000000000000000.0
710000000000000000000000000000000000.0
800000000000000000000000000000000000.0
900000000000000000000000000000000000.0
1000000000000000000000000000000000000.0
1120000000000000000000000000000000000.0
1250000000000000000000000000000000000.0
1400000000000000000000000000000000000.0
1600000000000000000000000000000000000.0
1800000000000000000000000000000000000.0
2000000000000000000000000000000000000.0
2240000000000000000000000000000000000.0
2500000000000000000000000000000000000.0
2800000000000000000000000000000000000.0
3200000000000000000000000000000000000.0
3600000000000000000000000000000000000.0
4000000000000000000000000000000000000.0
4500000000000000000000000000000000000.0
5000000000000000000000000000000000000.0
5600000000000000000000000000000000000.0
6300000000000000000000000000000000000.0
7100000000000000000000000000000000000.0
8000000000000000000000000000000000000.0
9000000000000000000000000000000000000.0
10000000000000000000000000000000000000.0
11200000000000000000000000000000000000.0
12500000000000000000000000000000000000.0
14000000000000000000000000000000000000.0
16000000000000000000000000000000000000.0
18000000000000000000000000000000000000.0
20000000000000000000000000000000000000.0
22400000000000000000000000000000000000.0
25000000000000000000000000000000000000.0
28000000000000000000000000000000000000.0
32000000000000000000000000000000000000.0
36000000000000000000000000000000000000.0
40000000000000000000000000000000000000.0
45000000000000000000000000000000000000.0
50000000000000000000000000000000000000.0
56000000000000000000000000000000000000.0
63000000000000000000000000000000000000.0
71000000000000000000000000000000000000.0
80000000000000000000000000000000000000.0
90000000000000000000000000000000000000.0
100000000000000000000000000000000000000.0
112000000000000000000000000000000000000.0
125000000000000000000000000000000000000.0
140000000000000000000000000000000000000.0
160000000000000000000000000000000000000.0
180000000000000000000000000000000000000.0
200000000000000000000000000000000000000.0
224000000000000000000000000000000000000.0
250000000000000000000000000000000000000.0
280000000000000000000000000000000000000.0
320000000000000000000000000000000000000.0
360000000000000000000000000000000000000.0
400000000000000000000000000000000000000.0
450000000000000000000000000000000000000.0
500000000000000000000000000000000000000.0
560000000000000000000000000000000000000.0
630000000000000000000000000000000000000.0
710000000000000000000000000000000000000.0
800000000000000000000000000000000000000.0
900000000000000000000000000000000000000.0
1000000000000000000000000000000000000000.0
1120000000000000000000000000000000000000.0
1250000000000000000000000000000000000000.0
1400000000000000000000000000000000000000.0
1600000000000000000000000000000000000000.0
1800000000000000000000000000000000000000.0
2000000000000000000000000000000000000000.0
2240000000000000000000000000000000000000.0
2500000000000000000000000000000000000000.0
2800000000000000000000000000000000000000.0
3200000000000000000000000000000000000000.0
3600000000000000000000000000000000000000.0
4000000000000000000000000000000000000000.0
4500000000000000000000000000000000000000.0
5000000000000000000000000000000000000000.0
5600000000000000000000000000000000000000.0
6300000000000000000000000000000000000000.0
7100000000000000000000000000000000000000.0
8000000000000000000000000000000000000000.0
9000000000000000000000000000000000000000.0
10000000000000000000000000000000000000000.0
11200000000000000000000000000000000000000.0
12500000000000000000000000000000000000000.0
14000000000000000000000000000000000000000.0
16000000000000000000000000000000000000000.0
18000000000000000000000000000000000000000.0
20000000000000000000000000000000000000000.0
22400000000000000000000000000000000000000.0
25000000000000000000000000000000000000000.0
28000000000000000000000000000000000000000.0
32000000000000000000000000000000000000000.0
36000000000000000000000000000000000000000.0
40000000000000000000000000000000000000000.0
45000000000000000000000000000000000000000.0
50000000000000000000000000000000000000000.0
56000000000000000000000000000000000000000.0
63000000000000000000000000000000000000000.0
71000000000000000000000000000000000000000.0
80000000000000000000000000000000000000000.0
90000000000000000000000000000000000000000.0
100000000000000000000000000000000000000000.0
112000000000000000000000000000000000000000.0
125000000000000000000000000000000000000000.0
140000000000000000000000000000000000000000.0
160000000000000000000000000000000000000000.0
180000000000000000000000000000000000000000.0
200000000000000000000000000000000000000000.0
224000000000000000000000000000000000000000.0
250000000000000000000000000000000000000000.0
280000000000000000000000000000000000000000.0
320000000000000000000000000000000000000000.0
360000000000000000000000000000000000000000.0
400000000000000000000000000000000000000000.0
450000000000000000000000000000000000000000.0
500000000000000000000000000000000000000000.0
560000000000000000000000000000000000000000.0
630000000000000000000000000000000000000000.0
710000000000000000000000000000000000000000.0
800000000000000000000000000000000000000000.0
900000000000000000000000000000000000000000.0
1000000000000000000000000000000000000000000.0
1120000000000000000000000000000000000000000.0
1250000000000000000000000000000000000000000.0
1400000000000000000000000000000000000000000.0
1600000000000000000000000000000000000000000.0
1800000000000000000000000000000000000000000.0
2000000000000000000000000000000000000000000.0
2240000000000000000000000000000000000000000.0
2500000000000000000000000000000000000000000.0
2800000000000000000000000000000000000000000.0
3200000000000000000000000000000000000000000.0
3600000000000000000000000000000000000000000.0
4000000000000000000000000000000000000000000.0
4500000000000000000000000000000000000000000.0
5000000000000000000000000000000000000000000.0
5600000000000000000000000000000000000000000.0
6300000000000000000000000000000000000000000.0
7100000000000000000000000000000000000000000.0
8000000000000000000000000000000000000000000.0
9000000000000000000000000000000000000000000.0
10000000000000000000000000000000000000000000.0
11200000000000000000000000000000000000000000.0
12500000000000000000000000000000000000000000.0
14000000000000000000000000000000000000000000.0
16000000000000000000000000000000000000000000.0
18000000000000000000000000000000000000000000.0
20000000000000000000000000000000000000000000.0
22400000000000000000000000000000000000000000.0
25000000000000000000000000000000000000000000.0
28000000000000000000000000000000000000000000.0
32000000000000000000000000000000000000000000.0
36000000000000000000000000000000000000000000.0
40000000000000000000000000000000000000000000.0
45000000000000000000000000000000000000000000.0
50000000000000000000000000000000000000000000.0
56000000000000000000000000000000000000000000.0
63000000000000000000000000000000000000000000.0
71000000000000000000000000000000000000000000.0
80000000000000000000000000000000000000000000.0
90000000000000000000000000000000000000000000.0
100000000000000000000000000000000000000000000.0
112000000000000000000000000000000000000000000.0
125000000000000000000000000000000000000000000.0
140000000000000000000000000000000000000000000.0
160000000000000000000000000000000000000000000.0
180000000000000000000000000000000000000000000.0
200000000000000000000000000000000000000000000.0
224000000000000000000000000000000000000000000.0
250000000000000000000000000000000000000000000.0
280000000000000000000000000000000000000000000.0
320000000000000000000000000000000000000000000.0
360000000000000000000000000000000000000000000.0
400000000000000000000000000000000000000000000.0
450000000000000000000000000000000000000000000.0
500000000000000000000000000000000000000000000.0
560000000000000000000000000000000000000000000.0
630000000000000000000000000000000000000000000.0
710000000000000000000000000000000000000000000.0
800000000000000000000000000000000000000000000.0
900000000000000000000000000000000000000000000.0
1000000000000000000000000000000000000000000000.0
1120000000000000000000000000000000000000000000.0
1250000000000000000000000000000000000000000000.0
1400000000000000000000000000000000000000000000.0
1600000000000000000000000000000000000000000000.0
1800000000000000000000000000000000000000000000.0
2000000000000000000000000000000000000000000000.0
2240000000000000000000000000000000000000000000.0
2500000000000000000000000000000000000000000000.0
2800000000000000000000000000000000000000000000.0
3200000000000000000000000000000000000000000000.0
3600000000000000000000000000000000000000000000.0
4000000000000000000000000000000000000000000000.0
4500000000000000000000000000000000000000000000.0
5000000000000000000000000000000000000000000000.0
5600000000000000000000000000000000000000000000.0
6300000000000000000000000000000000000000000000.0
7100000000000000000000000000000000000000000000.0
8000000000000000000000000000000000000000000000.0
9000000000000000000000000000000000000000000000.0
10000000000000000000000000000000000000000000000.0
11200000000000000000000000000000000000000000000.0
12500000000000000000000000000000000000000000000.0
14000000000000000000000000000000000000000000000.0
16000000000000000000000000000000000000000000000.0
18000000000000000000000000000000000000000000000.0
20000000000000000000000000000000000000000000000.0
22400000000000000000000000000000000000000000000.0
25000000000000000000000000000000000000000000000.0
28000000000000000000000000000000000000000000000.0
32000000000000000000000000000000000000000000000.0
36000000000000000000000000000000000000000000000.0
40000000000000000000000000000000000000000000000.0
45000000000000000000000000000000000000000000000.0
50000000000000000000000000000000000000000000000.0
56000000000000000000000000000000000000000000000.0
63000000000000000000000000000000000000000000000.0
71000000000000000000000000000000000000000000000.0
80000000000000000000000000000000000000000000000.0
90000000000000000000000000000000000000000000000.0
100000000000000000000000000000000000000000000000.0
112000000000000000000000000000000000000000000000.0
125000000000000000000000000000000000000000000000.0
140000000000000000000000000000000000000000000000.0
160000000000000000000000000000000000000000000000.0
180000000000000000000000000000000000000000000000.0
200000000000000000000000000000000000000000000000.0
224000000000000000000000000000000000000000000000.0
250000000000000000000000000000000000000000000000.0
280000000000000000000000000000000000000000000000.0
320000000000000000000000000000000000000000000000.0
360000000000000000000000000000000000000000000000.0
400000000000000000000000000000000000000000000000.0
450000000000000000000000000000000000000000000000.0
500000000000000000000000000000000000000000000000.0
560000000000000000000000000000000000000000000000.0
630000000000000000000000000000000000000000000000.0
710000000000000000000000000000000000000000000000.0
800000000000000000000000000000000000000000000000.0
900000000000000000000000000000000000000000000000.0
1000000000000000000000000000000000000000000000000.0
1120000000000000000000000000000000000000000000000.0
1250000000000000000000000000000000000000000000000.0
1400000000000000000000000000000000000000000000000.0
1600000000000000000000000000000000000000000000000.0
1800000000000000000000000000000000000000000000000.0
2000000000000000000000000000000000000000000000000.0
2240000000000000000000000000000000000000000000000.0
2500000000000000000000000000000000000000000000000.0
2800000000000000000000000000000000000000000000000.0
3200000000000000000000000000000000000000000000000.0
3600000000000000000000000000000000000000000000000.0
4000000000000000000000000000000000000000000000000.0
4500000000000000000000000000000000000000000000000.0
5000000000000000000000000000000000000000000000000.0
5600000000000000000000000000000000000000000000000.0
6300000000000000000000000000000000000000000000000.0
7100000000000000000000000000000000000000000000000.0
8000000000000000000000000000000000000000000000000.0
9000000000000000000000000000000000000000000000000.0
10000000000000000000000000000000000000000000000000.0
11200000000000000000000000000000000000000000000000.0
12500000000000000000000000000000000000000000000000.0
14000000000000000000000000000000000000000000000000.0
16000000000000000000000000000000000000000000000000.0
18000000000000000000000000000000000000000000000000.0
20000000000000000000000000000000000000000000000000.0
22400000000000000000000000000000000000000000000000.0
25000000000000000000000000000000000000000000000000.0
28000000000000000000000000000000000000000000000000.0
32000000000000000000000000000000000000000000000000.0
36000000000000000000000000000000000000000000000000.0
40000000000000000000000000000000000000000000000000.0
45000000000000000000000000000000000000000000000000.0
50000000000000000000000000000000000000000000000000.0
56000000000000000000000000000000000000000000000000.0
63000000000000000000000000000000000000000000000000.0
71000000000000000000000000000000000000000000000000.0
80000000000000000000000000000000000000000000000000.0
90000000000000000000000000000000000000000000000000.0
100000000000000000000000000000000000000000000000000.0
112000000000000000000000000000000000000000000000000.0
125000000000000000000000000000000000000000000000000.0
140000000000000000000000000000000000000000000000000.0
160000000000000000000000000000000000000000000000000.0
180000000000000000000000000000000000000000000000000.0
200000000000000000000000000000000000000000000000000.0
224000000000000000000000000000000000000000000000000.0
250000000000000000000000000000000000000000000000000.0
280000000000000000000000000000000000000000000000000.0
320000000000000000000000000000000000000000000000000.0
360000000000000000000000000000000000000000000000000.0
400000000000000000000000000000000000000000000000000.0
450000000000000000000000000000000000000000000000000.0
500000000000000000000000000000000000000000000000000.0
560000000000000000000000000000000000000000000000000.0
630000000000000000000000000000000000000000000000000.0
710000000000000000000000000000000000000000000000000.0
800000000000000000000000000000000000000000000000000.0
900000000000000000000000000000000000000000000000000.0
1000000000000000000000000000000000000000000000000000.0
1120000000000000000000000000000000000000000000000000.0
1250000000000000000000000000000000000000000000000000.0
1400000000000000000000000000000000000000000000000000.0
1600000000000000000000000000000000000000000000000000.0
1800000000000000000000000000000000000000000000000000.0
2000000000000000000000000000000000000000000000000000.0
2240000000000000000000000000000000000000000000000000.0
2500000000000000000000000000000000000000000000000000.0
2800000000000000000000000000000000000000000000000000.0
3200000000000000000000000000000000000000000000000000.0
3600000000000000000000000000000000000000000000000000.0
4000000000000000000000000000000000000000000000000000.0
4500000000000000000000000000000000000000000000000000.0
5000000000000000000000000000000000000000000000000000.0
5600000000000000000000000000000000000000000000000000.0
6300000000000000000000000000000000000000000000000000.0
7100000000000000000000000000000000000000000000000000.0
8000000000000000000000000000000000000000000000000000.0
9000000000000000000000000000000000000000000000000000.0
10000000000000000000000000000000000000000000000000000.0
11200000000000000000000000000000000000000000000000000.0
12500000000000000000000000000000000000000000000000000.0
14000000000000000000000000000000000000000000000000000.0
16000000000000000000000000000000000000000000000000000.0
18000000000000000000000000000000000000000000000000000.0
20000000000000000000000000000000000000000000000000000.0
22400000000000000000000000000000000000000000000000000.0
25000000000000000000000000000000000000000000000000000.0
28000000000000000000000000000000000000000000000000000.0
32000000000000000000000000000000000000000000000000000.0
36000000000000000000000000000000000000000000000000000.0
40000000000000000000000000000000000000000000000000000.0
45000000000000000000000000000000000000000000000000000.0
50000000000000000000000000000000000000000000000000000.0
56000000000000000000000000000000000000000000000000000.0
63000000000000000000000000000000000000000000000000000.0
71000000000000000000000000000000000000000000000000000.0
80000000000000000000000000000000000000000000000000000.0
90000000000000000000000000000000000000000000000000000.0
100000000000000000000000000000000000000000000000000000.0
112000000000000000000000000000000000000000000000000000.0
125000000000000000000000000000000000000000000000000000.0
140000000000000000000000000000000000000000000000000000.0
160000000000000000000000000000000000000000000000000000.0
180000000000000000000000000000000000000000000000000000.0
200000000000000000000000000000000000000000000000000000.0
224000000000000000000000000000000000000000000000000000.0
250000000000000000000000000000000000000000000000000000.0
280000000000000000000000000000000000000000000000000000.0
320000000000000000000000000000000000000000000000000000.0
360000000000000000000000000000000000000000000000000000.0
400000000000000000000000000000000000000000000000000000.0
450000000000000000000000000000000000000000000000000000.0
500000000000000000000000000000000000000000000000000000.0
560000000000000000000000000000000000000000000000000000.0
630000000000000000000000000000000000000000000000000000.0
710000000000000000000000000000000000000000000000000000.0
800000000000000000000000000000000000000000000000000000.0
900000000000000000000000000000000000000000000000000000.0
1000000000000000000000000000000000000000000000000000000.0
1120000000000000000000000000000000000000000000000000000.0
1250000000000000000000000000000000000000000000000000000.0
1400000000000000000000000000000000000000000000000000000.0
1600000000000000000000000000000000000000000000000000000.0
1800000000000000000000000000000000000000000000000000000.0
2000000000000000000000000000000000000000000000000000000.0
2240000000000000000000000000000000000000000000000000000.0
2500000000000000000000000000000000000000000000000000000.0
2800000000000000000000000000000000000000000000000000000.0
3200000000000000000000000000000000000000000000000000000.0
3600000000000000000000000000000000000000000000000000000.0
4000000000000000000000000000000000000000000000000000000.0
4500000000000000000000000000000000000000000000000000000.0
5000000000000000000000000000000000000000000000000000000.0
5600000000000000000000000000000000000000000000000000000.0
6300000000000000000000000000000000000000000000000000000.0
7100000000000000000000000000000000000000000000000000000.0
8000000000000000000000000000000000000000000000000000000.0
9000000000000000000000000000000000000000000000000000000.0
10000000000000000000000000000000000000000000000000000000.0
11200000000000000000000000000000000000000000000000000000.0
12500000000000000000000000000000000000000000000000000000.0
14000000000000000000000000000000000000000000000000000000.0
16000000000000000000000000000000000000000000000000000000.0
18000000000000000000000000000000000000000000000000000000.0
20000000000000000000000000000000000000000000000000000000.0
22400000000000000000000000000000000000000000000000000000.0
25000000000000000000000000000000000000000000000000000000.0
28000000000000000000000000000000000000000000000000000000.0
32000000000000000000000000000000000000000000000000000000.0
36000000000000000000000000000000000000000000000000000000.0
40000000000000000000000000000000000000000000000000000000.0
45000000000000000000000000000000000000000000000000000000.0
50000000000000000000000000000000000000000000000000000000.0
56000000000000000000000000000000000000000000000000000000.0
63000000000000000000000000000000000000000000000000000000.0
71000000000000000000000000000000000000000000000000000000.0
80000000000000000000000000000000000000000000000000000000.0
90000000000000000000000000000000000000000000000000000000.0
100000000000000000000000000000000000000000000000000000000.0
112000000000000000000000000000000000000000000000000000000.0
125000000000000000000000000000000000000000000000000000000.0
140000000000000000000000000000000000000000000000000000000.0
160000000000000000000000000000000000000000000000000000000.0
180000000000000000000000000000000000000000000000000000000.0
200000000000000000000000000000000000000000000000000000000.0
224000000000000000000000000000000000000000000000000000000.0
250000000000000000000000000000000000000000000000000000000.0
280000000000000000000000000000000000000000000000000000000.0
320000000000000000000000000000000000000000000000000000000.0
360000000000000000000000000000000000000000000000000000000.0
400000000000000000000000000000000000000000000000000000000.0
450000000000000000000000000000000000000000000000000000000.0
500000000000000000000000000000000000000000000000000000000.0
560000000000000000000000000000000000000000000000000000000.0
630000000000000000000000000000000000000000000000000000000.0
710000000000000000000000000000000000000000000000000000000.0
800000000000000000000000000000000000000000000000000000000.0
900000000000000000000000000000000000000000000000000000000.0
1000000000000000000000000000000000000000000000000000000000.0
1120000000000000000000000000000000000000000000000000000000.0
1250000000000000000000000000000000000000000000000000000000.0
1400000000000000000000000000000000000000000000000000000000.0
1600000000000000000000000000000000000000000000000000000000.0
1800000000000000000000000000000000000000000000000000000000.0
2000000000000000000000000000000000000000000000000000000000.0
2240000000000000000000000000000000000000000000000000000000.0
2500000000000000000000000000000000000000000000000000000000.0
2800000000000000000000000000000000000000000000000000000000.0
3200000000000000000000000000000000000000000000000000000000.0
3600000000000000000000000000000000000000000000000000000000.0
4000000000000000000000000000000000000000000000000000000000.0
4500000000000000000000000000000000000000000000000000000000.0
5000000000000000000000000000000000000000000000000000000000.0
5600000000000000000000000000000000000000000000000000000000.0
6300000000000000000000000000000000000000000000000000000000.0
7100000000000000000000000000000000000000000000000000000000.0
8000000000000000000000000000000000000000000000000000000000.0
9000000000000000000000000000000000000000000000000000000000.0
10000000000000000000000000000000000000000000000000000000000.0
11200000000000000000000000000000000000000000000000000000000.0
12500000000000000000000000000000000000000000000000000000000.0
14000000000000000000000000000000000000000000000000000000000.0
16000000000000000000000000000000000000000000000000000000000.0
18000000000000000000000000000000000000000000000000000000000.0
20000000000000000000000000000000000000000000000000000000000.0
22400000000000000000000000000000000000000000000000000000000.0
25000000000000000000000000000000000000000000000000000000000.0
28000000000000000000000000000000000000000000000000000000000.0
32000000000000000000000000000000000000000000000000000000000.0
36000000000000000000000000000000000000000000000000000000000.0
40000000000000000000000000000000000000000000000000000000000.0
45000000000000000000000000000000000000000000000000000000000.0
50000000000000000000000000000000000000000000000000000000000.0
56000000000000000000000000000000000000000000000000000000000.0
63000000000000000000000000000000000000000000000000000000000.0
71000000000000000000000000000000000000000000000000000000000.0
80000000000000000000000000000000000000000000000000000000000.0
90000000000000000000000000000000000000000000000000000000000.0
100000000000000000000000000000000000000000000000000000000000.0
112000000000000000000000000000000000000000000000000000000000.0
125000000000000000000000000000000000000000000000000000000000.0
140000000000000000000000000000000000000000000000000000000000.0
160000000000000000000000000000000000000000000000000000000000.0
180000000000000000000000000000000000000000000000000000000000.0
200000000000000000000000000000000000000000000000000000000000.0
224000000000000000000000000000000000000000000000000000000000.0
250000000000000000000000000000000000000000000000000000000000.

PINE, S., Honeywell, Inc., STEINHEISER, Rick., U.S. Army Research Institute, and KOCH, Chris., Honeywell, Inc., Minneapolis, Minnesota.

DEVELOPMENT OF A COMPUTER-BASED SKILL QUALIFICATION TEST (Tue A.M.)

The Skill Qualification Test (SQT) is the primary means by which the US Army evaluates the proficiency with which soldiers can perform the critical tasks in their job specialty. The SQT has both written and "hands-on" components. The present research developed an SQT for tactical computer operators in which the test was administered and scored via the actual operational tactical system. As a result, the soldier can take the test while performing actual job tasks on the actual (not simulated) operational equipment. Because the entire test is computerized, soldiers and their supervisors can receive immediate hard copy of test results, item by item. The CAI language used is PLANIT, which is machine transportable.

Benefits include increased test security, ease in making changes in the test, high fidelity, and hard copy of item analysis for immediate feedback.

Subsequent work will develop the embedded testing concept for air-defense missile control, and will also explore how tests might be "adaptively tailored" on an individual by individual basis.

Christopher G. Koch

Steven M. Pine

Honeywell, Inc.
Systems and Research Center
2600 Ridgway Parkway
Minneapolis, MN 55413

Frederick H. Steinheiser, Jr.

Army Research Institute for
the Behavioral and Social Sciences
Alexandria, VA 22333

Based upon work performed by Honeywell, Inc., under Contract MDA 903-79-C-0386.

The opinions expressed are those of the authors, and do not necessarily imply endorsement of the U.S. Army or the Dept. of Defense.

DEVELOPMENT OF A COMPUTER-BASED SKILL QUALIFICATION TEST

1.0 INTRODUCTION

1.1 Skill Qualification Test

The Skill Qualification Test (SQT) is a major diagnostic tool for the U.S. Army individual training system (TRADOC Reg 351-2). It contains performance-oriented and criterion-referenced tests on critical tasks selected from the Soldier's Manual. Test results feedback is provided to soldiers, training managers, and training/test developers. The SQT usually contains a written Skill Component (SC), a Hands-On Component (HOC), and a Job Site Component (JSC).

Although the SQT program was designed to maximize hands-on performance testing, problems of standardization, scoring reliability, extensive time and resource requirements, training of test administrators, and overall administrative feasibility have attenuated its success. Much skill qualification testing has become primarily written in nature, via the SC or a written Alternate HOC. The written tests allow for standardized administration and scoring, but often have low fidelity to actual job performance requirements.

One promising alternative to present SQTs for tactical data systems is the concept of a computer-based SQT administered (embedded) on the operational system or simulator to accomplish standardized, objective, hands-on skill qualification assessment. Once the feasibility of a computer-based SQT is demonstrated, advanced psychometric models of Computerized Adaptive Testing (CAT) can be investigated as vehicles for administering the SQT.

1.2 Application

1.2.1 TACFIRE--The testbed selected for development of an embedded SQT testing concept is the TACFIRE system. TACFIRE is a computerized tactical fire direction system designed to coordinate the command, control, and communications (C^3) functions among all levels of the Army field artillery system, from corps to forward observers. The Artillery Control Console Operator (ACCO), Military Occupational Specialty (MOS) 13C, is one element of the TACFIRE system whose functions include entering data required for artillery target intelligence, fire planning, tactical and technical fire control, and commander's criteria; controlling communications among the network of TACFIRE subscribers; and initiating, processing, and/or terminating artillery fire missions. The ACCO is a level E6 or E7 NCO who operates a specialized computer terminal (Figure 1) linked with the TACFIRE AN/GYK-12 computer.

1.2.2 PLANIT--The PLANIT (Programming Language for Interactive Teaching) computer software system and authoring language was selected to deliver the TACFIRE SQT. The PLANIT operating system is machine transportable and has been installed on TACFIRE equipment, as well as on commercial computers.

The PLANIT language follows a frame-oriented, self-paced instructional strategy useful for the preparation and administration of training and testing scenarios via computer terminal. The lesson author can arrange branching strategies that route the trainee or examinee through series of frames in an individualized manner contingent

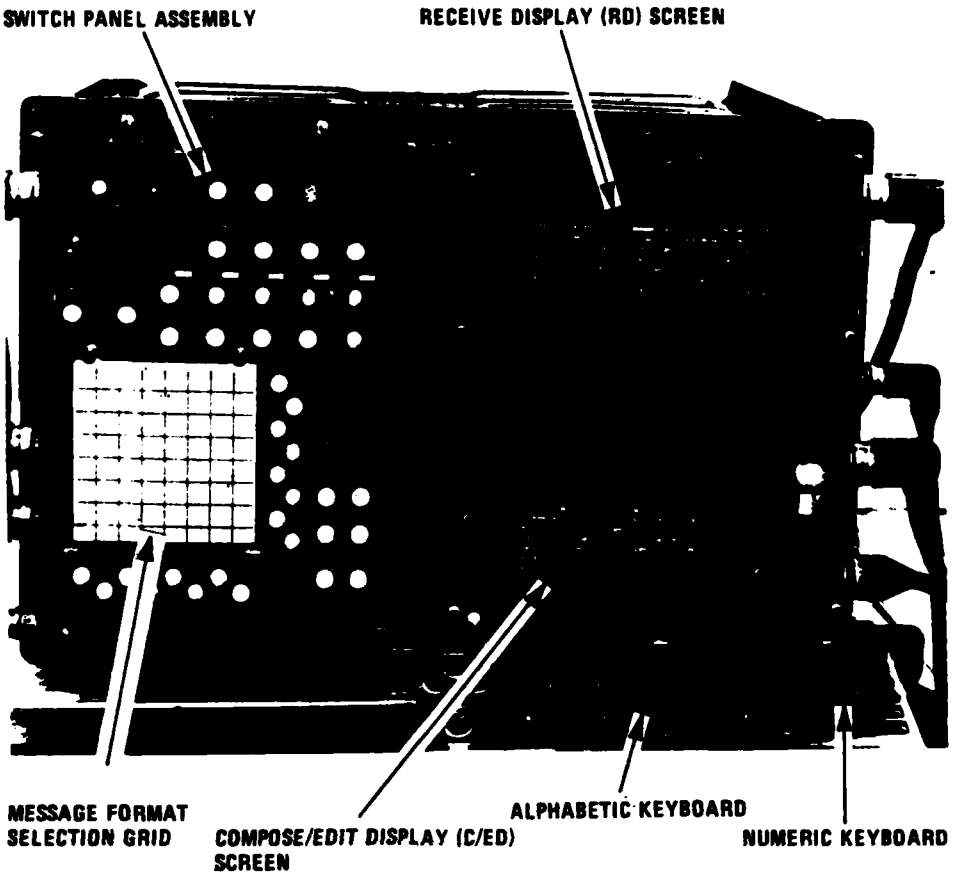


Figure 1. TACFIRE Artillery Control Console

on previous responses. Courseware is constructed either interactively or off line, and the author may test or edit a lesson at any time. PLANIT automatically compiles a performance record of each user's progress through the lessons.

PLANIT is currently operating at Army Field Artillery Schools at Forts Hood, Leavenworth, Benning, Sill, and Riley. It has been used for remedial instruction (e.g., high school arithmetic); however, lessons in the use of TACFIRE equipment comprise the most extensive set of PLANIT courseware developed to date. For the TACFIRE application, an "Enhancements" package for controlling the functions of the ACC has been appended to the PLANIT operating system.

2.0 TECHNICAL APPROACH AND RESULTS OF PHASE I: DEVELOPMENT OF COMPUTER-BASED SQT

2.1 Data Requirements for Hands-On Component Development

The Hands-On Component (HOC) of the SQT should require soldiers to actually perform job tasks under highly standardized conditions. This objective can be realized by means of PLANIT courseware which supplants the TACFIRE operating system software. Therefore, development of an HOC became the focus of the computer-based SQT concept.

The acquisition or generation of task analysis data which determine the components of job performance duties was the intial step in the test development process. A task within a designated job classification, such as MOS 13C, may be represented as a cluster of related actions which are observable, measurable, relatively independent of other actions, and have a beginning and ending point. Many of the required task data for this MOS are being prepared by the Directorate of Training Developments in a format which provides task conditions, standards, and performance measures. These task descriptions comprise the TACFIRE Soldier's Manuals for Skill Level 3 (Battalion Fire Direction Center) and Skill Level 4 (Division Artillery Fire Direction Center).

2.1.1 Task Selection--Since it is not practical, or necessary, to test all tasks for which a particular MOS is responsible, a task selection was required. The selection of tasks was based on the expert judgment of personnel at the Field Artillery School at Fort Sill, Oklahoma.

First, the domain of tasks for MOS 13C was identified to be the content of the Soldier's Manuals for Skill Levels 3 and 4. The size of the task domain was reduced by eliminating tasks which could not be tested on the TACFIRE ACC or which are done well by most soldiers. The remaining tasks were judged according to known performance deficiencies, criticality for successful job performance, frequency of occurrence, and representativeness. A set of 28 ACCO tasks at Skill Level (SL) 3 and 24 tasks at SL 4, representing about half of the task domain, was determined in this manner.

2.1.2 Task Performance Data--Once the set of critical tasks was selected for testing, data required to program task actions and behavior into computerized test items were needed. The computerized test imposes fewer limits on the examinee's actions than a written multiple-choice test item would, and consequently must account for all possible alternative action paths. Detailed information about TACFIRE ACCO procedures and the consequences of task actions was obtained from technical manuals, Job Performance Manuals, printouts from sample task performance, and observation of experienced ACC Operators performing the tasks.

2.2 Installation of PLANIT

One approach to programming the embedded SQT materials is to use the TACFIRE computer console, since this is the device on which the SQT will ultimately be administered. Certain disadvantages to this approach are evident:

- o TACFIRE operational or training sets are not generally available, nor are they cost-effective, for developmental work.
- o A 7-line screen size limitation on the ACC compounds the difficulty of author editing since PLANIT frame size averages 15-20 lines.

Consequently, the PLANIT Version 4.2 operating system was installed on Data General Eclipse S/200 16-bit computers in the Honeywell Man-Machine Sciences Computer Lab to permit on-line programming.

Remote test development on a commercial computer is possible because the PLANIT system operates in a manner identical to that on the TACFIRE computer. Although authoring and editing on the commercial terminal are facilitated, one important limitation is that TACFIRE hardware is not present. Switch assembly actions which would normally occur on TACFIRE cannot be made or observed directly. Thus, Enhanced PLANIT commands intended to operate the ACC are not exercised by the commercial computer, but merely printed out.

With the installation of PLANIT, the capability to develop and debug test items was established. Program activity was then focused on the design, i.e., format, scoring, feedback, and organization, of the SQT.

2.3 Development of PLANIT-Based Hands-On Component

The final product of the development process is a validated Enhanced PLANIT program which administers and scores the HOC. When loaded on the TACFIRE system, the program monitors and controls all the functions of the ACC, permitting assessment of soldiers performing critical job tasks exactly as they would be performed on the operational TACFIRE system. Table 1 summarizes the major characteristics of the embedded HOC.

2.3.1 Organization of Hands-On Tests--Tasks that operationally define job performance in a particular MOS are generally of unequal complexity, difficulty, and duration. Thus, tasks selected for SQT testing must be organized into logical units for testing and scoring purposes. Such units are termed hands-on tests and consist of a set of performance measures that are scored either pass or fail. In the context of the computerized HOC, the hands-on tests are based closely on the selected tasks to ensure high correspondence between test content and job tasks. Each hands-on test includes the prerequisite conditions of the task, the standards pertaining to the task, and the performance measures required for successful completion of the task.

Many of the tasks in the selected sample are covered in one hands-on test. More complex or lengthy tasks were divided into two tests, if each resulting segment represented a logical, independent set of actions. The intent of this organization was to devise hands-on tests that are of generally comparable complexity and length. This is desirable for test scoring purposes since the hands-on tests are weighted equally.

A second intent of the organization of tasks into hands-on tests was to reduce redundant testing of similar performance measures. An examinee should not be penalized repeatedly for not knowing how to correctly respond to similar performance measures, nor is it necessary to repeatedly pass similar measures to demonstrate competence. The approach which guided the formation of hands-on tests was to test only unique performance measures, with the qualification that redundant performance measures were retained when they were critical for fidelity or continuity of the task. A set of 37 hands-on tests was derived from the 28 critical tasks for Skill Level 3, and a set of 33 hands-on tests cover the 24 SL 4 tasks.

2.3.2 Format for Instructions and Required Data--The soldier taking the computerized HOC needs to be informed of what to do to complete each hands-on test. The "job" situation for each hands-on test is stated explicitly with text on the top ACC screen.

TABLE 1. CHARACTERISTICS AND REQUIREMENTS OF TACFIRE EMBEDDED SKILL QUALIFICATION TEST

Objectives

- o Assess qualification of Artillery Control Console Operator (ACCO), MOS 13C
- o Identify soldiers qualified for promotion
- o Assess individual and unit operational readiness

Content Domain

- o Operation and use of TACFIRE Artillery Control Console
- o Initialization and maintenance of Battalion and Division Artillery data bases, including system, ammunition and fire unit, support, meteorological, fire mission, fire support element, and artillery target intelligence data
- o Processing of fire missions

Test Population

- o MOS 13C, Grade E6 and E7 (Staff Sergeant, Sergeant 1st Class)
- o Test pipeline--100 to 200 per year
- o Approximately 30% tested at Field Artillery Schools; remainder tested locally in-unit

Conditions of Testing

- o Hands-On Component (HOC) delivered on operational tactical TACFIRE equipment using TACFIRE computer
- o All instructions, required data, and test feedback provided via ACC screens and line printer (i.e., no auxiliary test materials or guidance)
- o Capability of simultaneous testing of soldiers

Test Organization

- o Soldier's Manual tasks organized into hands-on tests
- o Hands-on tests each contain conditions, performance measures (PMs), and standards

Standards for Scoring

- o Each hands-on test scored "GO" or "NO-GO" based on:
  - a. Criterion % of PMs correct, and
  - b. Satisfaction of time limit standard
- o Qualification on SQT based on criterion % of hands-on tests scored "GO"

Test Results Feedback

- o Feedback provided only at end of HOC
- o Individual Soldier's Report lists % "GO"s, qualification standard, personal data, and date
- o Item feedback on line printer details each error committed:
  - a. Each hands-on test in which an error was made
  - b. Each PM in which an error was made
  - c. The correct response or typed entry for each erroneous one

This includes the task title and number, the prerequisite task conditions, stepwise requirements for successfully performing the task, and the standards pertaining to the task in terms of time limit and accuracy.

During the course of the hands-on test, data which in the authentic operating system are normally obtained through incoming messages on the ACC screens are presented in the same manner. However, many of the tasks in the TACFIRE HOC require that auxiliary data of various kinds (e.g., Executive Officer's Reports, map coordinates) be provided to the examinee. Options considered for presenting these data included a test data workbook, written reports to be handed to the examinee, the ACC Electronic Line Printer (ELP), or the ACC screens. The alternative chosen for the computerized SQT is to print out all auxiliary reports and data on the ELP. This allows the examinee to refer to any auxiliary data as required, and provides a hard-copy record that can be used later to review test performance.

One additonal source of auxiliary information is the ACCO's Fire Direction Officer (FDO). The FDO is responsible for supervising various ACCO activities and providing necessary authorization. In order to ensure a self-contained SQT, the presence of the FDO is simulated during the test by means of messages appearing on the lower ACC screen and copied on the ELP. Such messages are used to signify the FDO's voice authorization for the operator's next action but do not identify what the correct action is.

2.3.3 Test Scoring--Conventions for test scoring and criteria of successful task performance on the computerized SQT were developed to maximize compatibility with existing SQTs. However, the benefit of flexibility has been maintained--strategies for test scoring and assignment of performance criteria are simple to modify by virture of PLANIT's performance recordkeeping capabilities.

Each hands-on test in the computerized HOC is scored as "GO" or "NO-GO". The criteria for scoring "GO" on an SU are:

a) 100% of performance measures completed without error, and

b) test completed within the given time limit standard.

The overall test results would be reported to the examinee in terms of the percent of hands-on tests on which a "GO" was scored. A given prior percentage (e.g., 60%) would qualify the examinee to be eligible for promotion to the next skill level.

A significant advantage of PLANIT performance recordkeeping over manual scoring is the ease and reliability of collecting both product and process performance data. Examinee's actions are all recorded automatically without the participation of any SQT scorer. Thus, any appropriate combination of performance data can be tabulated by the computer for each hands-on test:

- o actions performed, single or multiple
- o sequence of performance of actions
- o end results of the actions performed
- o elapsed time.

The elimination of both the requirement for SQT scorers and the extensive training needed to achieve inter-scorer reliability permits substantial manpower and cost savings for the computer-based SQT.

2.3.4 Performance Feedback--One undesirable characteristic of conventional SQTs is the sometimes long delay between taking the SQT and obtaining test results. Furthermore, examinees are not typically given feedback detailed enough to help them diagnose their errors. Without more specific, timely performance feedback, the SQT is not a training experience since the examinees are given no opportunity to reinforce correct performance or to learn from their errors.

The computerized, embedded SQT is capable of providing detailed, immediate performance feedback. In the TACFIRE HOC a detailed Individual Soldier's Report is provided after the soldier completes all hands-on tests. No feedback is given during the HOC. In this way the soldier is not discouraged from finishing the HOC if he or she commits an error, and the independence of test items is maintained. At the end of the test, the Individual Soldier's Report is printed on the ELP for each examinee. The report summarizes the examinee's performance on the test just completed by listing the number of "GO" and "NO-GO" scores obtained and the hands-on test numbers where each were obtained. Then, feedback on errors committed (if any) are listed for each hands-on test.

2.3.5 Administrating the HOC--The HOC is developed to be a turnkey operation for SQT test administrators such that no programming experience is required to prepare and administer the HOC. The first step is to load the tapes containing the PLANIT operating system and the SQT materials into the TACFIRE computer. The test administrator then sits at the ACC, types in an authorized I.D. code, and executes a routine to prepare the SQT for administration. The routine operates in a self-explanatory fashion by means of menu selections. The following procedures are included:

- Enter identification data (unit number, Test Control Officer number).
- Obtain printout of complete list of hands-on tests.
- Select hands-on tests to be administered.
- Obtain printout of summarized initialization data.
- Try out selected hands-on tests.
- Perform test performance results tabulation.

To execute the HOC for examinees, an authorized student I.D. code is typed at the ACC. The HOC then begins, proceeding through the selected hands-on tests and performing scoring automatically and with no intervention from the test administrator.

3.0 TECHNICAL APPROACH AND RESULTS OF PHASE II: CAT IMPLEMENTATION OF EMBEDDED SQT

The concept of embedded testing is expected to greatly improve the way the Army conducts its SQTs for MOS 13C and similar MOSs. Recent research on Computerized Adaptive Testing (CAT) suggests that further improvements to SQT testing in the Army can be obtained by the application of CAT. Computerized Adaptive Testing is a generic name for certain models and techniques, administered by computer, for sequentially presenting test items to an examinee contingent on his/her previous item responses and ability level. Based on what is presently known about CAT, its application may further enhance the advantages of embedded SQTs over existing methods by: reducing testing time, making it more difficult to compromise the SQT, improving test scoring, and increasing the comparability of scores. The objective of this phase of the effort was to examine the models and techniques of the developing field of CAT and to determine the feasibility and desirability of using CAT to administer an embedded SQT.

3.1 Literature Review of CAT Models

The initial step in this endeavor was to carry out an extensive literature search of CAT models to identify and characterize those models which might be appropriate for the TACFIRE SQT application. Fifty-two studies describing 18 CAT models were reviewed. Each model was then carefully described and compared to the requirements of the TACFIRE SQT. Based on the comparisons, recommendations on the implementation of CAT were formulated.

Table 2 lists the 18 CAT models and summarizes the major characteristics of each in terms of its item selection model, the response format used, and the content domains in which it has been applied. The models can be divided into two broad categories: Item Characteristic Curve (ICC) models (Models 1-10, 14, 15, and 16), and non-ICC models (Models 11-13, 17, and 18).

The critical issue in question was how well each of these models, and class of models, satisfy the specific requirements of the TACFIRE SQT application. A second issue was what limitations, if any, the PLANIT language imposes in programming these models. To address these issues, each of the 18 models was evaluated with respect to the major potential limiting characteristics of SQT, TACFIRE, and PLANIT.

3.2 Evaluation of Models

There are three primary characteristics of the SQT Hands-On Component (HOC) that potentially restrict the choice of CAT models:

1. It is a performance test.
2. It is task-based.
3. The purpose of the HOC is to make qualification decisions about each examinee.

The potential constraints imposed by the characteristics of the TACFIRE system and PLANIT are:

1. The SQT pipeline size for MOS 13C would be 100-200 soldiers per year.

TABLE 2. CHARACTEPISTICS OF CAT MODELS REVIEWED

Model Number	Index Number	Item Selection Model	Response Format	Content Domain
1	4, 27	Two-Stage	Multiple Choice	Verbal Ability
2	12	Three-Stage (Fisher)	Matrix Completion	Mathematical Ability
3	33	Flexilevel	NA	Course Achievement
4	23	Pyramidal	Multiple Choice	Vocabulary Ability
5	1, 2, 28, 36, 46, 47	Stradaptive	Multiple Choice	Course Achievement, Vocabulary Ability
6	17, 30, 32, 34, 42, 51	Bayesian	Multiple Choice	Vocabulary Ability, Arithmetic Reasoning
7	21, 22	Adaptive Mastery Testing (AMT)	Multiple Choice	Course Achievement
8	48, 17	Maximum Information	Multiple Choice	Course Achievement Verbal Ability
9	24, 42, 10	Maximum Likelihood	Multiple Choice	Vocabulary Ability Arithmetic Reasoning Course Achievement
10	20	Empirical Bayesian (Confidence Interval Reduction)	NA	Vocabulary Ability
11	21	Sequential Probability Ratio Test (SPRT) (Wald)	NA	Mastery Testing
12	18	Sequential Consistency (Kalisch)	Multiple Choice	Course Achievement
13	11	Hierarchical Branching (Ferguson)	Free Response	Mathematics Ability
14	5	Multidimensional Maximum Information (Brown & Weiss)	Multiple Choice	Course Achievement
15	41	Compensatory Multi-dimensional (Sympson)	NA	NA
16	44, 45	Multiple Correlation (Urry)	NA	Ability
17	7, 8	Implied Orders (Cliff)	Free Response	Vocabulary Ability
18	43	Confidence Branching	Confidence Marking	Course Achievement

2. Tasks selected for the SQT HOC are comprised of hands-on tests which may not be independent.

3. The TACFIRE AN/GYK-12 computer uses single-precision arithmetic without floating point capabilities. Maximum integer size is 10^9 and maximum decimal size is four significant digits to the right of the decimal point.

4. The PLANIT system installation on TACFIRE has supplied somewhat limiting parameters for maximum matrix size, and the PLANIT language does not contain certain functions used in CAT procedures.

Table 3 summarizes the comparison of CAT models to these restrictions and limitations. A detailed decussion of this analysis is contained in the Honeywell Final Report (Koch, Pine, & Kingsbury, 1980).

3.3 Summary of CAT for Performance Testing

Based on the present review of the developing technology of Computerized Adaptive Testing, it is not yet clear how well CAT principles generalize to the domain of TACFIRE performance testing.

This study examined two broad classes of CAT models for possible application to TACFIRE performance testing: Item Characteristic Curve (ICC) models and non-ICC models. Most of the existing models are based on ICC theory and were originally developed for testing unidimensional ability traits, not task-based performance. Several of the basic principles underlying ICC models appear inappropriate for this application. These include the postulated existence of a unidimensional, underlying trait (in this case a performance trait), and the principle of "ability matching." Ability matching refers to the guiding principle of practically all adaptive testing strategies, whereby an attempt is made to match the difficulty of test questions to the current estimate of the examinee's ability level. This principle is of questionable validity for performance testing since the key issue here is not how difficult a task the examinee can perform, but whether he/she can perform certain critical tasks. In addition, ICC models require extensive data to estimate model parameters. Using a one-parameter model and/or an adaptive calibration procedure could greatly reduce these data requirements, however.

Several of the non-ICC models appear to offer an attractive alternative. These may implicitly suffer from similar shortcomings, however, and are not yet as fully developed or tested. It appears, therefore, that additional research is needed before CAT can be recommended for a fielded SQT performance test. Such research is warranted because of important potential benefits of applying CAT techniques.

TABLE 3. COMPATIBILITY OF CAT MODELS TO TACFIRE SQT CHARACTERISTICS

Characteristics	CAT Model																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
SQT																		
1. Performance Test	O		O	O	O						●	●	●				●	
2. Task Based	O	O		O	O		●				O	●	●	●	●	●	O	
3. Qualification Test	O	O	O	O	O	O	●	O	O	O	●	●	●	O	O	O	O	
TACFIRE/ PLANIT																		
4. Small Test Pipeline (100 to 300 per year)	●	●	●	●	O						●	●	●				●	●
5. Test Items Dependent	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
6. TACFIRE Computer Precision Limitations	●	●	●	●	●												●	●
7. PLANIT Limitations	●	●	●	●	●						●	●	●				O	●

● Compatible

O Could readily be made compatible or has been successfully used in a context similar to the SQT

4.0 FUTURE RESEARCH AND DEVELOPMENT

4.1 TACFIRE SQT

Findings obtained in this contractual effort offer the potential for use in several ways. The TACFIRE Artillery Console Operator (MOS 13C) is a new MOS for which final evaluation of trainees' job proficiency is not yet being conducted. To meet the requirement for such evaluation, an embedded computer-based SQT can be implemented for MOS 13C. Many of the tasks necessary for this implementation have been accomplished in the present embedded SQT concept demonstration effort, including: 1) selection of criterion-related and content-valid critical tasks, 2) development of test items, and 3) development of test scoring procedures.

4.2 Other Embedded Testing Applications

Another use of the findings of this effort is an investigation of other applications for embedded SQT testing. Many Army operational tactical computers are capable of delivering high-fidelity testing scenarios through which trainees' job proficiency on critical tasks can be effectively evaluated. The benefits of embedded SQTs possible for TACFIRE apply for any tactical computer system. These include:

- Performance testing is conducted on the operational tactical data system, ensuring content validity and maximum test fidelity.
- The control, monitoring, and scoring of the test is accomplished by the tactical computer and without the need for any changes to the system software.
- Immediate performance feedback and test results are available for soldiers and supervisory personnel.

4.3 CAT for Performance Testing

Benefits possible through the implementation of CAT techniques are significant in terms of testing efficiency, testing accuracy, and reduced chance of test compromise. However, it is not yet clear how well CAT principles generalize to the domain of TACFIRE performance testing.

The findings of the CAT feasibility study indicate the need for additional basic research before these techniques can be recommended for computer-based performance testing. A starting point would be to empirically examine the functional dependence and dimensionality of the SQT items by performing a correlational and factor analysis of test data. Results of this study should provide answers to several key issues: can a unidimensional model be assumed; is the item independence assumption valid? Research should also be undertaken to develop several new models designed specifically for performance testing. These models along with the more promising existing models should then be field validated to determine their costs and benefits relative to non-adaptive computerized techniques.

(748)

REFERENCES

Koch, C. G., Pine, S. M., & Kingsbury, G. G. "Development of a Computer-Based Skill Qualification Test: Final Report." Report 80SRC52, Honeywell, Inc., Systems and Research Center, Prepared for Army Research Institute, July 1980.

TRADOC Regulation No. 351-2, "Skill Qualification Test (SQT) Policy and Procedures," Department of the Army, Headquarters, U.S. Army Training and Doctrine Command, 21 April 1980.

POTTER, LCdr. Earl H., Department of Humanities, U.S Coast Guard Academy, New London, Connecticut.

SOME FACTORS WHICH LIMIT THE PREDICTABILITY OF EMPLOYEE PERFORMANCE
(Thu A.M)

Attempts to identify variables which successfully predict the future performance of military and civilian government employees have met with at best, modest success. In commenting on the observed moderate correlations between predictor and criterion variables authors often suggest that the "best predictor" remains to be found. An alternative explanation is that situational moderators such as the quality of leadership powerfully affect the level of performance achieved by employees. This paper discusses several factors which limit the predictability of employee performance. Examples are drawn from two studies of U.S. Coast Guard personnel in which stress between leader and subordinate moderates the subordinate's ability to utilize his native talent to accomplish assigned tasks.

Some Factors Which Limit the Predictability of Employee Performance

Earl H. Potter III

United States Coast Guard Academy

New London, Connecticut 06320

Abstract

Attempts to identify variables which successfully predict the future performance of military and civilian government employees have met with, at best, modest success. In commenting on the observed moderate correlations between predictor and criterion variables authors often suggest that the "best predictor" remains to be found. An alternative explanation is that situational moderators such as the quality of leadership powerfully affect the level of performance achieved by employees. This paper discusses several factors which limit the predictability of employee performance. Examples are drawn from three studies of U. S. Coast Guard personnel in which stress between leader and subordinate moderates the subordinate's ability to utilize his native talent to accomplish assigned tasks.

When ten people apply for one job it is to the obvious benefit of the employer to select the one person who will be the most effective on the job. In some cases, the hiring of an air traffic controller, for instance, it is critical that a person prone to ineffective performance not be hired. The surest method of selection would be a behavioral exam-like the swimming test - you throw the applicant in the water and see if he can swim. While this suggestion may seem faceitious, organizations in fact do this when they use attrition in basic level training and entry level jobs to screen out unsuitable employees. The costs of such a selection process are, however, very high - a fact which the General Accounting Office has noted with respect to the service academics. To cut these costs psychologists have attempted to develop sets of predictors which on the one hand can identify persons likely to fail and on the other hand persons most likly to succeed. Various combinations of aptitude, personality and attitude measures for a broad range

of jobs and schools have been developed which, at some acceptable level of statistical significance, predict the future performance of applicants. Yet many of these statistical successes offer small practical gains. Mobley, et al. (1979) developed an inventory of predictors which correlated with attrition from Marine Corps basic training with a multiple R of .30. While this finding represents an improvement over previous efforts to predict attrition, only 97% of the variance in attrition is accounted for. Coleman (1979) describes the development of a new test which predicts Air Traffic Controller performance. The correlation of this test to Air Traffic Controller performance is .26. Coleman estimates that this new test will cut down the number of predicted successes who fail from 35% to 25% of those hired. The widely used Scholastic Aptitude Test has increasingly come under fire with consumer advocate Ralph Nader claiming recently that the test predicts success only 12 percent better than random chance.

In this paper I would like to suggest that the less than satisfying results of our best efforts at prediction are in large part due to what Guion (1976) called situational factors and not soley to the tests themselves. In fact our tests may be more highly correlated with performance than our current methods of test validation reveal. If this is the case, then perhaps our efforts should be directed towards developing a complex paradigm of selection and assignment which reflects the complexity of the interacting factors which determine performance in school and on the job.

Study I

My first example is drawn from a study of 130 U. S. Coast Guard personnel assigned to one of the Coast Guard's 12 district offices. The purpose of this study was to uncover some of the factors which moderate the relationship between intelligence and job performance. While it is generally assumed that intelligent persons will do most jobs better than less intelligent persons, the research relating intelligence to job performance does not support

this assumption. Median correlations between various intelligence measures and job performance ranged between .26 and .30 (Mann, 1959). Considering the fact that most management jobs involve activities such as planning, organizing, co-ordinating and evaluating - which are cognitive functions - it seems strange that cognitive ability should not be more strongly related to performance. One reason that this might not be so is that intelligence does contribute to performance under some conditions but not others (Fiedler & Leister, 1977). One factor which emerges from the literature as a possible moderator of the intillegence performance relationship is stress. Lazarus (1966) reports that anxiety and stress narrow the individual's focus and inhibit creative thought. This could be the reason why Amabile (1979) found that persons who were apprehensive about evaluations of their work performed less well on creative tasks.

In this study the Wonderlic Personnel Test was used as the measure of intelligence. The median score (range 14 to 43) was 28.8 which corresponds to a score of about 113 on the Wechsler-Bellevue Test. The measure of job performance was an 18 item scale (Bons, 1975) completed by the staff member's immediate superior. Performance scores ranged from 33 to 126 with a median of 32.5. Stress was measured by a 34 item scale (Borden & Curtis, 1977) which focused specifically on stress between the staff member and his boss. A 6-item factor accounted for 72% of the variance in the scale and indicated that the relationship between boss and subordinate was most stressful when the boss demanded high performance but did not provide direction or support.

To examine the hypothesis that intelligence contributed to performance under conditions of low stress subjects were split at the median of stress with the boss. The correlation between intelligence and stress for those subjects with low stress with the boss was positive and not significant ($r = .16$, $N = 60$). For those subjects for whom stress with the boss was high

the correlation was negative and statistically significant ($r = -.27$, $N = 48$, $p < .01$).

In an effort to understand why intelligence was not significantly related to performance under conditions of low stress it was hypothesized that experience on the job might also effect the relationship. Those persons who had been in the organization longer may have learned the correct behaviors regardless of their intelligence (remember that we are dealing with a sample whose average intelligemce is above normal). Those persons with less experience may have to rely on their native talent to solve problems to which more experienced people already know the answer. To test this hypothesis subjects were split into high, middle and low experience groups (over 20 years, 10-20 years and under 10 years respectively). For staff members with less than 10 years of experience the relationship of intelligence to performance under conditions of high stress was $-.43$ ($N = 16$, $p < .05$) for low experience staff members with low stress with their boss the intelligence/performance correlation was .73 ($N = 13$, $p < .001$).

Study II

My second example is drawn from a study of 107 cadets enrolled in a required psychology course at the United States Coast Guard Academy. Of the cadets who participated 35 were seniors, 23 juniors, 44 sophomores and 5 freshmen. In this study the measure of intelligence was the mathematical score of the Scholastic Aptitude Test. Scores in this sample ranged from 500 to 780 with a mean of 655.2. Academic performance was indicated by the cadet's cumulative grade point average (GPA). GPA's for this sample ranged from 1.42 to 3.96 with a mean of 2.83. Experience was indicated by the cadet's class. Stress was measured by a set of four question scales which required cadets to reflect on certain significant relationships - those with instructors, company officers, parents, peers and cadets senior to themselves. For each relationship they were asked to indicate on

a seven point scale how much stress they felt in a particular relationship as a result of certain behaviors. An example for senior cadets is "He rates me low in aptitude but he doesn't tell me what I should do in order to improve my grade." This measure of stress corresponds closely to the concept of stress with the boss defined by the Coast Guard staff study. Scores for the several scales could range from 4 to 28. Actual ranges varied from 4 - 21 for stress with parents to 4 - 28 for stress with company officers. Average stress scores ranged from 7.65 for stress with parents to 13.04 for stress with senior cadets.

As in the first study it was by hypothesized that the magnitude of the relationship between intelligence and performance would decrease as stress increased. Furthermore, following the lead of Study I it was hypothisized that this trend would exist only for those cadets with relatively less experience. To test these hypotheses cadets were divided into high, middle and low stress groups. Looking first at junior cadets there is no difference in the SAT-M/GPA relationship for different levels of stress with parents, company officers or instructors. For cadets who reported high stress with peers the relationship of intelligence to performance was not significant ($r = .08$, $N = 15$). For cadets who reported moderate stress this relationship was significant ($r = .45$, $N = 18$, $p < .05$). For cadets who reported low stress with peers the relationship was also significant ($r = .76$, $N = 14$, $p < .001$). The difference between the intelligence/performance correlations for cadets with high and low stress was significant using a Fisher's Z test ($p < .05$). For cadets reporting high, moderate, or low stress with senior cadets the picture was the same: high stress ($r = .04$, $N = 15$, NS), moderate stress ($r = .60$, $N = 19$, $p < .01$), high stress ($r = .76$, $N = 13$, $p < .001$). A Fisher's Z test shows a significant difference between the intelligence/performance relationships for high and low stress with senior cadets ($p < .05$).

Clearly, not all relationships are equally salient to junior cadets. For those cadets who are relatively new to the Academy system their personal relationships in the barracks are most important. If freshmen and sophomore cadets are preoccupied with trying to meet the demands of senior cadets they appear unable to apply their intellectual ability to their academic tasks.

When we look at senior cadets we find that stress with peers does not moderate the intelligence/GPA relationship. For cadets who reported low stress with peers or cadets senior to them the SAT-M/GPA relationship was positive and non-significant ($r = .40$, $N = 21$, $p = .06$) for moderate stress and high stress the SAT-M/performance relationship was positive and significant ($r = .47$, $N = 18$, $p < .05$ and $r = .56$, $N = 19$, $p < .01$ respectively). It would appear that experience counter balances the impact of stress on the cadet's ability to use his intelligence.

Study III

A third study of 115 cadets at the U. S. Coast Guard Academy confirmed the previous finding that stress with peers moderates the SAT/GPA relationship. Under conditions of low peer stress cadets in the high SAT group have an average GPA of 3.12. Cadets in the low SAT group have an average SAT of 2.65 ($p < .05$). Under conditions of high peer stress there is no difference between these two groups in GPA (2.98 and 2.96 respectively).

Discussion

The results of these studies show that both stress and experience moderate the relationship between intelligence and performance. If we were to attempt

750

to predict job performance based solely on the results of an intelligence test, our success measured in terms of a validiy coefficient would be severly limited by the operation of these factors. When I consider the implications of these findings for the process of predicting future performance given a knowledge of an applicant's intelligence, I am struck by several things which I believe have relevance to the prediction process in general. First, as you look over the research on intelligence and performance it is easy to conclude that intelligence is not a very important factor in determining performance on the job. For the Coast Guard staff and which I have discussed the correlation between intelligence and performance is .02! The research evidence is often "confirmed" by our observations that some very bright people fail while less bright people succeed. These findings suggest, however, that we often err when we attempt to validate a predictor by relating it directly to a performance criterion without considering the impact of factors which moderate that relationship. In other words our prediction model is not nearly as complex as the emerging model of job performance. The consequence of this fact is that we underestimate the relevance of cognitive ability to job performance.

The practically minded person might well ask at this point if I would suggest that intelligent persons be assigned only to specific situations which facilitate the use of their intelligence. I might do that if I weren't also aware that people react to stressful situations differently - some coping with tension better than others. Sarason (1979) has discovered that you can identify people who will cope well with stress and that you can train those who don't cope as well to deal with stress more effectively. With this information in hand I now have several options (1) I can assign an intelligent person who also has new employee to a supportive environment where he can gain experience, (2) I may recognize that the position I have open is a stressful one and that I need an intelligent person who also has a high level of coping skills, (3) I can train the new employee to cope more effectively, or (4) I can intervene in the situation to tailor it more to the new employee's needs.

If this is beginning to sound like a description of a human resource management system and not an improved selection test, then I am beginning to make my point. Selection procedures ought to be based on an understanding the performance process. To be fair I have to say that our understanding of the performance process at this point is pretty sketchy; so, the relationship between those who develop selection procedures and those who study the performance process should be interactive. But there should be a relationship. At this time the people who develop selection instruments and the people who study performance are often not aware of each other's problems and progress. For example, there is very little overlap between the attendees at the Military Testing Association meetings and the Symposia on Psychology in DOD.

The consequence of a better understanding of the performance process is likely to be an increase in our ability to identify those personal characteristics which truly contribute to performance. Some of these will contribute directly and others will contribute in interactions. We are also likely to find, however, that maximizing personal assets will not insure employee success. At this point it will become necessary to develop assignment policies which place employees in optimum positions and to develop training programs which not only teach specific job skills but also add to the employee's ability to function in a wide range of job situations. The only alternative which I can see is to continue to place talented people in positions where they can only fail.

(76)

References

Amabile, T. M. Effects of external evaluation on artistic creativity. Journal of Personality and Social Psychology, 1979, 37, 221-233.

Bons, P. M. The effect of changes in leadership environment on the behavior of relationship - and task-motivated leaders. Unpublished doctoral disertation, University of Washington, 1975.

Borden, D. F. & Curtis, W. Personal communication, November 18, 1977.

Coleman, J. G. Prediction of success for airtraffic controllers. Proceedings of the 21st Annual Conference of the Military Testing Association, San Diego, California, 1979.

Fiedler, F. E., & Leister, A. Intelligence and group performance: A multiple screen model. Organizational Behavior and Human Performance, 1977, 20, 1-14.

Guion, R. M. Recruiting, selection, and job replacement. In M. D. Dunette (Ed.), Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally, 1976.

Lazarus, R. S. Psychological Stress and the Coping Process. New York: McGraw-Hill, 1966.

Mann, R. D. A review of the relationships between personality and performance in small groups. Psychological Bulletin, 1959, 56, 241-270.

Mobley, W. H., Hand, H. H., Baker, R. L., & Meglino, B. M. Conceptual and empirical analysis of military recruit training attrition. Journal of Applied Psychology, 1979, 64, 10-18.

Sarason, I. G., Johnson, J. H., Berberich, J. P., and Siegel, J. M. Helping police officers cope with stress: A cognitive-behavioral approach (Report No. SCS-CS-005). Arlington, Virginia: Office of Naval Research, February 1978.

PRESTWOOD, J. Stephen, Assessment Systems Corporation, St. Paul, Minnesota, and University of Minnesota, Minneapolis, Minnesota.

DESIGN OF AN EXAMINEE MONITORING STATION (Thu P.M.)

The state of the art of computerized testing has not yet reached the point where the computer can anticipate and correct everything the examinee can do wrong. An integral part of a computerized testing system is thus an examinee monitoring station. This part of the system is responsible for the administrative duties of initalizing a testing session for an examinee and, more importantly, detecting problems encountered by an examinee, alerting a test proctor, and providing the test proctor a means of communicating with the examinee and correcting the situation. After a general overview of the examinee monitoring function, this paper discusses in detail some methods of problem detection and correction. As well as the obvious problems, methods of detecting subtle problems such as random responding, collaborative responding, and coached responding are discussed. Finally, a general list of features desirable in such a station is presented and discussed.

DESIGN OF AN EXAMINEE MONITORING STATION

J. Stephen Prestwood

Assessment Systems Corporation
St. Paul, Minnesota

and

University of Minnesota
Minneapolis, Minnesota

In a traditional paper-and-pencil testing environment (e.g., a classroom or a military examining station), test administrators and proctors fulfill a number of functions. In addition to the routine functions of distributing tests, reading instructions, and answering questions regarding the instructions, they must also deal with problems which arise during the test. Among the various things an administrator or proctor must monitor during the test are collaboration between two or more examinees, the appearance of random responding (e.g., an examinee may appear to be answering items on an answer sheet without referring to the question booklet), and indications of insufficient examinee motivation for completing the test in an appropriate manner (e.g., an examinee might be looking around the room, or staring out a window). The state of the art of computerized testing has not yet reached the point where the computer can handle all of these monitoring functions without human assistance. The computer can, however, be used to assist an administrator or proctor in handling most of these functions.

Routine Functions

The computer can accomplish most of the routine administrative tasks without much human assistance. In a computerized testing situation, proctors would not normally distribute tests, booklets, or answer sheets. A proctor might, however, be responsible for assigning examinees to computer terminals, and for doing whatever is necessary to initiate the appropriate tests on each terminal. The test instructions would most likely be presented by the computer. Such instructions usually consist of a number of screen presentations ("frames") which present information and then require the examinee to act on that information. If the response is incorrect, the computer can often identify the source of error and deviate from the standard instructional sequence to administer additional frames in order to get around the problem. Intervention by a proctor, however, may be required. If the same mistake is repeated a number of times, the computer can alert the proctor at the examinee monitoring station to the terminal at which an examinee is having problems, the frame on which the problem occurs, and the type of incorrect response being made. A sophisticated system might even allow two-way communication between the examinee and the proctor seated at the monitoring station. The degree of sophistication should reflect the amount of monitoring the proctor is required to do. If only a few testing stations must be monitored, direct person-to-person communication

is probably not desirable. If a large number of stations must be monitored, a more sophisticated system may allow the monitoring to be accomplished with fewer proctors.

To facilitate answering questions the examinees may have about the test or response procedure, each testing terminal can be equipped with a prominent "panic button" which is connected to the monitoring station and with which the examinee can summon the proctor before the test begins or while the test is in progress. Because this panic switch could also be used to summon the proctor in the event of equipment failure, it should operate independently of the computer system administering the test. The switches, for instance, could be directly wired to a light panel which would allow the proctor to immediately identify the source of any request for assistance.

Monitoring Functions

The real value of a computerized examinee monitoring station is realized, not in dealing with routine administration functions, however, but in dealing with the more complex responsibilities of a proctor. Among these responsibilities are the identification of examinees who are cheating (copying from others, collaborating with others, or responding on the basis of a priori coaching by others), who are responding randomly to the test items, who appear to be trying to achieve a low score, or who appear to be having difficulties but who have, for one reason or another, not requested assistance.

Collaboration and Copying

Cheating in the form of collaboration between two or more examinees or cheating which involves one examinee's copying of answers from another can be automatically monitored in a variety of ways if all of the testing terminals are connected to a single computer. It is relatively simple, for instance, for a computer to detect simultaneous responding by two or more examinees at adjacent terminals. Another method for detecting these forms of cheating involves periodically sampling pairs of response records and calculating the probability of observing the incorrect response strings of those individuals. If two people are answering every question correctly, it is very difficult to say that collaboration is involved unless it is actually observed by a proctor. But if two individuals are answering only 70% of the items correctly, one can look at those items which are answered incorrectly. While it is possible that the same 30% of the items will be answered incorrectly by two individuals, it is very unlikely that the same distractors would be chosen on those incorrect items. The probability of observing identical incorrect-response strings can be statistically evaluated using the compound-binominal distribution (Weiss, 1980). If indications of such responding are detected by the computer, the computer can take actions to prevent it from continuing (e.g., varying the order of the items) or it can terminate the tests of the individuals involved, inform the proctor that the tests have been terminated due to probable collaboration, and can indicate the specific terminals involved. The proctor can then take whatever action is appropriate.

DTIC

Although the computer can assist the proctor in detecting such behaviors, it is important to note that monitoring may not be necessary when tests are designed especially for computerized administration. For instance, on adaptively administered tests, examinees may be assigned random entry points into an item pool. With most adaptive testing strategies this simple expedient would almost certainly insure that two examinees would rarely be administered the same items at the same time. One way to further assure that strings of identical items are not administered simultaneously on an adaptive test is to use two parallel item pools for item selection. At any single stage in testing, the computer would randomly choose which item pool to search for the next item to be presented. On a conventional test, where all examinees receive the same items, the order of item presentation could very easily be randomized. Some domain referenced test constructors go one step further and randomize the order in which alternatives are presented for each item. It is possible, however, that changing the order of alternatives will have an effect on the difficulty of the test items. Unfortunately, randomizing the order of test-item presentation can, in some cases, change the difficulty of the test as a whole. This might occur, for instance, on a highly speeded test.

Coached Responding

Coached responding is somewhat more difficult to detect. Coaching can take several forms. The simplist form involves instructing the examinee to memorize the response key. This form of coaching can be dealt with simply by *not* administering conventional, linear tests. If linear tests are to be administered, once again, the order in which the items are presented can be randomized. That will prevent someone from learning a response sequence, such as "ABBBACDAB." If conventional, fixed-order tests are to be administered, one can detect this form of coaching if the performance on a long test suddenly deteriorates. Such deterioriza-tion can reflect sequence errors in a memorized key but can also be due to other factors such as fatigue.

Coaching can take other forms as well. If test security has been lax, examinees may be given an opportunity to study and learn the actual items to be administered on the test. If one is administering linear tests of a fixed length, say less than 70 items, coached responding of this type is possible and can be very effective. If one is administering an adaptive or domain-sampling test from a large item pool, coached responding of this type may be desirable. After all, if we can define the domain for the item pool in such a way that all of the skills to be tested are covered, produce 500 to 600 items to measure that domain, and then coach people (i.e., teach them the answers to those 500 to 600 questions), we will, in fact, be changing the ability level of the examinee. If, however, conventional tests of fixed length are to be administered, procedures for detecting coached responding may be needed. Appropriateness measurement techniques (Levine & Rubin, 1979; Levine & Drasgow, 1980) and measures of person-item fit (e.g., Wright, 1977) may be employed to identify examinees responding above their ability level on items or sets of items within a test. Such procedures require a fair amount of statistical manipulation and will not be described in

detail here. They also usually require an examinee to respond to both coached and uncoached items. One method for decreasing the probable success of this form of coaching is to use facit theory (see Roid & Haladyna, 1980) to construct a number of homologous items which may be interchanged within specific tests. Homologous items are items which are similar in form and content but which differ in subtle ways that alter the correctness of the alternative responses. The two items below are homologous:

1. What is the primary advantage of tricycle landing gear over conventional landing gear?
   a. More ground stability in windy conditions.
   b. Typically longer intervals between tire replacment.
   c. Better handling on rough fields.

2. What is the primary advantage of conventional landing gear over tricycle landing gear?
   a. More ground stability in windy conditions.
   b. Typically longer intervals between tire replacement.
   d. Better handling on rough fields.

Although the items are identical in form and content, the answer to the first is a, whereas the answer to the second is c. An individual coached to respond to the first item may respond incorrectly if presented with the second, while an individual knowledgeable in the area would be more likely to answer either one correctly.

Random Responding

Random responding can be detected quite easily if an item pool has sufficient numbers of easy items. Random responding can then be identified as chance-level responding to the items being administered. That is, correctly answering 25% of a set of 4-alternative multiple-choice items (or 20% of 5-alternative multiple-choice items). If the test or item pool does not have sufficient numbers of very easy items, however, examinees performing at this level may be doing their best. If the test or item pool has a sufficient number of very easy items or if data concerning the examinee's ability (e.g., scores from previously administered subtests) indicate that the examinee has sufficient ability to perform adequately on the test, chance-level responding can be considered unusual. Random responding may also be indicated by appropriateness measurement techniques (Levine & Drasgow, 1980), by inadequate person-item fit (Wright, 1977), by convergence failures for maximum-likelihood scoring algorithms, or by unusually small decreases, item-by-item, in variance estimates computed by sequential Baysian scoring routines.

Other Problems

The latency of item responses (i.e., the elapsed time between the presentation of an item and the entry of a response to the item) is usually monitored during the administration of a computerized test. This information can be used to identify examinees whose rate of response is significantly slower than the large majority of individuals. Extremely slow responding can be indicative of a problem which may invalidate the

examinee's test score. Slow responding might indicate that the test is perceived as being too hard or is perceived as unrelated to its assumed purpose (maybe the wrong test was assigned to the examinee), that the examinee is too bored or too tired to perform at his or her actual level of ability, or that the examinee misunderstands the instructions or is having difficulty reading the items (Weiss, 1980). Slow responding could also indicate a machine malfunction. A set of questions designed to pin-point the reasons for slow response rates could be administered as soon as the rate falls below a specified value. The computer could then determine the probable cause of the delays before the proctor at the monitoring station is alerted. Once the proctor is alerted to the problem, appropriate measures can be taken to deal with the problem.

Station Hardware and Display

A complete examinee monitoring station might consist of a cathode-ray terminal (CRT), a printing terminal, a display panel for summoning help, and possibly a two-way voice communication device. The proctor could request several displays on the CRT screen. The first display would provide general information on the entire examinee group. The amount of information provided would be limited by the screen size and the need to include information about all stations on a single screen. The essential feature of this screen would be a signalling capability for each station to indicate when something needs further attention from the proctor.

This proctor screen might also contain the following information:

1) the terminals which are in use,
2) the test being administered at each terminal,
3) the elapsed time of test administration at each terminal (to assist in scheduling incoming examinees),
4) a code alerting the proctor to specific problems at individual terminals (problems such as unusually slow responding).

Other displays could be requested by the proctor. One such display might provide information on a specific examinee's progress:

1) the examinee's name,
2) the test being administered,
3) other tests administered to or scheduled for the examinee,
4) the elapsed time of administration for the current test,
5) details concerning the probabilities of such problems as random responding, and an indication if those probabilities exceed predetermined criteria for action by the proctor.

Another such display might contain scheduling information like the number of free terminals, the anticipated number of terminals to be free at a later time (based on the progress of individuals currently using the system), and the number of examinees scheduled for testing.

The printing terminal would allow a proctor to capture an examinee's status at a particular point in time (the CRT displays would be regularly updated) and also to print a summary at day's end of data such as the

number of examinees tested, the types of tests administered, and a list of problems encountered.

An examinee monitoring station, coupled with computer algorithms for avoiding and identifying problems during test administration, will greatly facilitate the performance of test administrators and test proctors and will allow large numbers of examinees to be served efficiently and effectively.

References

Levine, M.V. & Drasgow, F. Appropriateness measurement: Basic principles and validity studies. In D.J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1980.

Levine, M.V. & Rubin, D.B. Measuring the appropriateness of multiple-choice test scores. Journal of Educational Statistics, 1979, 4(4), 269-290.

Roid, G. & Haladyna, T. The emergence of an item-writing technology. Review of Educational Research, 1980, 50(2), 293-314.

Weiss, D.J. The state of the art of adaptive testing and latent trait-theory. In D.J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1980.

Wright, B.D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14(2), 97-116.

PRICE, Barbara J., Commissioner, Civil Service, Toledo, Ohio.

THE USE OF AUDIO-VISUAL MEDIA IN SELECTION (Thu A.M.)

The Toledo Civil Service Commission has, for some time, incorporated the use of audio-visual materials in its recruitment and promotional selection programs.

The primary use, to date, of audio-visual materials has been in the form of films which portray job situations requiring judgement or observational skills. Such films have been successfully used in large group settings as part of Police Officer recruitment examinations. In this application, a multiple-choice test item format was employed with satisfactory results.

The advantages of using film situations, as compared to case studies or other paper and pencil procedures, are the more realistic portrayal of events possible through the media of film and video tape and the relative freedom from educational or cultural restrictions when a lengthy or complex situation is presented on film. Additionally, the complexity available in film allows for the generation of a large number of examination items.

In practical use, film situations have exhibited less adverse impact than traditional selection procedures when used to assess observational skills and decision-making ability. For use in the United States, film situations have the distinct advantage of having withstood court review when used as one component of a content validity examination strategy. The use of this medium also exhibits many possibilities for use in other validation strategies calling for an effective measure of situational judgement or observational skills.

PUZICHA, Klaus J. Ph.D., and ROEPKE, Adelheid B., German Armed Forces Psychological Services Research Institute, Stroitkraefteamt, Germany.

ANALYSIS OF MOTIVATION TOWARDS THE FEDERAL ARMED FORCES (Mon P.M.)

The Federal Armed Forces is organized as a compulsory military service with a quota of 55% enlisted men. The personnel-development in the German Armed Forces is characterized by two main problems:

1. Since rearmament in 1956 there have been difficulties in covering requirements of the Armed Forces in quantity and quality.

2. In the late eighties we'll have a lack of soldiers because of decreasing birth rates.

On the basis of a representative panel-inquiry with young men determination of the decision-behavior between the alternatives:

- enlistment,
- military service under compulsory conditions,
- conscientious objections

was analysed.
Relevant predictors of this behavior are variables in the areas

- reference group influences,
- calculation of advantages and disadvantages,
- political and social attitudes,
- personality attributes,
- job motivations,
- unemployment.

Important are the influences of reference groups and - at first - the individual result of a rational calculation of costs and outcomes of the mentioned three alternatives. An a-priori-determination of decision by corresponding systems of values is very seldom. We presume a subsequent formation of such values as a type of rationalization of the individual's behavior.

ANALYSIS OF MOTIVATION TOWARDS THE FEDERAL ARMED FORCES

Klaus J. Puzicha and Adelheid B. Roepke
German Armed Forces Psychological Services Research Institute

The Federal Armed Forces are organized as a compulsory military service with a quota of 55 % enlisted men. The personnel development in the German Armed Forces is characterized by two main problems:

1. Since rearmament in 1956 there have been difficulties to cover requirements of the Armed Forces in quantity and quality.
2. In the late eighties we'll have a lack of soldiers because of decreasing birth rates.

1. Theoretical basis

Subject of our investigations in this context is the motivation-analysis of young men towards the Federal Armed Forces. At first we will present a model of description to explain young men's motivation in favour of or against military service within a network of conditional variables. These four interdependent variables are

+ integration of the Armed Forces into the civil society around;
+ attractiveness of the military profession respectively the job of a young soldier under compulsory conditions;
+ effectiveness of the socialisational agencies family, school, and Bundeswehr with regard to the learning object: "Making plausible why the norms of our society are worth to be defended";
+ the morale of the soldiers, i.e. their confidence in the military and political leaders, in the efficiency of their own arms, in the efficiency of the alliance.

In addition to these determinants of motivation one has to keep in mind a historical component. In the FRG we can feel the after-effects of the political discussions in the fifties dealing with the relevance of a German rearmament after World War II.

2. Methods

Based on this model of description we have managed a representative panel inquiry:

DESCRIPTION OF SAMPLES

SAMPLES	MEN 1979	PANEL 78/79	VOLUNTEERS 1979
N =	1784	1167	2229
AGECLASSES			
1957 OR EARLIER	-	-	7 %
1958	-	-	14 %
1959	17 %	24 %	48 %
1960	27 %	37 %	25 %
1961	28 %	39 %	4 %
1962 OR LATER	28 %	-	2 %
EDUCATIONAL LEVEL			
5-YEAR SECONDARY SCHOOL ED.	47 %	47 %	36 %
INTERMEDIATE SCHOOL EDUCATION	29 %	29 %	21 %
UNIVERSITY ADMISSION CERTIFICATE	24 %	24 %	43 %
SELECTION OF SAMPLE	RANDOM		RANDOM

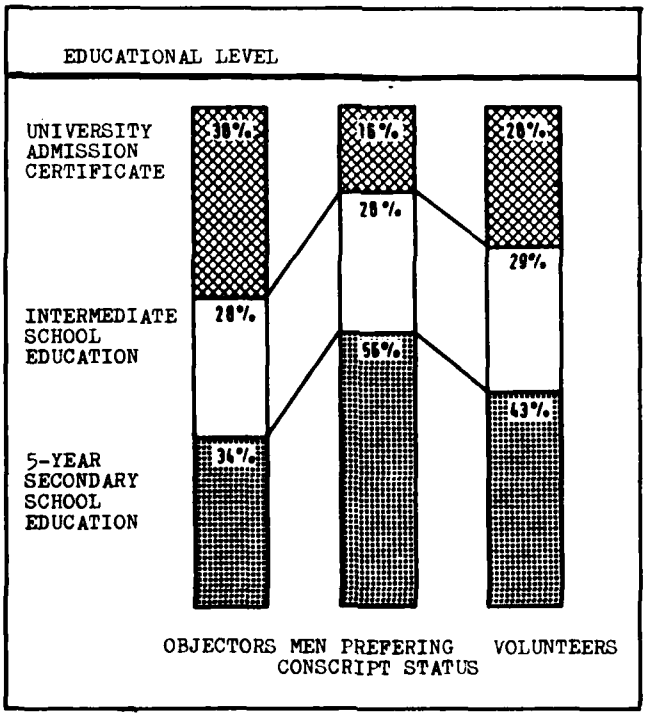
This table gives some informations about size, distribution of age and educational levels of the panel as well as of a sample of volunteers.
In yearly inquiry cycles and in a panel design during the years 1976 to 1980 we questioned a representative sample of young men who had not served with the Armed Forces.
Subjects of this study were the attitudes of the final decision for one of the three alternatives

+ enlistment as a volunteer;
+ military service of 15 months under compulsory military service conditions;
+ conscientious objection.

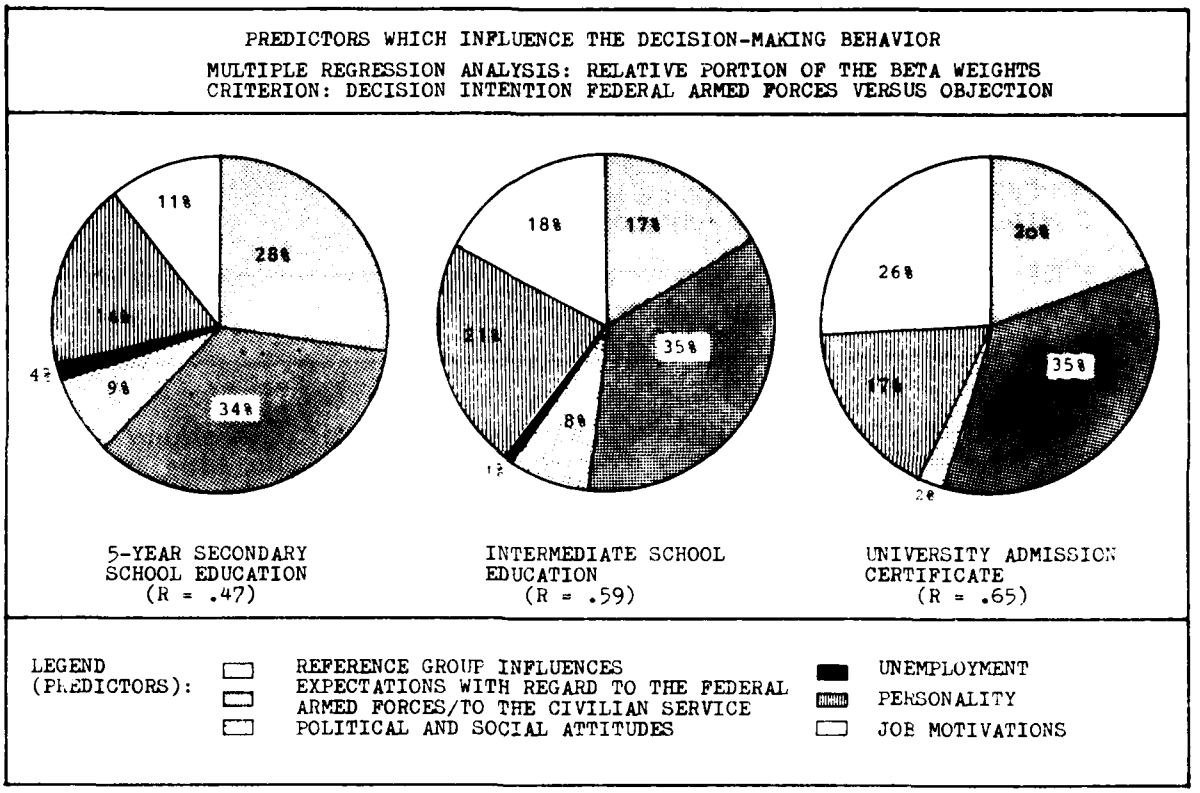
In the last years the rate of CO's has increased continously in the FRG. In 1979 we had more than 44.000 applicants.
The tendency today is increasing furthermore.

"THERE ARE DIFFERENT POSSIBILITIES FOR YOUNG MEN IN FAVOUR OF OR AGAINST THE FEDERAL ARMED FORCES. WHICH OF THESE POSSIBILITIES COMES NEXT TO YOUR OPINION?"	
I MADE AN APPLICATION FOR CO	3 %
I WILL MAKE AN APPLICATION FOR CO	8 %
I AM NOT SURE IF I WILL JOIN THE ARMED FORCES UNDER COMPULSORY CONDITIONS OR IF I WILL MAKE AN APPLICATION FOR CO	14 %
I WILL JOIN THE ARMED FORCES UNDER COMPULSORY CONDITIONS	54 %
I AM NOT SURE IF I WILL JOIN THE ARMED FORCES UNDER COMPULSORY CONDITIONS OR IF I WILL JOIN THEM AS A VOLUNTEER	11 %
I WILL MAKE AN APPLICATION FOR VOLUNTEER	9 %
I MADE AN APPLICATION FOR VOLUNTEER	1 %

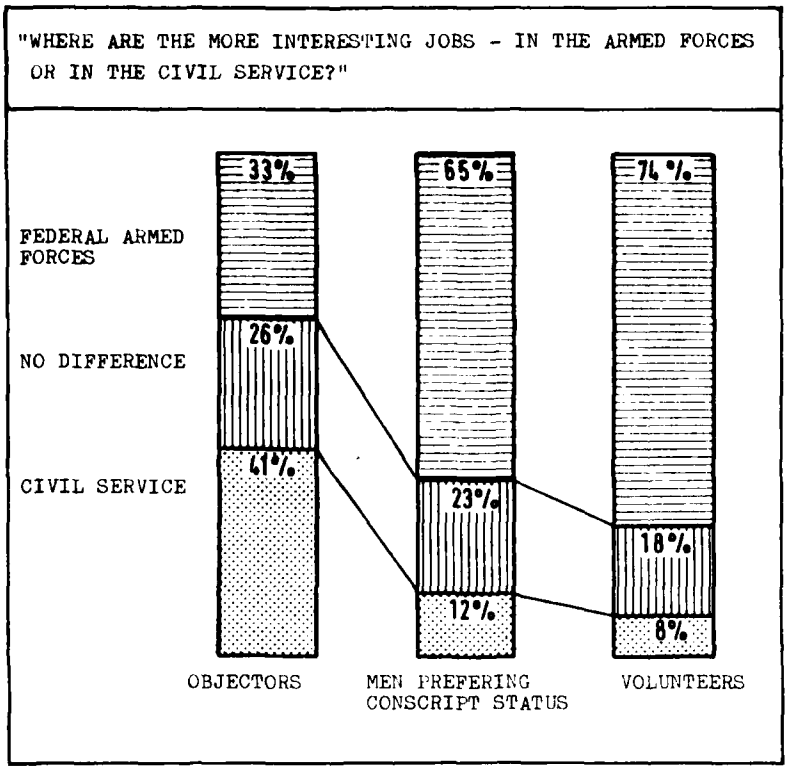
This table shows the behavioral tendencies of the young men we questionned. Every fourth of them has the idea of objection, every fifth of them has the tendency to enlist as volunteer.



This statement in general is to be differentiated by controlling the educational level. Objection extensively is characterized by a high level of education. The preference to serve under compulsory conditions is typical for young men with a lower level. So the ambitions of our German brothers in the GDR to establish Armed Forces with workmen and farmers perhaps is surpassed by the westgerman army.



This graph shows the most central result, a multiple regression analysis of six groups of predictors with a behavior criterion of a choice between the aformentioned three decision alternatives.
In summary, this display shows that the decision of a young man essentially depends on a rational consideration of the advantages and disadvantages of the alternatives. Reasons of conscience (as prescribed for a conscientious objection by our constitution) and the conviction that our social order is worthwhile defending (as requested from volunteers at the selection centers) in most cases are of minor importance only.
In the following we will report some results out of the two most important fields of determination.



This graph demonstrates the relevance of calculating costs and profits for the decision. The more interesting a young man anticipates the military job the more he tends to enlist. The more interesting he anticipates the civilian alternative service the more increases the probability of objection.
For the two extreme alternatives - objection and enlistment - we have analysed whether young men intend to one of these alternatives would stick to their intention even under the loss of some profits. For the CO's we analysed the loss of usual advantage to sleep at home, a possible prolongation of civilian service or an obligatory assignment to nursing services - for most of them a very unattractive job.

If such an extremely burdensome alternative service with all disadvantages threatened only every fifth person willing to object would be prepared to finally realize his intention.

For enlisted men we have analysed the effects of three advantages, for example the loss of the training for a civilian second career or the necessity to serve far away from home. Compared with the results of objection - analysis exact the same quota of steadiness - only every fifth - can be shown looking at the results.

THE MOST IMPORTANT PREDICTORS IN FAVOUR OF OR AGAINST THE FEDERAL ARMED FORCES. MULTIPLE REGRESSION ANALYSIS (CRITERION: DECISION INTENTION FEDERAL ARMED FORCES VERSUS OBJECTION) R = .52

RANKING	
1	"WHERE ARE THE MOST INTERESTING JOBS?"
2	+"ANTICIPATING A LIFE IN BARRACKS I FEEL **FAINT**"
3	EXTRINSIC JOB MOTIVATION
4	"ARE THERE ANY OBJECTORS AMONG YOUR FRIENDS?"
5	"WHAT DOES YOUR FATHER PREFER: OBJECTION OR JOINING THE FEDERAL ARMED FORCES?"
6	+"THE MANY SUPERIORS IN THE FEDERAL ARMED FORCES GET ON MY NERVES"
7	+"I LIKE TO WEAR UNIFORM"
8	"WHAT DO YOUR FRIENDS PREFER FOR YOU: OBJECTION OR JOINING THE FEDERAL ARMED FORCES?"
9	AGGRESSIVITY

The considerable overweight of extrinsic motivations is to be shown again in the schedule of the results of another multiple regression analysis. Among the nine most important predictors six belong to this area of determination, two of them belong to the field "reference groups".
Very important for the final decision of a man is his individual reaction to three attributes of military life characterizing the military in the whole world.
These components are living in the barracks, the use of uniform, and military superiors. Nearly all young men do not like any of these aspects.
Their reactions vary from

"I don't like it, but I will cope with it"

vs.

"That will not come into question for me at all".

The individual score in that dimension determines essentially the final behavior.

4. Discussion

We analysed six groups of determinants for the decision-behavior of young men in Western-Germany between the alternatives enlistment, service under compulsory conditions, and CO. The most relevant predictors were find in the field of calculating costs and outcomes of the mentioned three alternatives. An a-priori-determination of decision by corresponding systems of values is very seldom. We presume a subsequent formation of such values as a type of rationalization of the individual's behavior.

We can't find any indicators that a system of value arising in a process of ex-post-rationalization should have less behavioral relevance than values originated from early child socialisation. We find it characteristic for the ageclass of 16 to 18 developing a lot of political and ideological attitudes during this period and integrating them into personality. The social-psychological valuation of the dominance of cost-outcome-calculations in the young man's decision process cannot correspond with the idealistic expectations of public opinion and - in our case - of our constitution. One cannot expect from a young man in our society that he exclusively deals in an altruistic manner while the adult society is characterized by egoism. We are sure that most of the young men enlisting in the Armed Forces or those objecting the service have "egoistic" arguments to do so. Nevertheless we are sure that most of them will become good soldiers respectively good co-workers in the alternative civilian service.

Attitudes and values will change during the decision process and after decision. The necessity to proof their decision face to face with their reference-groups and official institutions reinforces this process. As far as the institutions Bundeswehr and civil alternative service are able to communicate their legitimations to their "freshmen", an extensive identification with the norms of these institutions can be expected.

5. Appendix

Compilation of the most important single "reasons" for the decision for the Federal Armed Forces or for conscientious objection, respectively. Presentation in the form of a modal personality.

The typical conscientious objector / volunteer is / has:

conscientious objector	volunteer
higher school education	
....comes from middle-class families (subordinate and small employees)	....comes from middle-class families (skilled workers, subordinate and small employees)
pessimistic assessment of the development on the labour market	
	unemployment among friends and acquaintances above average
....is affected by role conflicts: girl friend, friends are in favour of, parents, employers against conscientious objection	
	....has more often than the average a volunteer among his acquaintances
....the most important reference persons are: friends, employer (A, B*); girl friend and father (C*)	....the most important reference persons are: mother, father, brothers and sisters (A, C); friends, father (B)

---

*applies only to
A = persons with 5-year secondary school education
B = persons with intermediate school education
C = persons with university admission certificate

	scarcely aggressive
introverted	extraverted
scarcely dominant	dominant
scarcely unpolitical attitude	scarcely unpolitical attitude
	affinity for the military
pacifistic	less pacifistic
more often than the average suicidal ideas	
higher consumption of alcohol and fashionable drugs	higher consumption of alcohol
small need for achievement	high need for achievement
small extrinsic and intrinsic professional motivation	strong extrinsic, intrinsic and social professional motivation
dissatisfied with school or profession	satisfied with school or profession
negative assessment of the activities and the superiors in the Federal Armed Forces	positive assessment of the activities and the superiors in the Federal Armed Forces
concrete ideas and information on the Federal Armed Forces	concrete ideas and information on the Federal Armed Forces

....considers the Federal Armed Forces to be important	....considers the Federal Armed Forces to be important
....regards the value of the Federal Armed Forces for his future life as little to insignificant	....regards the value of the Federal Armed Forces for his future life as great to little
	....considers service in the Federal Armed Forces to be more interesting than civilian alternative service
....considers civilian alternative service to be more exhausting than military service	....considers military service to be more exhausting than civilian alternative service
....thinks that in the Federal Armed Forces chicanery occurs more often than in the civilian alternative service	....thinks that in the Federal Armed Forces chicanery occurs more often than in the civilian alternative service
....has an unrealistic opinion regarding the number of home leave trips on weekends in military service	....has a realistic opinion regarding the number of home leave trips on weekends in military service
....would refrain from conscientious objection of: - a prolongation of the civilian alternative service to 2 years (A,B, C) - care of patients confined to bed (A, B) - immediate draft to the civilian alternative service (C)	....would refrain from voluntary enlistment in case of: - a garrison far away from home (A) - without separation allowance and readjustment allowance (B, C)

(776)

RAUCH, Martin, Chief Psychologist, German Armed Forces, Bonn, West Germany.

"DEVELOPMENT OF SELECTION SIMULATORS IN THE GERMAN MILITARY AVIATION PSYCHOLOGY" (Tue A.M.)

In the German Military Aviation Psychology a so-called sequential selection strategy has been established. It consists of three steps:

1. Paper-pencil-tests (examination of basic psychological functions)
2. Examination of complex functions (e.g. psychomotor functions) by the use of special apparatus (e.g. ICA = Instrument Coordination Analyzer; PCA = Precise Coordination Analyzer)
3. Screening

The screening consists of three phases:

a. psychological screening;
b. theoretical screening;
c. practical screening.

For the special purpose of psychological screening German military aviation psychology is going to develop a so-called selection simulator system APAMS (Automated Pilot Aptitude Measurement System). The whole system will consist of four simulators (each with motion system and CGI = Computer Generated Imagery) and one CPU.

The applicant will have enough time to learn how to "fly" simple missions on the simulator. Later on in the examination situation he will have to solve simple flying tasks. His performance is measured automatically by the differences between reference inputs and actual values.

Since this is a relatively new technical and methodological approach, a close cooperation with other NATO countries - above all Canada, which plans a similar APAMS development - is highly desirable.

Development of Selection Simulators in the German Military Aviation Psychology

Disposition:

1. Aeronautical/aviation-psychological Step-by-step Selection System for FAF Flying Personnel

1.1 Aviation-psychological Pre-selection

1.2 Initial Aviation-psychological Test

1.3 Aeronautical/aviation-psychological Screening

1.4 Quantitative Data on the Aeronautical/aviation-psychological Selection Practiced in the Federal Armed Forces (FAF)

2. Automated Selection Simulators

2.1 Purpose

2.2 US-APAMS (Automated Pilot Aptitude Measurement System)

2.3 Further US Developments and Approaches

3. FAF Aviation-psychological Selection System (FPS 80)

3.1 First Approach

3.2 Second Approach

3.3 Canadian Approach/German-Canadian Cooperation

4. Conclusions

"Development of Selection Simulators in the German Military Aviation Psychology"

1. Aeronautical/aviation-psychological Step-by-step Selection System for FAF Flying Personnel

Flying personnel for the three FAF services is selected in accordance with the so-called sequential aeronautical/aviation-psychological selection strategy. This selection strategy step-by-step selection system consists of the following stages:

1.1 Aviation-psychological Pre-selection

Both the Officer Applicants Test Center and the Volunteers Recruiting Stations use paper and pencil tests to determine a series of aeronautically relevant basic functions (e.g. intelligence, attention). Those who pass this comparatively coarse-meshed screen, take part in the next stage of the selection system, i.e. in the

1.2 Initial Aviation-psychological Test

The initial aviation-psychological test is carried out within the so-called Military Pilot Aptitude Test (FBA Wing 49/Aviation Psychology Test Center). Who successfully passes the initial aviation-psychological test, takes part in the medical Military Pilot Aptitude Test. During the initial aviation-psychological test, complex psychomotor functions, decision reactions and the flexibility of the attention distribution are tested. To be able to do this, the use of part task trainers is required, which imitate the reality of flying to a certain extent, though neither motion systems nor computer-generated images (CGI) are employed. It has been shown that the existing quality of simulation is methodically sufficient. The most important part task trainers are the ICA (Instrument Coordination Analyzer) and the PCA (Precise Coordination Analyzer). In 1981/82 the existing devices, which have been in permanent use since about 1970, will be replaced by a complex ICA/PCA combination model. The 1.3 million Deutschmarks required have already been approved. Once the procurement will finally have been completed, standardization of the

(780)

system will at least take another year. This standardization to be realized must be seen in connection with the computer-assisted aeronautical/aviation-psychological longterm success control already in the planning stage. The BE and NE air forces have already shown interest in a procurement of the German system. Swissair follows a similar path, however, with a slightly modified methodical approach.

Those applicants who have passed the "initial aviation-psychological test" stage, will go on to screening, which again consists of three stages.

1.3 Aeronautical/aviation-psychological Screening

1.3.1 Psychological screening with selection simulators.
The following arguments will focus on them (Para 4.)

1.3.2 Theoretical screening
(Aeronautical theory, aircraft technology, flight safety, meteorology, etc.)

1.3.3 Practical Screening
(in a training aircraft; in the FAF-owned Piaggio 149 D)

Each of the steps and stages finishes with a selection conference, in which an aviation psychologist, who is a permanent member with the right to vote, takes part.

1.4 Quantitative Data on the Aeronautical/aviation-psychological Selection Practiced in the Federal Armed Forces

- approx. 10 percent of the applicants are eliminated during the so-called aviation-psychological preselection;
- elimination rate of the initial aviation-psychological test (Aviation Psychology Test Center): approx. 26 percent;
- the elimination rate of the medical Military Pilot Aptitude Test is about the same: approx. 26 percent;

- within the total screening about another 29 percent of the applicants are eliminated;
- in the UPT (Undergraduate Pilot Training) used in the USA, the elimination rate is about 15 percent (these 15 percent refer to pilots; in the case of backseaters the percentage is about 10 percent).

From 100 applicants, who are interested in a pilot's career, only 25 (one quarter), as a rule, achieve the target and become pilots/backseaters after having completed this step-by-step selection process. Each year the Aviation Psychology Test Center tests about 800 applicants.

2. Automated Selection Simulators

2.1 Purpose

Of the step-by-step selection system described, Para. 1.3.1, "Aviation-psychological Screening",has not yet been realized so far, since a corresponding selection simulator is still in the conceptual planning stage in the Federal Armed Forces. The objective of such automated selection system is to subject the total aeronautical potential of an applicant to an objective and quantitative analysis. It is expected to recognize unfit applicants at this stage with a higher degree of precision than that achieved formerly. It is expected that with this additional "selection simulator" stage the subsequent elimination rates in the screening and UPT stages and the accompanying costs can be reduced considerably.

2.2 US-APAMS

A first successful experimental test on a selection simulator was carried out in 1976 with the USAF. The results of the tests with the so-called APAMS (Automated Pilot Aptitude Measurement System) are discussed in the AFHRL-TR-75-58 report.

The experimental device constructed consisted of a simulator with a display unit and an automated instruction input mechanism as well

as a computer-assisted evaluation unit, which also operated automatically. In the experiment each of the applicants had five hours of exercise in order to learn simple flight missions. On the basis of the flight missions learnt, simple "missions" had to be flown in the test situation. Then the performance achieved was objectively calculated as the difference between actual values and desired values.

The APAMS experiment was carried out with 178 applicants; the results obtained were compared with those of the UPT (T-41 and T-37).

The results clearly show that the achievements of stages T-41 and T-37 can be predicted with sufficient accuracy by means of the APAMS values. The APAMS values also allow unequivocal predictions with regard to Manifestation of Apprehension (MOA), Self-Initiated Elimination (SIE) and Flying Training Deficiency (FTD).

2.3 Further US Developments and Approaches

At present USAF carries out experiments with a considerably simplified APAMS system (without motion systems and GCI). It is regarded as possible that high predictive validities similar to those obtained with the complex test device of the first APAMS can be achieved with this simplified version.

The US Navy also follows the APAMS approach. At the Naval Aerospace Research Institute (Department of Aerospace Psychology) tests are carried out at present with a so-called NAPAMS (Naval Automated Pilot Aptitude Measurement System).

APAMS and NAPAMS have been conceived for the aviation-psychological selection of flying personnel for fixed-wing aircraft.

Especially for the selection of helicopter personnel, the US Army additionally developed the so-called PASS system (Rotary Wing Proficiency-Based Aviator Selection System). First experimental results are discussed in the ARI - Technical Report TR 79-A1,

dated January 1979. The tests were carried out by McDonnell/ Douglas on behalf of the U.S. Army Research Institute for the Behavioral and Social Sciences.

The core of the PASS system is a UH-1 FS flight simulator with a motion system having five degrees of freedom. Simple and frequently occurring helicopter-specific flight missions were included in the program both during the training and during the test stage. Operational functional readiness of PASS was proven in a first test with 11 experienced helicopter pilots and 11 helicopter pilot applicants. Still in 1980 the predictive validity of the systems for the army aviation-related UPT will be studied experimentally.

3. FAF Aviation-psychological Selection System (FPS 80)

3.1 First approach

The German aviation psychology was fast in recognizing the advantages of the US APAMS approaches (cost reduction, marked improvement of predictive validity, increase in objectivity and reliability, savings in personnel). Already at the beginning of 1979 a tactical requirement for the so-called AVIATION-PSYCHOLOGICAL SELECTION SYSTEM (FLIEGER-PSYCHOLOGISCHES SELEKTIONSSYSTEM)) FPS 80 was approved. Methodically the tactical requirement closely followed the first US APAMS.

According to this methodical conception, the FPS 80 should serve the purpose of obtaining aviation learning test results during the so-called aviation-psychological screening (Para. 1.3).

The aim of this approach is to test in a standardized objective form both the learning rate given complex aeronautical conditions and the stress capacity under these conditions which in each specific case comprise a maximum of one-hour missions throughout several days. Even if the entire test situation largely develops under automated conditions, the additional observation of attitudes by experienced aviation psychologists, who ideally are pilots themselves, cannot be dispensed with.

In accordance with tactical requirements, the simulator should be equipped with a motion system in order to increase the impression of reality and in order to be able to recognize in time signs of sensitiveness to motion in the applicant.

FPS 80 should have the following technical structure:

- a central digital computer;
- 4 cockpits;
- a digital computer-generated imagery in each cockpit;
- a motion system in each cockpit;
- a sound simulation device in each cockpit;
- an audio-visual system for automatic instruction, error feedback and for supplying the applicant with information, one in each cockpit;
- a device for voice communications between aviation psychologists and applicants, one in each cockpit;
- a central simulator control station for control and supervision of the entire system with a Visual Display Unit (VDU) for the continuous presentation of test parameters and results, with one monitor for the supervision of the computer-generated imagery, with one monitor for the VDU installed in the cockpit and with one set of repetition instruments;
- 4 sets of peripheral equipment (data terminal for program input and modification, hard-copy unit for recording the test results, magnetic tape unit for storing the test results);
- software.

The expansion of costs drove the original budget estimate of approx 4.0 million Deutschmarks (at the beginning of 1978) to approx. 16.0 million Deutschmarks by the end of 1979.

3.2 Second Approach

The main cost-increasing parameters were to be found in the field of the motion systems and the computer-generated images. It proved

to be necessary to order a revision of the existing FPS 80 tactical requirement. Still in 1980 the tactical requirement will be revised bearing in mind technical, methcdical and financial aspects. If necessary, motion systems and electronic computer-generated imagery (CGI) have to be dispensed with. At this juncture we would like to refer again to analogous efforts undertaken by USAF which at present also largely simplifies the first experimental APAMS (Para.2.3).

Even given a largely reduced system, a satisfactory predictive validity is assumed in aviation psychology as well. According to conservative estimates a realization of FPS 80 can be expected by about the mid-eighties.

3.3 Canadian Approach/German-Canadian Cooperation

The Canadian Air Force has also advanced comparatively far with the conceptual preparation of an APAMS. A corresponding Program Development Proposal was approved by the Department of National Defence as early as in the spring of 1978. In accordance with the then scheduling a delivery of the Canadian system to its user should be expected in October 1982.

The Canadian approach also makes the attempt to keep costs as low as possible (approx. 1 million Canadian dollars). The core of the Canadian APAMS will be a GAT simulator (in 1978 the Canadian Air Force procured 4 GAT simulators). The original German FPS 80 calculation was also based on GAT; however, series production of this system suddenly ceased in 1979; this also contributed to an enormous cost increase of the German approach.

On the international market there is at present no system similar to GAT in sight.

A close exchange of information in the field of APAMS/FPS 80 development was agreed upon between the German aviation psychology and the Canadian Air Force.

4. Conclusions

In future automated selection simulators will play an important role in the aviation-psychological methodology. A marked increase of the aviation-psychological predictive validity and an accompanying cost reduction is expected.

This assumption is justified both on the basis of experimental results available stemming from US sources and on the basis of international follow-up developments in this field (USA, Canada, Federal Republic of Germany, Netherlands, Belgium).

RUCK, Hendrick W., and EDWARDS, Capt. John O. Jr., Air Force Human Resources Laboratory, Manpower & Personnel Division, Brooks AFB, Texas.

THE DEVELOPMENT OF ORGANIZATIONAL EFFECTIVENESS MEASURES FOR SECURITY POLICE UNITS (Thu P.M.)

Traditionally, Air Force Security Police Units are evaluated by their performance on simulated exercises and operational readiness inspections. Although exercises and inspections serve quite useful purposes, they suffer from at least two deficiencies. First, they do not normally allow for valid quantitative comparisons between units. Second, they do not measure intangibles, such as morale, leadership or planning, that could lead to improved effectiveness. An additional difficulty encountered in attempting to measure security effectiveness is lack of quantifiable output associated with security units.

This paper presents an approach to developing measures of organizational efffectiveness that precludes some of the aforementioned difficulties. Specifically, the approach taken involved:

a) the development of a large number of potential security police unit-effectiveness indicators; b) the development of empirically -derived weights for subsets of the effectiveness indicators; c) the development of an overall unit-effectiveness score; and d) a comparison of policy capturing and policy specifying results. The techniques used in this research borrowed heavily from test construction theory, training objective/task writing methodology, organization theory, and policy measurement methodology.

THE DEVELOPMENT OF ORGANIZATIONAL EFFECTIVENESS MEASURES FOR SECURITY POLICE UNITS

Hendrick W. Ruck
John O. Edwards, Jr.
Grace R. Hrymoc

Manpower and Personnel Division
Air Force Human Resources Laboratory
Brooks AFB, Texas 78235

Traditionally, Air Force Security Police Units are evaluated by their performance on simulated exercises and operational readiness inspections. Although exercises and inspections serve quite useful purposes, they suffer from at least two deficiencies. First, they do not normally allow for valid quantitative comparisons between units. Second, they do not measure intangibles, such as morale, leadership, or planning, that could lead to improved effectiveness. An additional difficulty encountered in attempting to measure security effectiveness is lack of quantifiable output associated with security units.

The measuring of security police organizational effectiveness is made even more difficult by the fact that easily measurable indices may not be very valid indices. For example, a base with a low crime rate may have such a crime rate because the crime suppression program is very effective, or because the crime reporting program is poor, or because there is little crime in the area regardless of police efforts, or some combination of the three. Similar problems are found when considering other easily measured security indices.

This paper presents an approach to developing measures of organizational effectiveness that precludes some of the aforementioned difficulties. Specifically, the approach taken involves (a) the development of a large number of potential security police unit-effectiveness indicators, (b) the analysis of management policies regarding the relationship of individual indicators to overall unit effectiveness, and (c) the development of empirically-derived scoring systems for weighting subsets of the indicators to measure overall effectiveness.

Since "hard" criteria of security police unit effectiveness are exceedingly difficult to measure reliably, and interpretation of such measures is generally ambiguous, a two-tiered approach toward measuring unit effectiveness was adopted. The first tier was composed of measures that could be reliably collected for the various activities for which the units are responsible. The measures were developed from interviews and group interaction with security police functional experts. The second tier was composed of policy decisions about the relative importance and direction of each of the measures in terms of overall effectiveness within functional areas. Composite weights for the measures were developed by the application

of policy development procedures using senior-level headquarters personnel. Thus, the measurement problem was addressed using operationally-oriented functional experts, and the evaluation of those measures was addressed using senior-level policy makers. A major implication of this approach is that unit effectiveness is not necessarily a fixed construct. Given different policy makers, different units could be judged optimally effective.

Two separate activities were involved in constructing measures of organizational effectiveness: (1) development of the Security Police Organizational Effectiveness Measures (SPOEM) Booklet specifying observable measures, and (2) development of policy-based scoring procedures for the SPOEM. Each activity will be discussed in turn.

Development of the SPOEM Booklet

A team of six experienced security police personnel, together with two research psychologists and an occupational analyst, formed the nucleus for the development of the SPOEM. The SPOEM items were developed in conference with team members being augmented by subject area experts whenever necessary. The process model used to develop the SPOEM borrowed heavily from occupational analysis, task analysis, and Specialty Knowledge Test (SKT) development methods. The steps used in developing the SPOEM were as follows:

1. All functions performed within a Security Police (SP) squadron were delineated. The delineation of functions served several purposes. First, it provided the psychologists with an understanding of the scope of the measurement problem. Second, it forced the SP personnel to consider SP functions independent of unit organization. Third, it provided the beginning of the road map which the group would use in developing the SPOEM.

2. The functions were grouped into seven major areas (see Table 1). This process resulted in an agreed-upon outline of the SPOEM.

3. Importance weights for each of the seven functional areas were assigned by each of the six SP team members and the project officer. The weights were discussed in terms of percent of contribution to overall unit effectiveness. The total weight of the seven areas was 100 percent. Weights were discussed publicly and an effort to achieve consensus was made. Unfortunately time constraints precluded full consensus; however, substantial agreement was achieved. The purpose of assigning weights was to target the number of effectiveness items to be written in each of the seven areas. This precluded the writing of many items in a less important area due to ease of describing items. It also caused the team to search for additional items in the more important areas. Since no more than 200 organizational effectiveness criteria could be measured in a reasonable amount of time at each unit, each of the seven areas was assigned a proportion of the items based on the area's weight.

4. Types of effectiveness criteria were reviewed. The concepts behind and forms of behaviorally anchored rating scales (BARS) (Fogli, Hulin, & Blood, 1971), behavioral expectation scales (BES) (Smith & Kendall, 1963), criterion objectives for training (AFP 50-58), test items, inspector general

TABLE 1

Number of Items in the Functional Areas of the Security Police Organizational Effectiveness Measures

Functional Area	Original Outline Weights	Number of Items Measured	Sample Items Measured in the Area
Security Flight Operations	33%	23	Percent of required responses force exercise conducted in the last 60 days Variance within the rank distribution among security flights
Personnel Management and Training	15%	25	Percent of functional inspections made by other organizations that are passed satisfactorily by SP unit Percent of SP owned POV accidents
Law Enforcement Flight Operations	19%	19	Percent of LE assigned personnel actually working law enforcement flights Number of reported cases of minor crimes per 90 day period
Facilities and Equipment Management	13%	16	Percent of properly equipped SP vehicles in operation versus total number assigned Percent of M-16s with carbon or dirt in barrel
Planning, Programming, Budgeting, Supply, and Procurement	11%	11	SP success compared to Combat Support Group success in obtaining General Support Division Equipment (628) money. (100 is equal) Scaled effectiveness of unit budgeting procedures and internal operations
Police Services	5%	5	Percent of attendance of required members at RPC meetings Number of incidents reported on lost government property per 90 day period
Crime Suppression	4%	5	Percent of cases solved per all cases closed Number of major security police community relations presentations or information items given in the community per 60 day period
TOTAL	100%	104	

(IG) inspection items (USAFEP 125-40), and productivity criteria (Smith, 1976) were discussed. The criterion model chosen was an adaptation of the training criterion objective. Generally, effectiveness measures would include conditions, behaviors, and standards. Furthermore, the measures would be expressed as a percentage whenever possible.

5. Each of the seven functional areas was addressed by the group as a whole. Experts in each area were called in when necessary to provide information and guidance. Prior to including a specific function as an effectiveness item, group consensus was achieved on the following: (a) the item to be measured was, indeed, related to SP unit effectiveness, (b) the item was not included in another measure, (c) the item belonged to the area under discussion, (d) the item was unambiguously related to organizational effectiveness, and (e) the item was measurable. Ratings and subjective opinions were generally not allowed as measures, unless the SP personnel felt quite strongly about an item, and there were no other options.

6. Once the SP effectiveness measures were drafted, the whole list of items was reviewed by the team, project officer, and interested headquarters staff members. Approximately 125 SPOEM items remained after this review.

7. The team spent one day at the SP squadron measuring as many of the effectiveness criteria as possible. The pilot test resulted in the rewriting of a small number of items and the deletion of a number, so that 104 remained. Sample items measured in each of the given functional areas have been included in Table 1. Several examples of operational definitions of SPOEM items are presented in Table 2.

Development of a Scoring Protocol for the SPOEM

The approach used to measure the policy maker's judgments regarding organizational effectiveness was policy capturing (Christal, 1968). In applying the policy capturing technique, synthetic unit profiles for each of the seven areas were presented to headquarters personnel. The headquarters personnel rated the effectiveness of each hypothetical unit in each functional area. Regression equations were developed to determine the weights of each measure within each functional area. These weights represented, mathematically, the views of the headquarters judges regarding the relationship of each measure to unit effectiveness for that functional area. As a result of policy capturing, a score for each of the seven areas of the SPOEM was derived. Overall unit effectiveness measures, combining the seven functional areas into an overall effectiveness score, were also developed using the same procedure.

Policy Capturing

Simulated unit profiles were developed for each of the seven functional areas included in the SPOEM. In addition, an eighth profile set, using hypothetical scores from the seven functional areas, was developed for assessing overall unit effectiveness.

TABLE 2
Examples of the Operational Definitions of SPOEM Items

1. Percentage of persons actually on the same post as indicated on CSC status board.

Instruction: Check every post on three shifts, two hours or more after shift change.

Measures:	SHIFT 1	2	3
A. Number of posts	_____	_____	_____
B. Number of persons in correct post	_____	_____	_____

33. Percentage of possible communication channels being utilized.

Time: 90 days

Definition: Communication channels are Airman's Advisory councils, Dormitory councils, wives clubs, and squadron newsletters.

Measures:	None Exists	Yes	No
A. Did the Airmen's Advisory Council meet?	_____	_____	_____
B. Did the Dormitory council meet?	_____	_____	_____
C. Did an SP wives club meet?	_____	_____	_____
D. Was a squadron newsletter published?	_____	_____	_____

72. Percent of security flight airmen fully equipped with SP equipment while posted.

Definition: Fully equipped with SP equipment - possessing appropriate foul weather gear, a flashlight (during swing and mid shift), and a whistle (all three items).

Instructions: Randomly sample another 10% of each flight (day, swing, & mid) on a different day; physically inspect.

Measures:

A. Number personnel sampled	_____
B. Number fully equipped	_____

The generation of unit profiles for each of the seven functional areas was accomplished in several steps.

1. Four SP units were measured on the SPOEM items. Means, standard deviations, and the direction of item interrelationships were inferred from analyses of these data.

2. The total number of profiles generated for a functional area was equivalent to twenty-times-the-number-of-items included in that area.

3. Security experts reviewed each profile for believability. Profiles containing obvious contradictions in items were eliminated.

4. From the remaining profiles, the number of profiles randomly selected for each functional area for subsequent evaluations was equal to ten-times-the-number-of-items in an area.

Twenty headquarters-level security police experts from a single major command served as subjects for the policy capturing phase of this study. Each subject rated the unit effectiveness of hypothetical units on at least one of the profile sets. Table 3 displays the number of hypothetical units rated for each of the profile sets and the number of Security Police experts that rated each set.

Statistical Analysis

For each rater and each profile set, a multiple linear regression model was developed using SPOEM items as predictors and the raters' overall effectiveness ratings as criteria. Regression models developed for different raters for the same profile set were statistically compared to determine whether or not the various raters had common policies.

One measure of success of policy capturing is the size and statistical significance of the squared multiple correlation (R^2) values obtained across raters. The correlation indexes the amount of variance in the criterion ratings accounted for by the individual elements in the simulated profiles. Table 4 summarizes the results of the policy capturing model development for each of the seven functional areas and for the overall score. The R^2 values for individual policies ranged from a low of .36 to a high of .93, and all were significant beyond the .001 level.

Since the models were judged to have effectively captured the rating policies of the experts, policy models were statistically compared for similarity. Individuals who were quite dissimilar from the others were excluded from subsequent analyses. The number of raters remaining in each functional area is also shown in Table 4.

For each group of raters with similar policies, a new composite regression model was developed using the combined ratings on a profile as the criterion and the profile variables as predictors. The last column in Table 4 shows the final R^2 values for the composite model associated with each functional area. The appendix presents the individual SPOEM items

TABLE 3

Number of Raters for Each Profile Set for Policy Capturing

Profile Set		Number of Hypothetical Units to be Rated	Number of Raters
I.	Security Flight Operations	230	4
II.	Personnel Management and Training	250	5
III.	Law Enforcement Flight Operations	190	5
IV.	Facilities and Equipment Management	160	5
V.	Planning, Programming, Budgeting, Supply, and Procurement	110	4
VI.	Police Services	50	5
VII.	Crime Suppression	50	5
VIII.	Overall Effectiveness	70	7

selected for use in the composite models along with their correlations with the effectiveness ratings, regression weights, and other pertinent data. Table 5 shows the results of the composite model for the overall effectiveness score.

Of the 104 items originally measured using the SPOEM, only 48 appear to be required to produce an overall score of unit effectiveness. This is because the stepwise analysis reduced the number of items scored within each of the seven functional areas; and the stepwise solution for the overall effectiveness model eliminated two functional areas: Law Enforcement Flight Operations; and Planning, Programming, Budgeting, Supply, and Procurement. If it were required to obtain unit effectiveness scores for each of these areas (even though the scores would not affect overall effectiveness), an additional 18 items would have to be measured for a total of 66.

Discussion

Despite the difficulties expected in developing a valid measure of Security Police unit effectiveness, the approach adopted in this paper seems to have been successful. Although considerable effort was required to develop measures that might be indicators of unit effectiveness, it was indeed possible to develop over 100 measures in a 3-week period. Undoubtedly, additional measures could be developed; however, the measures that were developed in this study cover the effectiveness areas that would affect large, aircraft-oriented Security Police units. This was shown by the success of the policy capturing portion of the study. Whether the items could have been more easily measured or more clearly stated are moot questions. The items were sufficiently understood so that senior managers were able to make consistent judgments based on them.

It is necessary to stress that the policy capturing approach to developing effectiveness weights for scoring the items was quite successful. The rating strategy of individual managers participating in the study was adequately reflected in the models developed. However, the composite models for all of the managers were not all the same. Decisions were made by the researcher regarding the actual models to be used. The decisions were generally clear-cut; but the necessity to make such decisions highlights a critical characteristic of the SPOEM--namely, that the scoring of the SPOEM is dependent on the perceptions of managers. Managers having identically structured units with different missions may have different perceptions of unit effectiveness. This hypothesis was not tested in the present study. Nevertheless, the fact that there was some lack of agreement among the managers in the present study indicates the hypothesis to be tenable.

As an example of the characteristics that might be found in the manager effectiveness policies, the overall effectiveness score model is cited. Not only was the functional area of Law Enforcement Flight Operations given zero weight, but the functional area of Crime Suppression was given negative weight regarding overall effectiveness. The functional area of Police

TABLE 4
Regression Results for Each Profile Set

Profile Set		R2 for Individual Models Low	High	Number of Raters in Original Sample	Number of Raters Used In Final Model	Final Model R2
I.	Security Flight Operations	.50	.70	4	3	.17
II.	Personnel Management and Training	.47	.73	5	4	.20
III.	Law Enforcement Flight Operations	.39	.88	5	2	.42
IV.	Facilities and Equipment Management	.61	.93	5	5	.50
V.	Planning, Programming, Budgeting, Supply & Procurement	.48	.80	4	4	.27
VI.	Police Services	.54	.78	5	5	.38
VII.	Crime Suppression	.36	.76	5	4	.50
VIII.	Overall Effectiveness	.62	.87	7	4	.47

TABLE 5

SPOEM Overall Effectiveness Score

Functional Area		Zero Order Correlation	Standard Weight	Raw Weight	SD of Raw Weight	Independent Contribution
I.	Security Flight Operations	.56**	.3890	.0242	.0039	.0735
II.	Personnel Management & Training	.48**	.3520	.0251	.0044	.0636
VI.	Police Services	.29**	.1569	.0094	.0028	.0216
IV.	Facilities and Equipment Management	.09	.1657	.0113	.0032	.0241
VII.	Crime Suppression	-.16**	-.1360	-.0086	.0041	.0085
	- Constant = 1.4255					

* P < .05

** P < .01

Services received positive weight. Anecdotal evidence, based on interviews with Headquarters managers, indicates that the emphasis in terms of mission importance in their command is on security, and that law enforcement is secondary. The overall effectiveness model appears to support such evidence. Apparently Security Police unit organization effectiveness is measured by the managers in terms of the output of the planning process rather than the perceived effectiveness of the process itself. As noted earlier, another set of managers might have a different policy.

The SPOEM were developed as part of an organizational design study presently being conducted (Ruck & Edwards, 1979). In response to reported difficulties in the management and conduct of SP operations, the United States Air Forces in Europe (USAFE) decided to test a new organizational structure for SP squadrons. Although the Air Force has regulations controlling the evaluation of organizational change, the USAFE/SP was interested in gathering objective data about the effects of the structural change. Specifically, they wanted to measure the effects of reorganization on jobs performed, individual job satisfaction, organizational climate, and unit effectiveness. The SPOEM and the associated scoring models were developed to assess unit effectiveness for the USAFE study. Not only can the SPOEM be used to measure the effectiveness of other Air Force Security Police units, but also the techniques used are quite promising and could be used to develop additional organizational effectiveness measures.

REFERENCES

Christal, R.E. Selecting a harem - and other applications of the policy capturing model. *Journal of Experimental Education*, 1968, 36(4), 35-41.

Department of the Air Force. Training Handbook for Designers of Instructional Systems (AF Pamphlet 50-58). Headquarters U.S. Air Force, Washington, D.C., 15 July 78, 3.

Department of the Air Force. USAFE Security Police--Guidelines and Evaluation (USAFE Pamphlet 125-40). APO New York 09012; Headquarters US Air Forces in Europe, 15 December 1977.

Fogli, L., Hulin, C.L., & Blood, M.R. Development of first-level behavioral job criteria. Journal of Applied Psychology, 1971, 55, 3-8.

Ruck, H.W., & Edwards, J.O., Jr. Measurement of changes in organizational effectiveness of security police squadrons resulting from unit reorganization. Paper presented at Fourth Annual Workshop on the Role of Behavioral Science in Physical Security, Defense Nuclear Agency, Washington, D.C., July 1979.

Smith, P.C. Behavior, results, and organizational effectiveness: The problem of criteria. In M.D. Dunnette (Ed.), Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally, 1976.

Smith, P.C., & Kendall, L.M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47(2), 149-155.

400

SAKO, S., Officer Training School, USAF OTS/MTCM, Lackland AFB, Texas.

VALIDATION STUDIES OF WRITTEN ACHIEVEMENT TESTS USED AT USAF OTS
(Wed P.M.)

The purpose of this study is to investigate the results of different procedures of test validation currently in use at the Officer Training School, USAF (OTS), Lackland Air Force Base, Texas. The study was made on written achievement test scores of over 6,000 officer trainees undergoing training at OTS during the period of 1978 - 1980. The procedures consisted of analysis of criterion objectives, item analysis, pretest/posttest analysis, and statistical analysis of test scores. Information from these validational procedures was used in making decisions for conducting changes on specific items of the written tests. A total of over 140 item revisions were processed for test improvement during the period. These changes have resulted in refining the written tests by increasing the curricular validity of the items, lowering the standard error of measurement, producing more reliable test scores, and reducing the failure rates on criterion objectives of the tests.

The VULCAN is a high rate of fire, 6 barrel, 20mm gun with explosive rounds. It is manually controlled with a range radar which allows a computer calculated, automatically inserted lead angle. The gunner must be highly proficient in manually acquiring the target, tracking smoothly, and utilization of the radar directed lead angle. In addition, he must be proficient in boresighting the weapon, visual aircraft recognition, and other related tasks. This study evaluates 891 gunners from units all over the world and points out pertinent problem areas and their solution.

VALIDATION STUDIES OF WRITTEN ACHIEVEMENT TESTS USED AT USAF OTS

INTRODUCTION

This paper describes the validational procedures for determining the validity and reliability of the consolidated written tests (CWTs) used in the Officer Basic Military Training Precommissioning Course at the Officer Training School, USAF (OTS), Lackland Air Force Base, Texas. The tests are developed and administered at OTS as part of the criteria for determining those officer trainees (OTs) who will graduate and those who will not, and for selecting honor graduates and distinguished students. It is important that these measurements possess the highest degree of validity and reliability so that the test results provide accurate indication of student academic performance. In order to ascertain these important qualities, several different methods of test analysis have been investigated. They include: (1) item analysis; (2) criterion objective (CO) analysis; (3) pretest/posttest analysis; and, (4) statistical analysis of test scores. The end results of these analyses are used primarily to assist in improving the written examinations at the School.

POPULATION SAMPLE

The population sample used in this study consisted of over 6000 OTs enrolled at OTS during the calendar years 1978, 1979, and 1980. On the average, about two-thirds of the population have had no previous military service, and the remainder have served in enlisted status, or active duty under the Airman Education and Commissioning Program. The median age is approximately 27 years. About half of the OTs are married. All of them have bachelor degrees, with master or doctorate degrees for approximately 10%.

MEASUREMENT

Although both written tests and performance rating measurements are used at OTS, we will limit our discussion to the written tests. These written tests are designed to measure student academic achievement in the areas of communicative skills (CS), leadership and management (LM), human behavior (HB), professional knowledge (PK), defense studies (DS), and field leadership (FL). There are five different CWTs used at OTS,

with four alternate forms of each test. Each test is administered at a specific point in training; namely, day 15, 25, 34, 42, and 52. The CWTs are constructed as criterion-referenced tests and not norm-referenced. There are 36 COs and 295 test items in each set of five tests. The minimum achievement standard for meeting OTS criterion objectives is set at 80%.

PROCEDURES AND RESULTS

The procedures used in test validation include calculations of item analysis, CO analysis, pretest/posttest analysis, and statistical analysis. From item analysis we can tell how each item is functioning by way of ease indexes and tabulations of alternate responses. Since .80 is the minimum cutoff, any item with an ease index of .79 or below is checked for causes of low index reading. The cause may be due to poor item construction, inadequate coverage of material, or insufficient assimilation/retention by students. In item analysis of 900 questions based on results of the past 16 classes (Class 79-06 through 80-04), we have observed 130 items which fell below the ease index of .80. Fifty-eight percent of them were due to construction of test questions, 19% to coverage in instruction, 5% to material in the curriculum, and the remainder to assimilation or retention of material by students. These items were reviewed by the Test Item Review Board consisting of Chiefs of the Academic Division, Measurement Branch, Curriculum Branch, and Academic Instruction Branch.

CO analyses consist of computation of CO failure rates classified according to classes and test forms. In a CO analysis categorized by class, the failure rate should not be more than 15%. If it is greater than 15%, we must again determine whether the high failure rate is due to instruction, the test itself, or the curriculum material. In CO analysis by class, the failure rate was calculated on 36 different COs. The study showed that the failure rate of 15 of them was less than 16%. The remaining 26 COs had failure rates of 16% or greater distributed throughout the 16 classes analyzed. The COs which showed the largest number of item failures were PK-6 with a total of 9, followed by LM-7 with 5, and CS-2 with 4 (see Attachment 1).

In CO analysis by test forms, the failure rate was compiled according to the test form. Each form was administered to four or five classes. The differences of the form averages were obtained by subtracting the lowest average from the highest. The mean of the differences amounted to 5.7. A total of 13 COs indicated differences greater than 5.7. Thirty-nine items from the 13 COs were thus interchanged among the forms in order to balance the difficulty indexes.

Another technique used at OTS to improve the quality of test items is to conduct analysis of achievement gain. This procedure focuses on the proportion answering incorrectly on the pretest, and then correctly on the posttest. If an item showed no significant gain after classroom instruction, that item is discarded or revised. One hundred eighty-five students in Class 78-03 were given a pretest and a posttest. Eight items indicated no significant gain after instruction. The eight items were reviewed and then revised by joint efforts of subject-matter experts and testing specialists.

Statistical analysis of test scores can indicate the reliability of the tests by the amount of variation shown on the mean score (M), standard deviation (SD), reliability index (RI), and standard error of measurement (SE). The amount of differences will depend on whether the tests are standardized or still in the experimental stage. The statistical results of the standardized and experimental tests are shown in Fig 1. STANDARDIZED VS EXPERIMENTAL TESTS.

The amount of variation of the standardized tests was determined by subtracting the statistics of those tests from the average statistics of the total standardized tests. The amount of variation of the experimental tests was obtained by subtracting the statistics of those tests from the results of the standardized tests. The differences of the experimental and standardized tests are indicated in Fig 2. VARIATION OF STANDARDIZED VS EXPERIMENTAL TESTS.

Tests of significance indicated that most of the variation in the statistical parameters of the experimental tests were significantly greater than those of the standardized tests. By revising a total of 62 poor items after the first pretest, 52 after the second, and 26 after the third, we were able to put the experimental tests into the operational inventory. Statistical analyses indicated a significant improvement of the new operational tests as shown in Fig 3. STANDARDIZED, EXPERIMENTAL, AND OPERATIONAL TESTS.

REFERENCES

1. Anatasi, Anne. Psychological Testing, 4th Edition, 1976. Mac-Millan Publishing Co., Inc.

2. Cronbach, Lee J. Essentials of Psychological Testing, 3d Edition, 1970. Harper and Row Publishers.

3. Edwards, Allen L. Statistical Methods, 3d Edition, 1973. Holt, Rinehart, and Winston, Inc.

4. "Evaluation" in Principles and Techniques of Instruction, Air Force Manual 50-62, 1 April 1974. Department of the Air Force.

5. Guilford, J. P. and Bengamin Fruchter. Fundamental Statistics in Psychology and Education, 6th Edition, 1978. McGraw-Hill Book Co.

6. Huck, Schuyler W. "Test Performance under the Condition of Known Item Difficulty." Journal of Educational Measurement, Vol 15, No 1, Spring 1978. National Council of Measurement in Education.

7. Knapp, Thomas R. "The Reliability of a Dichotomous Test Item: A 'Correlationless Approach.'" Journal of Educational Measurement, Vol 14, No 3, Fall 1977. National Council of Measurement in Education.

8. Livingston, Samuel A. and Marilyn Wingersky. "Assessing the Reliability of Tests Used to Make Pass/Fail Decisions." Journal of Educational Measurement, Vol 16, No 4, Winter 1979. National Council of Measurement in Education.

9. Objectives and Tests, Vol III, AFP 50-58, Handbook for Designers of Instructional Systems, 15 July 1978. Department of the Air Force, Washington DC, HQ USAF.

10. Terwilliger, James S. and Kaustubh, Lele. "Some Relationships among Internal Consistency, Reproducibility, and Homogeneity." Journal of Educational Measurement, Vol 16, No 2, Summer 1979. National Council of Measurement in Education.

11. Warm, Thomas A. A Primer of Item Response Theory. Technical Report 941078, October 1978. Department of Transportation, Coast Guard, Oklahoma City.

(806)

CO	DESCRIPTION	79-06	79-07	79-08	79-09	79-10	79-11	79-12	79-13	79-14	79-15	79-16	80-02	80-03	80-04
CS-2	Comm Process	5	4	(23)	5	11	(20)	10	7	(24)	6	11	6	(23)	5
CS-4	Categories of AF Pub	4	4	12	12	4	(17)	(22)	6	11	6	11	3	4	3
CS-5	Fall and Emotional Appeals	7	1	2	1	6	4	1	5	4	2	2	4	4	2
CS-9	Pub Opin & Role of the Media	4	4	12	6	5	4	8	2	4	5	6	1	6	4
LM-5	Lead Authority	0	1	5	1	1	7	3	5	8	4	3	6	10	2
LM-6	Lead Behavior	2	5	7	7	15	13	7	14	(16)	6	4	(16)	13	4
LM-7	Manag Functions and Duties	7	4	(22)	14	15	(19)	(18)	6	(20)	4	4	4	(17)	13
LM-8	Resource Mgt	4	2	9	9	2	2	4	5	2	6	4	2	1	2
LM-11	Mil Justice	(20)	6	13	8	10	4	6	8	10	7	9	11	6	2
HB-3	Gp Dyn;Moti Theo	10	5	(18)	11	10	(26)	11	10	9	7	10	9	11	8
HB-6	Amer Min Exp	5	3	8	6	15	12	7	13	(16)	6	5	(16)	9	5
HB-7	AF Pers & Pers	0	1	0	1	6	3	1	4	2	1	7	4	1	2
HB-8	Drug/Alcohol Abuse Education	0	0	1	2	2	3	3	1	1	3	-	2	2	1
PK-1	AF Cust & Court	4	5	7	2	9	10	7	5	8	6	7	8	2	10
PK-2	Inf/Pers Sec	7	14	3	7	7	7	14	14	10	14	14	9	7	5
PK-6	Pay, Allow&Lv	(19)	(16)	15	15	(25)	(19)	(18)	(18)	11	(21)	-	(24)	13	(23)
PK-8	AF Pers Sysm	4	3	7	13	9	10	12	2	6	3	4	7	7	7
PK-10	AF Entitlements	4	7	3	6	(19)	5	5	(17)	5	6	-	(20)	5	8
DS-2	Man&Conflict	1	3	6	1	7	12	2	7	11	2	2	5	10	1
DS-3	Pow in Conflict	9	4	5	13	14	7	(16)	13	10	14	(26)	12	11	(26)
DS-4	Depart of Defense	4	9	6	(24)	14	6	9	12	4	10	14	7	5	8
DS-5	Employ of Aero Forces I	8	7	9	10	15	11	7	(17)	12	8	9	9	4	6
DS-6	Employ of Aero Forces II	4	11	(25)	6	(19)	15	9	5	12	9	14	13	11	5
FL-1	Orient to Fld Ldr	2	2	4	2	1	10	1	3	6	1	0	2	3	1
FL-6	Drill Orient	14	(18)	9	9	15	15	10	10	14	9	10	12	2	11

SAK-5

(103) Atch 1

Fig 1 STANDARDIZED VS EXPERIMENTAL TESTS
(S) (E)

	No Cases	Mean	Standard Dev	Reliability Index	Standard Error
CWT 1 (S)	400	92.2	5.4	.91	1.6
CWT 1 (E)	400	78.0	8.3	.79	3.7
CWT 2 (S)	355	88.8	6.6	.88	2.3
CWT 2 (E)	355	72.9	9.6	.76	4.7
CWT 3 (S)	330	92.8	5.5	.93	1.5
CWT 3 (E)	330	75.1	9.7	.78	4.6

Fig 2 VARIATIONS OF STANDARDIZED VS EXPERIMENTAL
(S) (E)

	Diff of Mean	Diff of Standard Dev	Diff of Reliability Index	Diff of Standard Error
CWT 1 (S)	1.1	1.3	.10	.4
CWT 1 (E)	14.2	2.9	.12	2.1
CWT 2 (S)	3.5	2.4	.03	.8
CWT 2 (E)	15.9	3.0	.12	2.4
CWT 3 (S)	1.9	1.2	.02	.5
CWT 3 (E)	17.7	4.2	.15	3.1

Fig 3 STANDARDIZED & EXPERIMENTAL VS OPERATIONAL
(S) (E) (O)

	No Cases	Mean	Standard Dev	Reliability Index	Standard Error
CWT 1-1 (S)	400	92.2	5.4	.91	1.6
CWT 1-4 (E)	400	78.0	8.3	.79	3.7
CWT 1-4 (O)	268	91.0	8.1	.90	2.2
CWT 2-2 (S)	355	88.2	6.6	.88	2.3
CWT 2-4 (E)	355	72.9	9.0	.76	4.7
CWT 2-4 (O)	229	88.4	4.9	.88	1.7
CWT 3-1 (S)	330	92.8	5.5	.93	1.5
CWT 3-4 (E)	330	75.1	9.7	.78	4.6
CWT 3-4 (O)	210	87.7	8.1	.87	2.9

SANDS, William A., Acquisition and Initial Assignment Program, Navy Personnel Research and Development Center, San Diego, California.

THE AUTOMATED GUIDANCE FOR ENLISTED NAVY APPLICANTS (AGENA) SYSTEM (Tue A.M.)

The Automated Guidance for Enlisted Navy Applicants (AGENA) system is a computer-based, interactive, vocational guidance system which, in conjunction with a management information system capability, is being developed under the Navy Personnel Accessioning System (NPAS) project.

The AGENA system described in this paper is a demonstration version designed for use with male, non-prior service applicants for enlistment in the U.S. Navy. Three broad functions are provided: vocational guidance, aptitude testing, and personnel assignment. The AGENA system is composed of thirteen modules: (1) Entry, (2) Background Information, (3) System Introduction, (4) Aptitude Screening Test, (5) Interest Inventory, (6) Career Planning, (7) Initial Summary, (8) ASVAB Interpretation, (9) Navy Ratings Available, (10) Navy Training Instrumentality, (11) Final Summary, (12) System Evaluation, and (13) Exit. The presentation sequence and content of each of the modules are discussed.

The Automated Guidance for Enlisted Navy Applicants (AGENA) System

William A. Sands

Acquisition and Initial Service Program

Navy Personnel Research and Development Center

San Diego, California 92152

INTRODUCTION

The Automated Guidance for Enlisted Navy Applicants (AGENA) system is a computer-based, interactive, vocational guidance system which, in conjunction with a management support capability, is being developed under the Navy Personnel Accessioning System (NPAS) project. This research and development project is designed to address the present and future operational requirements of the Navy Recruiting Command. As can be seen in Figure 1, two major areas are addressed by Project NPAS: management support and person-job matching. The management support portion includes data entry, forms generation, and report generation. The person-job matching side of Project NPAS is the AGENA system, which incorporates three major functions: aptitude testing, vocational guidance, and personnel assignment.

AGENA is a system which is presently under development. The version described herein is a prototype, designed for field testing in a small number of recruiting stations. The target population consists of male, non-prior service applicants for enlistment into the U.S. Navy.

SYSTEM MODULES

The demonstration version of the AGENA system is shown in the macro-level flowchart of Figure 2. The system is composed of nine modules: (1) System Introduction, (2) Aptitude Screening Test, (3) Interest Inventory, (4) Career Planning, (5) ASVAB Interpretation, (6) Navy Ratings Available, (7) Related Civilian Occupations, (8) Session/Final Summary, and (9) System Evaluation.

Module #1 - System Introduction

The purpose of the first module is to introduce the applicant to the AGENA system. Following an opening statement of welcome, the applicant will be introduced to the operation of the cathode ray tube (CRT) computer terminal. An explanation of the control keys and functions will be presented and the applicant will have an opportunity to practice using the control keys. A brief description of each of the three AGENA system functions (aptitude testing, vocational guidance, and personnel assignment) will follow. Finally, a list of the nine modules will be displayed to provide the applicant with a "roadmap" of the AGENA system.

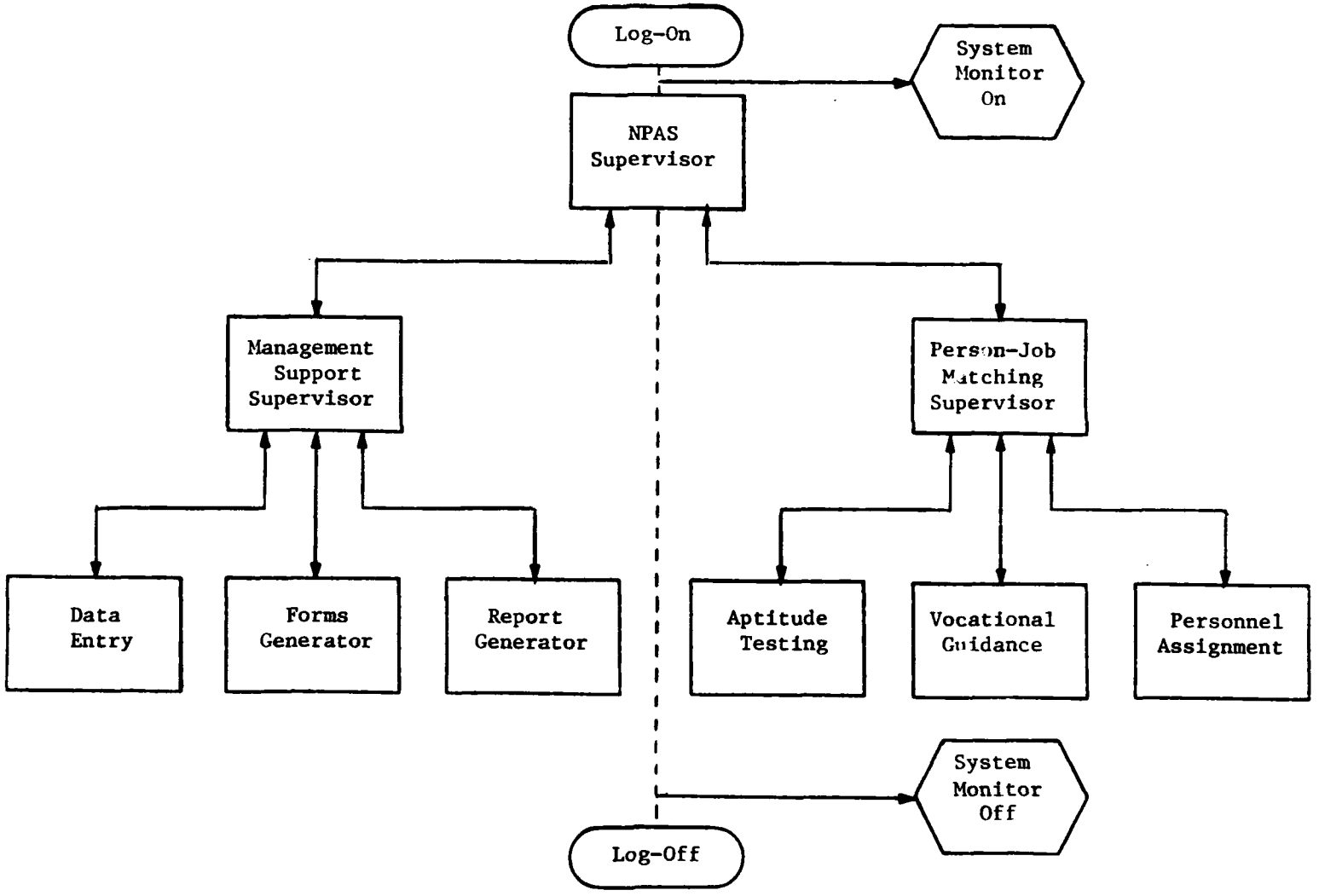


Fig. 1 - Flowchart of the Navy Personnel Accessioning System (NPAS) demonstration

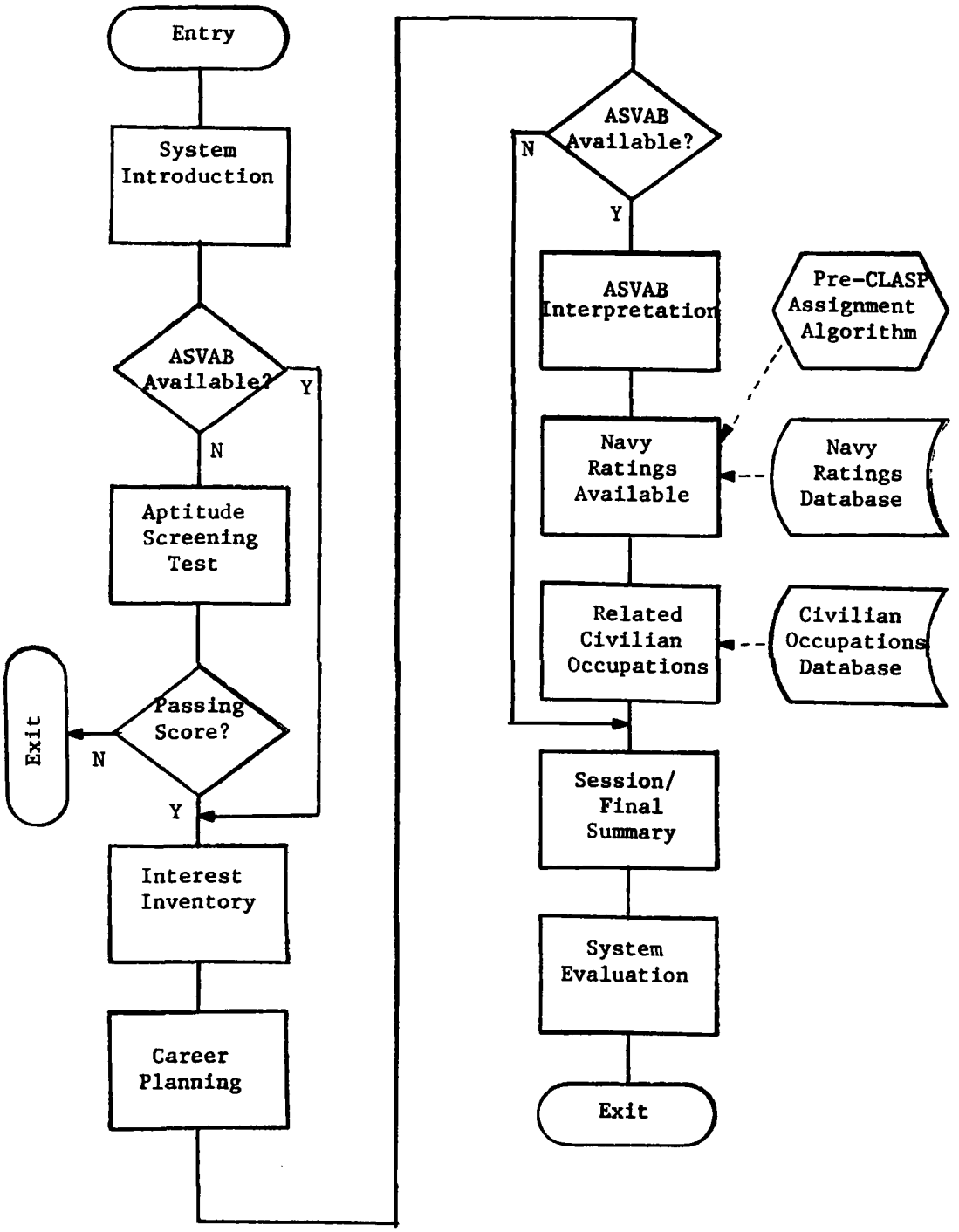


Fig. 2 - Flowchart of the demonstration version of the Automated Guidance for Enlisted Navy Applicants (AGENA) system.

Module #2 - Aptitude Screening Test

As shown in Figure 2, this module will be presented only to those applicants who do not have scores on the Armed Services Vocational Aptitude Battery (ASVAB) available. The module contains the Computerized Adaptive Screening Test (CAST), an adaptively-administered, aptitude screening instrument.[1] The purpose of the CAST is to predict an applicant's performance on the ASVAB. More specifically, the CAST is designed to predict the applicant's Armed Forces Qualification Test (AFQT) score, based upon a linear composite of ASVAB scores. This AFQT score determines the applicant's eligibility for enlistment in the U.S. Navy. Accurate prediction of an applicant's AFQT score is important because significant transportation costs are entailed in sending an applicant to an Armed Forces Entrance Examining Station (AFEES) or a Mobile Examining Test (MET) site for ASVAB testing. The CAST provides the recruiter with an index of an applicant's chances of qualifying for enlistment. If the chances are not promising, the recruiter may elect to save the expense of transporting the applicant for ASVAB testing. For the purpose of the AGENA system demonstration, CAST will replace the Enlisted Screening Test (EST), a conventionally-administered, paper-and-pencil test (Jensen & Valentine, 1976).

Module #3 - Interest Inventory

The third module of the AGENA system contains Form C of the Vocational Interest Career Examination (VOICE).[2] The 245 items consist of occupational titles, work tasks, etc. and will be computer-administered. Scores will be obtained on the following eighteen basic interest scales: (1) office administration, (2) electronics, (3) heavy construction, (4) science, (5) outdoors, (6) medical service, (7) aesthetics, (8) mechanics, (9) food service, (10) law enforcement, (11) audiographics, (12) mathematics, (13) agriculture, (14) teacher/counseling, (15) marksman, (16) craftsman, (17) drafting, and, (18) automated data processing. The norm group which will be used is a nationwide probability sample of 12146 high school students tested during the 1975-76 academic year. The sample was stratified on the basis of race, sex, grade in school, and geographic region (Alley, 1978). An applicant's scores on these basic scales will be displayed on the CRT using bar graphs.

Module #4 - Career Planning

The purpose of the fourth module is to convey to the applicant the value of career planning. The idea that career planning should incorporate information about aptitudes and interests will be presented. If the applicant has

---

[1]The Computerized Adaptive Screening Test (CAST) is being developed by the Psychometric Methods Program at the University of Minnesota under contract to the Navy Personnel Research and Development Center.

[2]The Vocational Interest Career Examination (VOICE) was originally developed by the Educational Testing Service under contract to the Air Force Human Resources Laboratory.

(r14)

not taken the ASVAB at this point, a discussion of ASVAB (Forms 8, 9, and 10) will be presented. This discussion will include the general purpose of the battery and a list and brief description of the ten ASVAB subtests. Finally, the applicant will be encouraged to schedule an appointment for ASVAB testing, as the next module in the AGENA system requires ASVAB scores.

Module #5 - ASVAB Interpretation

The fifth module of the AGENA system will interpret the results of the ASVAB testing to the applicant. Initially, a list and brief description of the component tests will be shown. Then the applicant's performance on the battery will be interpreted in terms of strengths and weaknesses, using bar graph CRT displays. An option will be presented to the applicant to obtain hardcopy output from the printer.

Module #6 - Navy Ratings Available

The enlisted personnel assignment system currently used by the U.S. Navy is the Personalized Recruiting for Immediate and Delayed Enlistment (PRIDE) system. This system has a number of shortcomings, and will be modified by the introduction of the CLASP model, an acronym for Classification and Assignment within PRIDE. CLASP is a real-time, interactive computer model which allows classifiers at AFEES to assign enlisted applicants to ratings in a near-optimal way. Strictly optimal personnel assignment is possible only when the entire pool of people is available simultaneously. The CLASP system makes assignments for people in a sequential fashion, as they arrive.

Two types of factors influence personnel assignments in the CLASP system: individual factors and institutional factors. The individual factors used by CLASP include ASVAB scores and the applicant's occupational preferences. The Navy ratings have been organized into fourteen occupational groups. The applicant will be shown a brief description of each group and then asked to select between one and five of the groups which seem most appealing. Then the preferred groups will be rank-ordered.

The institutional factors which influence personnel assignments in the CLASP system include: task complexity, Navy need, distribution of opportunity for minorities, and level loading of Class "A" schools.

The CLASP utility function includes five components: (1) school success, (2) aptitude/complexity, (3) Navy priority/preference, (4) minority fill-rate, and (5) fraction fill-rate. The first component, school success, is a multiple linear regression estimate of final school grade based upon an optimal set of ASVAB scores. The aptitude/complexity component is a joint payoff based upon technical aptitude and the complexity of the rating. The technical aptitude portion is a linear sum of unit-weighted ASVAB scores, while the complexity portion is based upon paired-comparison judgments of Navy ratings. The third utility component is also a joint payoff based upon the Navy's priority for the rating and the person's preference for the occupational groups. Minority fill-rate, the fourth utility component, represents the payoff to the Navy in assigning a minority to a rating in order to meet parity requirements. This minority fill-rate component acts like a negative feedback loop by employing the discrepancy (positive or negative) from the desired level as a weight. The last component, fraction fill-rate represents the

payoff to the Navy of assigning an individual to a rating in order to meet goaling requirements. Again, like the minority fill-rate component, this component is influenced by the discrepancy between the current and the desired level.

These five utility components are combined to yield a weighted composite payoff that reflects the total utility of assigning a particular person to a particular Navy rating. The weighted composite payoffs (one for each entry level rating) are transformed into optimality indicators using the decision index method introduced by Ward (1958, 1959). The resulting optimality indicators are used to rank-order the total set of entry level Navy ratings from most to least optimal. Finally, consideration is given to the mental, physical, and other qualification standards (e.g., U.S. citizenship) and to the current availability of the rating.[3]

As previously mentioned, the AGENA system is designed for use at Navy recruiting stations. Since actual rating assignments will be made later by a Navy classifier at an AFEES, the AGENA system will use a model called Pre-CLASP. The purpose of Pre-CLASP is to predict what the CLASP system will show as rating options for an individual during the subsequent classification interview at AFEES.

In addition to the data required by the CLASP model, the Pre-CLASP model needs two other inputs: (1) projected arrival date(s) at AFEES for classification, and (2) preferred ship month(s). After extensive computations, the Pre-CLASP model employed in the AGENA system will produce three types of output information for an applicant, given a specified AFEES classification date and ship month: (1) a set of ratings options, (2) an associated set of optimality indicators, and (3) an associated set of probabilities that the ratings will be open.

The top three rating options (based upon optimality indicators) will be displayed to the applicant for consideration. The applicant will have the option to access the Navy Ratings Database[4] and obtain information on any of his rating options. Information on the entry level Navy ratings will be available in two formats. The abbreviated version is designed for CRT display and includes five sections: (1) general description, (2) related civilian jobs, (3) qualifications, (4) working conditions, and (5) Navy opportunities. An extended description, which will be available in hardcopy as an option, includes the five sections of the abbreviated description and three additional sections which cover: (1) what the people in the rating do, (2) sea/shore rotation, and (3) the training provided by the Navy.

---

[3]More detailed information on the CLASP model is available elsewhere (Kroeker, 1979).

[4]The Navy Ratings Database is being developed by the Institute for Behavioral Research, Inc. under contract to the Navy Personnel Research and Development Center.

FILE

If none of the top three ratings appeals to the applicant, he can elect to examine another set of three ratings (optimality indicator ranks four through six). Again, on-line access to the Navy Ratings Database will be available. This process can continue until one of three conditions occurs: (1) the applicant finds a rating that seems a good choice, (2) the rating options for which he is eligible are exhausted, or (3) a total of fifteen ratings (five sets of three) have been examined. Before exiting from this module, the applicant will be requested to choose one or two Navy ratings which seem most promising.

Module #7 - Related Civilian Occupations

The purpose of this module is to communicate to the applicant that Navy training can make a substantial contribution to total career development. The module will begin with a brief discussion of the general value of Navy training and experience. Then, the applicant will have on-line access to the Civilian Occupations Database.[5] Descriptions of civilian occupations (or clusters of occupations) related to the Navy rating(s) he has selected will be available on the CRT and/or as hardcopy output from the printer. Descriptions of civilian occupations will include five sections: (1) general description, (2) qualifications and training, (3) pay and working conditions, (4) employment outlook, and (5) related Navy jobs. Finally, a brief discussion of additional benefits of Navy enlistment (e.g., medical benefits) will be presented.

Module #8 - Session/Final Summary

This module summarizes the results of the present session on the AGENA system. If the applicant has completed the first seven modules, a final summary is presented. This final summary covers the results of all sessions on the system, and includes information on the aptitude screening test (CAST), the interest inventory (VOICE), the classification test battery (ASVAB), the Navy ratings explored, the rating(s) selected, and the related civilian occupations examined.

Module #9 - System Evaluation

The purpose of this final module is to obtain the applicant's evaluation of the AGENA system. A series of multiple-choice questions will be administered on-line. The information obtained will be used as feedback to help modify and improve the AGENA system.

SUMMARY

The AGENA system should provide a number of benefits. The Computerized Adaptive Screening Test (CAST) should enhance the accuracy of predicting the Armed Forces Qualification Test (AFQT) score, thereby reducing decision errors

---

[5]The Civilian Occupations Database is being developed by the Institute for Behavioral Research, Inc. under contract to the Navy Personnel Research and Development Center.

and the attendant unnecessary transportation costs. The vocational guidance and opportunity for applicants to explore the enlisted Navy world of work should result in improved person-job matching, benefiting both the individual and the Navy. Finally, the management support side of the demonstration system being developed under Project NPAS should provide a significant reduction in the paperwork for recruiters, while, at the same time, reducing clerical errors. In summary, the demonstration system promises to have substantial benefits, both for the enlisted applicant and for the Navy.

REFERENCES

Alley, W.E. Vocational Interest-Career Examination: Use and Application in Counseling and Job Placement. Technical Report AFHRL-TR-78-62. Brooks Air Force Base, Texas: Air Force Human Resources Laboratory, October 1978.

Jensen, H.E. & Valentine, L.D., Jr. Development of the Enlistment Screening Test - EST Forms 5 and 6. Technical Report AFHRL-TR-76-42. Brooks Air Force Base, Texas: Air Force Human Resources Laboratory, May 1976.

Kroeker, L. Policy Specifying, Judgment Analysis, and Navy Personnel Assignment Procedures. Proceedings of the 21st Annual Military Testing Association Conference, October 1979.

Ward, J.H., Jr. The Counseling Assignment Problem. Psychometrika, 1958, 23, 6-16.

Ward, J.H., Jr. Use of a Decision Index in Assigning Air Force Personnel. Technical Note WADC-TN-59-38. Lackland Air Force Base, Texas: Personnel Laboratory, Wright Air Development Center, Air Research and Development Command, April 1959.

SCANLAND, Dorothy VonK., Ed.D., and SCANLAND, Worth, Ph.D., Defense Activity for Non-Traditional Educational Support, and Naval Education and Training Command, Pensacola, Florida.

EVALUATION CRITERIA FOR PERSONNEL SELECTION (Tue A.M.)

RESEARCH PROCEDURE

In order to establish with a fair amount of validity those traits and characteristics which most employers, supervisors and others who need to rate people believe to be important in the selection of others to work for them, several steps were initiated, as follows:

(1) A set of words and phrases which are descriptive of people was assembled using a previously assembled list of 700 such words.

(2) Enough meaningful and non-meaningful words and phrases were assembled to allow for selection of about ten pairs of these. Those selected from the list must have equal social acceptability.

(3) Sets of two pairs of words were assembled into tetrads.

(4) The tetrads were then arranged on a form and the rater required to make a check mark opposite two of the words or phrases in each set, indicating his choice of that which is most descriptive of the ratee, or indicating that which is least descriptive, depending upon the instructions.

The key to the selection of the words and phrases is that they must have significant differences in their respective discrimination indices, while at the same time having equal preferential indices.

The conduct of the research and the outcomes are described.

"EVALUATION CRITERIA FOR PERSONNEL SELECTION"

Dorothy von K. Scanland, Ed.D.
Worth Scanland, Ph.D.

A paper for
presentation before the annual conference
of the Military Testing Association
Toronto, Canada, 27-30 October 1980

INTRODUCTION

In all areas of personnel management, in and out of the military services, the selection of the best and most qualified people to fill position vacancies, or for promotion to higher levels of responsibility, is a matter of concern to both those in competition and those charged with making the selections. Yet in those categories of criteria which one might label as traits and characteristics much is left to intuition when selecting those to be measured and how best to obtain reliable evaluations. The measurement of capabilities is a much simpler task—one can often administer a task related performance test, or at least a verbal (written or oral) test designed to demonstrate within acceptable validity the capabilities of the subject to execute certain functions which are job-relevant or predictive of future performance on the job. But who can say whether seriousness is a more valuable trait than, say zeal? And if one chooses to take a position on such a question, on what research or experience is he able to justify that position? This paper offers one method by which these questions can be answered for the design of a new Supervisor's Appraisal form for evaluating applicants for a new position.

BACKGROUND

The making of the wisest selection possible in the choice of new personnel to fill vacant positions in any business, profession, trade or occupation, as well as the planning for training and education of the mind and hands for the achievement of superior proficiency in one's chosen field are matters of deep concern to all supervisors and other personnel managers. These vital problems should be solved, as Balkin (1931) says, in a "careful scientific way, with due regard for each person's attributes." (Underscoring is the writer's.) When St. John Baptist de la Salle (1651-1719), the Patron of all instructors, felt the need for a guide in the choice of teachers, he wrote his "Conduct of the Christian Schools." In this book he included a simple list of the eleven virtues for a good teacher: seriousness, silence, humility, prudence, wisdom, patience, restraint, zeal, watchfulness, piety and generosity. The Boy Scouts have their twelve attributes of perfection which are, we are confident you will remember, "trustworthy, loyal, helpful, friendly, courteous, kind, obedient, cheerful, thrifty, brave, clean and reverent." While one may now, some years later, take issue with these lists, the significant point is that even in St. John's time there was an appreciation for the fact that personal attributes suit or unsuit one for certain activities.

The significant question for which we need the answer is a two part one: what are the important attributes, or capabilities, or traits which bear upon an individual's qualifications or suitability for a certain endeavor, and if we can arrive at some such list, are these capabilities measurable or comparable between individuals?

The judgment of character (another way of saying the measurement of capabilities and traits) is an indirect process in which, from a momentary observation or a cross-section of the life of another, one attempts to estimate its general quality or tenor. As Hollingsworth (1922) describes it, this judgment is a sort of diagnosis, relying on symptoms, clues and signs and other very incomplete evidence. This diagnosis of human character is an enterprise in which we engage all our lives, yet we are far from being accurate, even in judging the characters of those very near to us over long periods of time. Furthermore, according to Wechsler (1952), the variability of human characteristics, when compared to that of other phenomena in nature, is extremely limited, and the differences which separate human beings from one another with respect to whatever trait or ability we may wish to compare, are far smaller than one would suppose. If then, despite years of practice we seem to gain little proficiency in the judgment of others, and if too the differences we are attempting to judge are slight, or at least small, compared with other natural phenomena, are we faced with an impossible task in designing a system for this very purpose? Rugg (1922) adds further pessimism to this gloomy picture by reminding us of the difficulty—almost impossibility—of securing agreement among several people in judging character, except erroneously when there is a strong (usually objectionable) personality trait, in which case all traits are judged low on a measurement scale.

But we are not facing a hopeless undertaking. Blackford and Newcome (1916) point out for us that the perfection of a science of knowing ourselves will require the united efforts of many investigators, experimenters and practical workers, such as teachers, employers, parents and other men and women everywhere. When some uniform method of studying human attributes is adopted and made a part of our school systems, colleges and universities, industry, government, the military, etc., uniform records kept and analyzed, careers studied, we can in one generation have sufficient information on which to base a science of human character analysis. So far we are a long way from this happy day, but meanwhile we feel justified in attempting to add to the general experience in the field.

As a start at designing a rating form it might be well to find a definition of "rating." Good (1959) says that a rating is "an estimate made according to some systemized procedure, of the degree to which an individual person or thing possesses any given characteristic; it may be expressed qualitatively or quantitatively." A rating scale, then, would be some device for expressing this estimate along a continum. The measuring device itself is not a format or scale, of course, but a human rater. So there are limitations which it is important to understand, limitations besides those imposed by the rating form itself, no matter how fine that instrument. These are the limitations of the rater. His inevitably selective perception, memory (and lack of it), lack of sensitivity to what may be psychologically and socially important, plus his inaccuracies of observation all limit his ability to measure truly. Obviously,

therefore, it is imperative that every effort be made to design a rating system which will be as uninfluenced by these limitations as possible. It is appropriate at this point, then, to examine some of the characteristics of rating systems which will be of concern to the designer.

Remmers (1963) lists five characteristics of rating systems which are significant to the effectiveness with which they may be used. These are:

(1) Objectivity - meaning that use of the instrument should yield verifiable, reproducible data not a function of the peculiarities of the rater;

(2) Reliability - meaning that it should yield the same values, within the limits of allowable error, under the same set of conditions;

(3) Sensitivity - meaning it should yield as fine distinctions as are usually made in communicating about the subject of the rating;

(4) Validity - meaning that the categories in the scale should be relevant to a defined area of investigation and to some relevant behavioral science construct;

(5) Utility - meaning it should efficiently yield information pertaining to the practical requirement.

These characteristics, if incorporated into the rating system properly, will help to eliminate errors in judgment which are found to be common among many raters. The most pronounced of these tendencies to err are described by Guilford (1954); as he has said, they will probably invalidate the measurement if not controlled. These errors may be stated as follows:

(1) the error of leniency, where the rater tends to rate those whom he knows well higher than others whom he does not know well;

(2) the error of central tendency, where the rater hesitates to rate others at the extremes of the scale; and

(3) the "halo effect," which is a tendency to rate individual traits in the direction of the overall opinion the rater has of the person being rated.

The major task in designing a rating system, then, would seem to be to arrive at one which contains the maximum number of desirable characteristics and which allows the minimum amount of error from the rater to creep into the evaluation. Obviously one form of scale will tend less to error of some kind than will another, and there are strong points and weak points which may be found in each. The perfect scale on which total reliance may be placed has not yet been invented. When casting about for a scale or form for a certain requirement, one must carefully consider the various kinds commonly in use, and find that one which will render minimum error under the conditions anticipated.

X23

Guilford (1954) has divided rating scales into five basic classifications, as follows: numerical, graphic, standard, cummulative points and forced choice. These may be briefly described this way:

> The numerical method assigns a number of each of several descriptive words or phrases arranged in some suitable ascending or descending order;
>
> The graphic method arranges descriptive adjectives concerning a trait in sequence so that the rater may check one;
>
> The standard scales are a set of standards by which the rater may compare the traits of the subject;
>
> The cummulative points system is a compilation of good (positive) and bad (negative) traits and characteristics which are checked for the person being rated, and the algebraic sum of the checks is the rating;
>
> The forced choice system forces the rater to decide whether the person being rated has more of one trait than another of a pair. One of the members of the pair is valid for predicting some total quality, and the other is not, but both are equally socially acceptable.

Of these five basic systems, the only one which seems to find support in theoretical construct is the forced choice system. In this system the rater is presented with several sets of adjectives and phrases which characterize people. Rather than indicating how much or how little of these characteristics the ratee, in the rater's opinion, possesses, the rater is required to choose from among the several adjectives in each set that one which comes nearest to describing the ratee. Or he may be required to choose that adjective which in his opinion is least descriptive of the ratee. Because of the way in which the sets of characteristics are constructed, there is a strong reduction in the rater's ability to produce any desired outcome by the choice of obviously good or obviously bad traits. For this reason there is almost no room for personal bias or favoritism. Tests conducted by Sisson (1948) show the forced-choice system to be free from pile-ups at either end of the scale, and that it produces ratings which are objective and high in both reliability and validity. Although there are some who have taken issue with the claims for this system, as for example Travers (1951), it is the conclusion of these authors that it offers the least chance for serious error of the several systems possible.

To these writers the most serious defect of the forced-choice system is the tremendous amount of work which is involved in fully developing a particular form for a particular purpose. It is believed, however, that a good deal of this work may be reduced to a manageable level if one is agreeable to making certain assumptions which do not appear to be risky. Remmers (1963) goes into considerable detail on the techniques of constructing the forced-choice scale; in deciding upon the forced-choice system for this discussion the acceptance of some deviations from the full technique were accepted for the sake of financial savings as well as time savings. These shortcuts do not appear, however, to be fatal to the usefulness of the method. The steps involved in the construction of the scale for a proposed new Supervisor's Appraisal to replace the current form (Appendix A) for evaluation of applicants for new positions are described hereinafter.

PROCEDURE AND DATA

The initial step in the procedure for constructing a forced choice scale was to review the full recommended procedures described in Remmers (1963), and to re-describe this procedure in the light of our requirements, as well as the time and effort available for the job. The re-description of the full Remmers procedure provided a set of steps which are set forth as follows:

(1) There must be collected a list of words and phrases which are descriptive of people; this list must be of sufficient length to assure that all meaningful words and phrases are included;

(2) The above list must then be reduced to a manageable length, after which it must be subjected to some treatment which will separate the truly meaningful descriptive words and phrases from those which are not meaningful (that is, meaningful in the context in which the rating form is to be used);

(3) Enough meaningful and non-meaningful words and phrases must be arrived at to allow for the selection of about ten pairs of these. Those selected from the list must have equal or nearly equal social acceptability, that is to say, pairs of words or phrases must give about equal praise or condemnation between the two words or phrases in the pair;

(4) Sets of two pairs of these words and phrases are then to be assembled into tetrads; about five or six tetrads will be sufficient for the rating scale envisioned, particularly if one does not wish to risk losing the interest of the rater using the form;

(5) The tetrads will then be arranged on a form and the rater required to make a check mark opposite two of the words or phrases in each set, indicating his choice of the word or phrase which is most descriptive of the ratee, or indicating that which is least descriptive of the ratee, as the instructions may indicate he is to do.

The key to the selection of the words and phrases which comprise the sets is that they must have significant differences in their respective discrimination indices (that is, one word must be very meaningful as a trait when considered in a future employee, while the other word has relatively no meaningfulness), while at the same time having equal preferences indices (that is, be descriptions which are of equal social significance or desirability). The steps outlined above, leading to the selection of the important words and phrases, were performed as follows:

(1) Wherry's (1951) original work for the U.S. Army Personnel Research Division is devising a forced-choice rating system for the Army officer performance rating forms produced a list of about seven hundred words and phrases which might be used in describing another person's traits and characteristics. The Army was requested to

X25

supply a copy of this exhaustive list, which it very graciously did. This contribution, of course, saved many hours of research.

(2) The 700 adjective words and phrases obtained from the Army study were subjected to a careful review with the objective of reducing them to a manageable number which could be used in a survey. In the end one hundred fifty words and phrases were selected for compilation into a questionnaire-type form.

(3) Approximately one hundred fifty commercial employers were randomly selected from the files of the Office of the Florida Secretary of State, and to this list were added the names of all county school supervisors in the State of Florida, the presidents of all the colleges and junior colleges in Florida and Georgia, and a number of other educators. A letter of instructions together with the list of one hundred fifty words and phrases were then mailed to these individuals with the request that the list be reviewed and a check made against those fifteen to twenty-five traits which they considered as being significant and those they considered insignificant to their hiring decisions. Responses were received from approximately 50% of the addressees, response being bettter from the educational fraternity than from the business firms.

(4) The lists returned from the respondents were tallied on a master sheet to determine which, if any, traits and characteristics stood out as being significantly meaningful to a majority of employers, and which, if any appeared to hold universal insignificance to the same people.

(5) The fifteen adjective words and phrases indicated as meaningful to the most respondents were extracted from the total list. The fifteen adjective words and phrases indicated as the least meaningful to the most respondents were likewise extracted. An unpatterned mixture of these thirty words was assembled into a modified Q-sort and distributed to approximately 400 education specialists in the Navy, with the request that the words be re-sorted into an order of social import. A copy of this listing is shown as Appendix B. The result of this survey, from which almost 200 responses were received and analyzed, was to indicate that there is practically no way in which these thirty words and phrases could be placed in a descending order of social acceptability which would represent the consensus of those surveyed.

(6) From the above information five tetrads of adjective words and phrases were constructed. Each tetrad included four words or phrases, two of which were known from the survey to hold high significance to employers as traits in a prospective employee, and two of which were likewise known to have little or no such significance. All four of the words or phrases, however, were of essentially equal social acceptability as shown by step five above. These five tetrads formed the basis of a trial rating sheet.

(7) The trial rating sheets were subjected to two evaluative procedures designed to determine if they can fulfill the purpose for which developed. The results of these evaluations will be included in a later discussion paper.

RESULTS

The survey of prospective employers

The most interesting result of the survey of prospective employers was the unanimity with which they view a small number of traits as being very significant to their hiring decisions, and a large number of other traits as being wholly insignificant to this decision. This can be interpreted to mean that there are a few (ten to fifteen) traits of the one hundred fifty listed which are, in the opinion of the majority of those responding, considered significant to the hiring decision, and there is a wide range of traits which by and large are considered unimportant in a prospective employee. This information is sufficient to provide a reasonable basis for matching of traits into pairs which are highly discriminatory. Below are listed the traits, attributes and characteristics which at least sixty per cent of the respondents indicated as significant to their hiring decisions (listed in the descending order of their popularity):

Enthusiastic
Emotionally mature
Self-controlled
Cooperative
Tactful
Possessing initiative
Well informed
Tolerant
Having sound judgment
Well organized
Can speak well
Energetic
Ethical

Now here are the traits, attributes and characteristics which less than 10 per cent of the respondents indicated were of significance in their hiring decisions (listed in the descending order of their popularity):

Conservative
Intense
Debonair
Quiet, subdued
Reserved
Neat in dress
Suave
Handsome
Sophisticated
Modest
Devout
Generous
Politically unbiased
Good posture
Good athlete
Perfectionist
Refined
Altruistic
Kind
Relaxed
Polite
Amiable
Methodical
Gregarious

The final operation upon the material received as a result of the survey was to select pairs of traits which are equally acceptable from a social point of view but which are in fact discriminatory from the standpoint of being desirable in hiring future employees. To do this the two lists above were thoroughly and randomly intermingled so that one unfamiliar with the original listing would have no way of determining (at least from the survey's standpoint) which are discriminatory and which are not. This composite list was then given to several hundred Navy civilian education specialists with the request that they re-arrange the words and phrases in a descending order of social acceptability. Because there was no unanimity among the results of this effort (and this is exactly what was hoped for), it served to demonstrate that, with minor exceptions, the majority of the traits in this composite list is acceptable from a social standpoint. From here, then, the pairings of traits and characteristics were made without any special discrimination on the part of the form drafter except to be sure that each pair contained a trait from the "significant" list and a trait from the "not significant" list. Ten pairs were made in this fashion, then grouped into five tetrads as follows:

First tetrad	Is well informed	Second tetrad	Is unwavering, steadfast
	Is methodical		Has initiative
	Is tolerant		Is a perfectionist
	Is amiable		Shows sound judgment
Third tetrad	Is kind	Fourth tetrad	Has self-control
	Is tactful		Is emotionally mature
	Is cooperative		Is relaxed
	Is generous		Is conservative

Fifth tetrad	Is enthusiastic
	Is quiet, reserved
	Is modest
	Is well organized

These tetrads are to undergo an evaluation for the reviewers criteria, and if the new system proves superior to the current form, it will be recommended to the Navy's Civilian Personnel Office for consideration.

BIBLIOGRAPHY

Baier, D. E. "Reply to Travers' 'A Critical Review of the Validity and Rationale of the Forced-Choice Technique'," Psychol. Bull., Vol.48 (September, 1951), pp. 421-434.

Balkin, Harry H. The New Science of Analyzing Character. Phila.: David McKay Co., 1931.

Blackford, Katherine M. and Newcomb, Arthur. Analyzing Character, the New Science of Judging Men. New York: The Review of Reviews Co., 1916.

Edwards, Allen L. The Social Desirability Variable in Personality Assessment and Research. New York: Dryden Press, 1957.

Good, C. V. (ed.). Dictionary of Education. 2d ed. New York: McGraw-Hill, 1959.

Grande, Luke M. Twelve Virtues of a Good Teacher. New York: Sheed and Ward, 1962.

Guilford, J. P. Psychometric Methods. 2d ed. New York: McGraw-Hill, 1954.

Highland, R. W. and Berkshire, J. R. "A Methodological Study of Forced-Choice Performance Rating," Research Bulletin (1951), pp. 51-59. San Antonio, Tex.: Human Resources Research Center.

Hollingsworth, Harry L. Judging Human Character. New York: Appleton-Century-Crofts, 1922.

Paterson, D. C. "Methods of Rating Human Qualities," Ann. Amer. Acad. Pol. and Soc. Sci., Vol. 110 (November, 1923), pp. 81-93.

Remmers, H. E. "Rating Methods in Research on Teaching," Handbook of Research on Teaching. New York: Rand McNally & Co., 1963.

Rugg, H. O. "Is the Rating of Human Character Practicable?" J. Educ. Psychol., Vol. 13 (January, 1922), pp. 31-42.

Shen, E. "The Reliability Coefficient of Personal Ratings," J. Educ. Psychol., Vol. 16 (April, 1925), pp. 232-236.

Sisson, E. D. "Forced-Choice - The New Army Rating," Personal Psychol., Vol. 1 (Autumn, 1948), pp. 365-381.

Slawson, J. "The Reliability of Judgments of Personal Traits," J. Applied Psychol., Vol. 6 (June, 1922), pp. 161-171.

Taylor, E. K. and Wherry, R. W. "A Study of Leniency in Two Rating System," Person. Psychol., Vol. 4 (Spring, 1951), pp. 39-47.

Travers, R. M. W. "A Critical Review of the Validity and Rationale of the Forced-Choice Technique," Psychol. Bull., Vol. 48 (January, 1951), pp. 62-70.

Uhrbrook, R. S. "Rating Tendencies of Personally Selected Judges," J. Educ. Psychol., Vol. 23 (November, 1932), pp. 594-603.

Wechsler, David. The Range of Human Capabilities. 2d ed. Baltimore: Williams and Wilkins Co., 1952.

(830)

APPENDIX A: ANNUAL MERIT PROMOTION SUPERVISOR APPRAISAL

-10 AND ABOVE IN GRADED SERVICE FOREMAN AND ABOVE IN WAGE GRADE SERVICE/WAGE SUPERVISORY PRODUCTION FACILITATING

EMPLOYEE'S NAME (PRINT LAST NAME AND INITIALS)

SOCIAL SECURITY NUMBER

SOC. SECURITY NUMBER 1 2 3 4 5 – 6 7 8 9

DO NOT DENT OR CREASE THIS SHEET.
USE ONLY NO. 2 PENCIL.
PLACE SHEET ON A SMOOTH, HARD SURFACE.
DO NOT LET MARKS EXTEND INTO ADJOINING BOXES.
MARK ONE BLOCK IN EACH HORIZONTAL LINE IN EACH DATA GROUPING

NAME (FIRST 3 LETTERS OF LAST NAME)

NAME 1 2 3

DATE PREPARED

DATE D D M M Y Y

APPRAISAL
(USE APPRAISAL ELEMENTS ON REVERSE SIDE OF FORM)

BASED ON PAST PERFORMANCE

	UNSATISFACTORY	BELOW AVERAGE	AVERAGE	ABOVE AVERAGE	OUTSTANDING
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					

BASED ON SUPERVISORY POTENTIAL

	UNSATISFACTORY	BELOW AVERAGE	AVERAGE	ABOVE AVERAGE	OUTSTANDING
SUPERVISORY 1					
SUPERVISORY 2					
SUPERVISORY 3					
SUPERVISORY 4					
ORG & MET 1					
ORG & MET 2					
ORG & MET 3					
ORG & MET 4					
COM & DEC 1					
COM & DEC 2					
COM & DEC 3					
COMMUNICATION 1					
COMMUNICATION 2					
COMMUNICATION 3					
COMMUNICATION 4					
O/A PERSONAL 1					
O/A PERSONAL 2					
O/A PERSONAL 3					
O/A PERSONAL 1					

PRIVACY ACT STATEMENT

GENERAL: This information is provided pursuant to Public Law 95-579 (Privacy Act of 1974), December 31, 1974 for individuals completing this form.

AUTHORITY: Sections 3301 and 3302 of Title 5 of the United States Code, E.O. 10577; and 3 CFR, 1954-1958 Comp., p. 218 gives activities authority to transact promotions and internal placements, which includes the evaluation and ranking of employees qualifications for promotion.

PURPOSE AND USE: The purpose of this form is to collect information needed to aid in determining the quality of qualifications of current Federal employees for promotion purposes.

EFFECTS OF NONDISCLOSURE: Ommission of an item may eliminate you from consideration for a position you are seeking.

INFORMATION REGARDING DISCLOSURE OF YOUR SOCIAL SECURITY NUMBER: Disclosure by you of your Social Security Number (SSN) is mandatory to obtain the services, benefits or processes that you are seeking. The SSN is used as an identifier throughout your Federal career from the time of application through retirement. The information gathered through the use of the number will be used only as necessary in personnel management processes carried out in accordance with established regulations and published notices of systems and records. The request for the SSN is necessary because of the large number of present and former Federal employees and applicants who have identical names and birth dates and whose identities can only be distinguished by the SSN.

I HAVE REVIEWED THIS RATING IN AN INTERVIEW WITH MY SUPERVISOR. MY SIGNATURE DOES NOT NECESSARILY INDICATE MY CONCURRENCE WITH THE EVALUATION OR RATING.

SIGNATURE OF EMPLOYEE | DATE

SIGNATURE OF IMMEDIATE SUPERVISOR | DATE | SIGNATURE OF ENDORSING OFFICAL | DATE

NASP 12340/4 (REV. 12-78)

937

SUPERVISORY POTENTIAL

APPRAISAL ELEMENTS

BASED ON PAST PERFORMANCE	1. CAN PRESENT & SUPPORT AND ADJUST BY IDEAS 2. HAS GOOD JUDGMENT ON WHOM TO ASSIGN TO WHICH JOBS ARE GOOD 3. IS ABLE TO DEAL EFFECTIVELY WITH OTHERS EVEN WITH PEOPLE WHO ARE OPPOSED TO HIM 4. ABILITY TO OBTAIN AND MAINTAIN COOPERATION 5. PERCEPTIVE IN ESTIMATING QUALITIES OF OTHERS 6. ABILITY TO DEAL WITH ORGANIZED GROUPS INCLUDING MINORITY GROUPS ETC 7. WHEN THINGS GO WRONG HE WORKS TO RECTIFY THEM INSTEAD OF MAKING EXCUSES 8. RELIABILITY AND DEPENDABILITY 9. CAN HANDLE SEVERAL DIFFERENT PROBLEMS AT THE SAME TIME 10. WOULD GIVE AN HONEST REPORT ON A PROBLEM EVEN IF IT WOULD NOT BE TO HIS ADVANTAGE 11. EXEMPLIFIES FIRST RATE TRADE KNOWLEDGE AND SKILL 12. WILLING TO LISTEN TO IDEAS OF OTHERS 13. EXHIBITS GOOD JUDGMENT IN ALL FACETS OF DECISION MAKING 14. PLANS AND ORGANIZES APPROACH TO JOB DUTIES SPECIAL PROJECTS AND PROBLEMS 15. HAS CONFIDENCE IN HIS ABILITY AND KNOWLEDGE OF JOB 16. SEEKS ACCEPTS AND APPROPRIATELY DISCHARGES RESPONSIBILITY 17. FOLLOWS THROUGH TO SEE THAT WORK IS ON SCHEDULE 18. EFFECTIVE IN THINKING OF NEW IDEAS AND SOLUTIONS 19. SETS PRIORITIES EFFECTIVELY 20. OVERALL EVALUATION
SUPERVISORY	1. THE CANDIDATE WOULD DEFINE ASSIGNMENTS OR PROJECTS CLEARLY 2. DELEGATE AUTHORITY AND RESPONSIBILITY AND WORK WITH AND THROUGH OTHERS EFFECTIVELY 3. ESTABLISH AND MAINTAIN HIGH STANDARDS OF QUALITY AND QUANTITY FOR THE WORK PRODUCED 4. BE FAIR AND OBJECTIVE IN DEALINGS WITH AND JUDGMENTS OF SUBORDINATES 5. MOTIVATE TRAIN DEVELOP AND GUIDE EMPLOYEES OF VARIED BACKGROUNDS AND SKILL LEVELS EFFECTIVELY
ORGAN & MGT	1. THE CANDIDATE WOULD DEVISE ECONOMICAL AND EFFECTIVE ORGANIZATIONAL OR OPERATIONAL PLANS AND PROCEDURES 2. ESTABLISH PROGRAM OBJECTIVES OR PERFORMANCE GOALS AND ASSESS PROGRESS TOWARD THEIR ACHIEVEMENT 3. ADJUST WORK ACTIVITIES AND SCHEDULES TO MEET EMERGENCY CONDITIONS OR UNANTICIPATED REQUIREMENTS 4. COORDINATE AND INTEGRATE THE WORK OF SUBORDINATE EMPLOYEES OR ORGANIZATIONAL SEGMENTS EFFECTIVELY
COMMIT/DECI	1. THE CANDIDATE WOULD ANALYZE COMPLEX ISSUES OR PROBLEMS THOROUGHLY AND QUICKLY 2. ASSESS THE ADVANTAGES AND DISADVANTAGES OF ALTERNATIVE PLANS OR COURSES OF ACTION 3. MAKE SOUND DECISIONS [illegible] BASED ON PAST EXPERIENCE, PRESENT EFFORT AND FUTURE OUTCOME
COMMUNICATIONS	1. THE CANDIDATE WOULD COMMUNICATE EFFECTIVELY WITH MANAGEMENT, AND EMPLOYEES, AND WHERE APPROPRIATE EMPLOYEE GROUPS 2. BE SKILLFUL IN ORAL AND WRITTEN COMMUNICATIONS 3. MAINTAIN POISE, HANDLE CONTROVERSIAL OR DELICATE MATTERS SKILLFULLY 4. PROVIDE MANAGERIAL LEADERSHIP AND FULL PARTICIPATION IN ACTIVITIES WITHIN OR OUTSIDE HIS ORGANIZATION WHICH FOSTERS GOOD GOVERNMENT
PERSONAL	1. THE CANDIDATE WOULD ADJUST TO CHANGE WORK PRESSURES OR DIFFICULT SITUATIONS WITHOUT UNDUE STRESS 2. BE ABLE AND WILLING TO [illegible] PEOPLE AND SITUATIONS AS THE FACTS 3. HAVE A POSITIVE OUTLOOK TOWARD THE WORK AND THE EMPLOYING [illegible]
OVR	1. WHAT [illegible] OVERALL LEVEL OF ABILITY TO SUPERVISE

(832)

APPENDIX B

LIST OF ATTRIBUTES

1.	Ethical	BOX I	1
2.	Modest		2
3.	Quiet, subdued		
4.	Enthusiastic		1
5.	Neat in dress		2
6.	Self-control		3
7.	Methodical	BOX II	4
8.	Emotionally mature		5
9.	Conservative		
10.	Speaks well		1
11.	Extroverted		2
12.	Cooperative		3
13.	Reserved		4
14.	Sound in judgment		5
15.	Sophisticated		6
16.	Well organized		7
17.	Tactful	BOX III	8
18.	Politically unbiased		9
19.	Relaxed		10
20.	Has initiative		11
21.	Generous		12
22.	Not lazy		13
23.	Kind		14
24.	Very polite		15
25.	Tolerant		16
26.	Refined		
27.	Well informed		1
28.	A perfectionist		2
29.	Devout	BOX IV	3
30.	Amiable		4
			5
		BOX V	1
			2

SEUBERLICH, Col. H.E., DBwV (Federal Armed Forces Association),
Bonn, West Germany.

A GERMAN MODEL OF ASSESSMENT PROBLEMS (Tue P.M.)

The present model is being introduced as an aid to reaching a decision on the assessment and pay grading for the functions of officials and servicemen according to specific requirements criteria. This model has been elaborated by the Federal Ministry competent for pay matters.

This model has been set up on the basis of a "normal performance" of a job. The assessment consists in the investigation according to specific characteristics of the positions, the comparison between same and the gradation thereof.

The system covers seven assessment characteristics, differently evaluated according to their importance. The components of the assessment characteristics have been established according to "designation rating".

Moreover, there is an assessment table for each assessment characteristic. This table is subdivided into "assessment degrees", which again are correlated to a value grading figure. From the total of the value grading figures the "position value" is obtained.

In order to enable the classification of the individual positions within an assessment order, the positions have been classified according to the order of succession of the relevant value in a "series of grades".

The main features of the model and of the assessment characteristics system bearing on the actual situation as in 1980 with advantages and disadvantages for the servicemen, are being discussed from the viewpoint of the Federal Armed Foces Association.

A GERMAN MODEL OF ASSESSMENT PROBLEMS

Col. H.E. Seuberlich, Federal Armed Forces Association

1. Introduction into the Assessment Model

The Federal Pay Act provided under Art. 18 for a pay according to function for officials, judges and servicemen. This includes three phases:

(1) the assessment of the functions,
(2) the coordination of the functions into positions,
(3) the coordination of the positions into pay grades.

The following Model serves as an aid to reaching a decision on phase (1). It has been elaborated by the Federal Ministry competent for Pay Matters and has been drawn up according to the analytical viewpoints, requiring the assessor to get thouroghly acquainted with the positions to be assessed as well as the surroundings thereof. This Model has been set up as a "framework" with a view to enable the application thereof in all sectors of public utility.

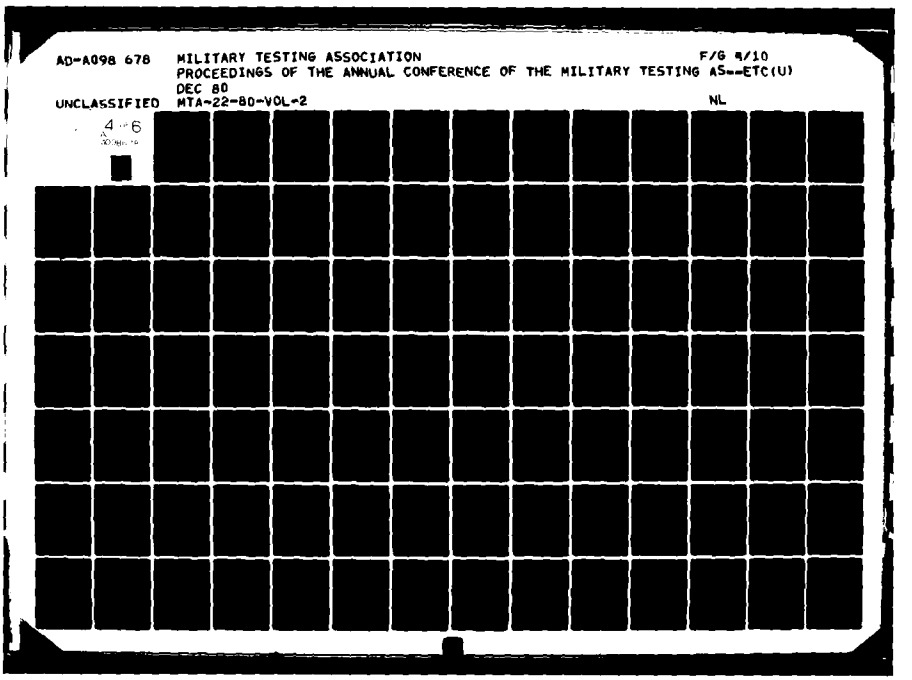
2. Main Features of the Model

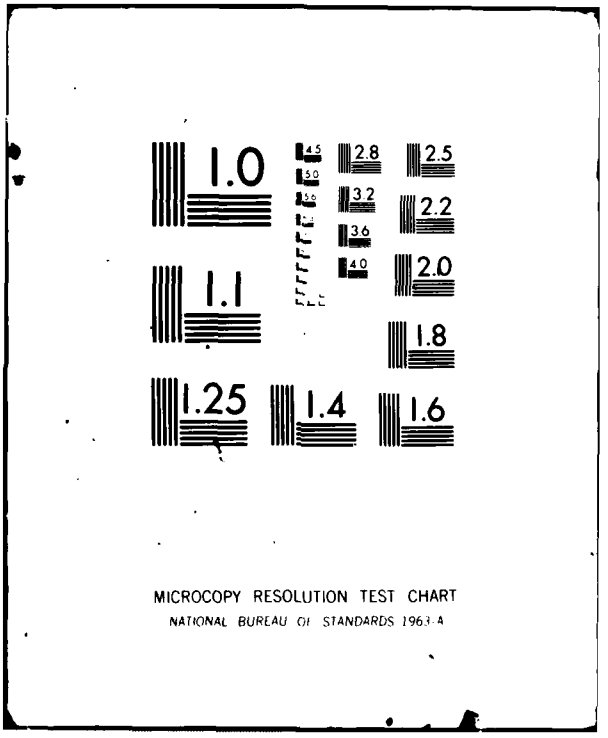
2.1 The assessment of the positions is made according to the envisaged "normal performance" of the job.

2.2 The basis is a description of the position.

2.3 The assessment has been made in three steps, namely positions in certain sectors

(1) to be investigated according to "assessment characteristics"
(2) to be compared together
(3) to be set according to an "order of rank"


AD-A098 678
MILITARY TESTING ASSOCIATION
F/G 5/10
PROCEEDINGS OF THE ANNUAL CONFERENCE OF THE MILITARY TESTING AS--ETC(U)
DEC 80
UNCLASSIFIED
MTA-22-80-VOL-2
NL
4 of 6


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

2.4 The system consists of:

(1) seven <u>assessment characteristics</u> covering:
- designation rating
- assessment grades
- percentage of importance

(2) an <u>Assessment Table</u> (foil 2) for each assessment characteristic, including:
- up to ten assessment grades
- description of the individual grades
- examples of positions
- up to 250 value grading figures.

2.5 Each individual "position value" is obtained from the total of the value grading figures corresponding thereto. In conclusion,the individual positions are classified according to the order of succession of the relevant "position value" in a "series of grades".

3. <u>The seven Assessment Characteristics and the Designation Grading thereof</u> (foil 1)

3.1 The degree of difficulty of information processing is determined by:
- the extensiveness of the information field and the penetration depth
- the degree of abstraction of the contents
- novelty towards the information to be processed
- speed of reception and reaction.

3.2 The degree of difficulty of the"official relations" is determined by
- indigence of explanation
- possibilities of conflict
- frequency and versatility of the contacts.

838

3.3 The degree of "independence" is determined by

- the extent of the latitude of action
- the frequency of decisions.

3.4 The degree of "responsibility" is determined by

- the impact of the working procedure
- the extent of the field of performance and supervision
- the type and extent of the authority of management and supervision
- degree of responsibility of the directly subordinated collaborator.

3.5 The degree of the "expenditure of energy" is determined by

- the intensity of the expenditure of energy
- the duration thereof

3.6 The degree of the "preparatory trainirg and education" is determined by the required

- qualification

3.7 Finally, the degree of"experience" by

- the extensiveness and thouroghness of the knowledge and abilities required.

4. <u>The Assessment Table under the example of "Degree of Independence"</u> (foil 2)

4.1 components of the assessment characteristics

Independence means the authority to decide between several admitted possibilities.

Independence requires the ability to take decisions and initiative.

4.2 Designation rating

The degree of independence is determined by
. the extent of the latitude of action
. the frequency of decisions

4.3 The extent of the latitude of action is to be judged on according to

. to what extent is the fulfilment of duties determined by other (i.e. with regard to time of distribution of the work, the type of procedure, the employment of personnel and means or with regard to the outcome of the work) and the possibilities of choice- with reference to the duties in their entirety - which is left for the holder of the position and
. to what extent within the general framework of the goals set up are there tasks to be taken up at own initiative and processes to set going.

Decisions are not only to be assessed when an authorization therefor subsists, but also when same are independently prepared und worked out until being ready for signature.

Also verbal decisions are to be valuated.

4.4 The description of grades for six grades is as follows

(1) The fulfilment of duties is pre-determined in detail, at the most, there is the possibility of the timely distribution thereof.

(2) The fulfilment of duties is pre-determined, enabling in part limited possibilities of choice.

(940)

(3) The fulfilment of duties is pre-determined enabling in part great possibilities of choice.

(4) a) The fulfilment of duties is only partially determined by the general objectives designated.

b) Within the framework of the latitude of action described under grade (3), particularly often own decisions have to be taken.

(5) a) The fulfilment of duties is predominantly determined by the general objectives designated alone and enables a wide-reaching freedom of disposition.

b) There is a latitude of action according to grade 4 point a). Within the framework of the fulfilment of duties in the whole, particularly often own decisions have to be taken.

(6) There is a latitude of action according to grade 5 point a). Within the framework of the fulfilment of duties in the whole, particularly often own decisions have to be taken.

4.5 The examples of positions such as:

- Clerk competent for the pay of servicemen at the WBGA (Assessment grade 1 and value grading figure 10), or:
- Departmental expert for professional promotion at an administrative section of the Armed Forces (Assessment grade 5 and value grading figure 76)

describe two appointments/positions in the Federal Armed Forces, yet same are function of administrative officials; whereas functions of servicemen in the Forces or even

partially comparable function, like for example in the Federal Frontier Police, are entirely lacking.

5. Advantages and Disadvantages of the present Model

At this point, I wish to lead over to a summarized evaluation of the Model according to the viewpoint of the Federal Armed Forces Association. I emphasize - according to the viewpoint of the Federal Armed Forces Association, which is an independent organization having a quarter of a million of members. Its taks is to attend to the ideal, social and professional interests of the servicemen vis-a-vis the Parliament, the Government and the public opinion in the Federal Republic of Germany. The Federal Armed Forces Association is not striving for neither privilages nor advantages for the servicemen in comparison with other groups in the society or in the public utilities. Yet it is taking great care that the social symmetry in the Federal Repubic does not shift much to the disadvantage of the servicemen. This has namely occurred in part during the last years through the big success achieved by the German Trade Unions especially in two sectors: First- the introduction of the 40 hours week (whereas the servicemen have to perform their duties over 60 and more hours per week) and second - enlargement of the possibilities of the personnel representation in the Public Utilities. Thus it is for example practically impossible that an official or an employee be transferred against his own will. In comparison to this, servicemen and especially the officers are being transferred every 3-5 years in the average with all the disadvantages deriving therefrom for the families. The shifting in the social symmetry resulting alone herefrom should adequately be taken into consideration in a special pay system for servicemen.

(542)

The Federal Armed Forces Association intends during this year to proceed with the thorough examination of this complex of problems through a committee which I have the honour to lead.

The foregoing has the purpose of enabling the comprehension of the summarized judgement of the Model under discussion according to the viewpoint of the Federal Armed Forces Association set out hereunder.

In any case it is worth appreciation and is thus of advan- that after all a Model for the analytical assessment which can be used as reference in all the Public Utilities has finally been elaborated by the Federal Ministry competent for Pay Matters.

However, the result of an expenditure of work for about five years is rather poor indeed to the extent of being described by the Ministry itself merely as a Model offering a "framework".

The limitation to only seven assessment characteristics is of course advantageous. This small number is easy to survey and enables to a great extent - at least from a theoretical viewpoint - a more easy and a more strict application. On the other hand, this goes cogently at the expense of exactness, which can be particularly essential for comparative assessments and validation.

In particular, an assessment element for "the influence of the environment" is lacking. Such element cannot be absolutely dispensed with at least with regard to part of the specific peculiar service characteristics of the servicemen.

Apart from the foregoing, the formation and assignment of permanent assessment committees for each specific sector in the Public Utilities is equally indispensible as the appropriate preparation or training of same. Also the sub-

sequent regular reciprocal harmonisation following the example of particularly typical "corner-positions" to be established in the different sectors of the Public Utilities, in order to enable an as fair as possible classification according to the pay in the "series of grades" foreseen for the positions. This would lead to the possibility of eliminating - at least in part - numerous existing inconsistancies especially with regard to the servicemen. In this respect for instance, the assessment factor "time", lacking in the Model, has to be duly considered. This factor which is of special importance in view of the demands and requirements the servicemen have to cope with, has not hitherto been taken into consideration either in the assessment or in the pays of the servicemen.

In conclusion, we'll have to wait and see how this Model will be used as an assessment aid and how is it going to work out. The Federal Armed Forces Association intends to start with the first attemps during this year.

SURVEY OF THE SYSTEM

Assessment characteristic	Designation grading	Number of assessment grades	Importance in percent
Degree of difficulty of information processing	Exentension of the information field and the penetration depth	10	25
	Degree of abstraction of the contents		
	Novelty towards the information to be processed		
	Speed of reception and reaction *)		
Degree of difficulty of the official relations	Indigence of explanation of the information	6	10
	Possibilities of conflict		
	Frequency and versatility of the contacts		
Degree of independence	Extent of the latitude of action	6	10
	Frequency of decisions		
Degree of responsibility	Impact of the working procedure	10	20
	Extent of the field of performance and supervision **)		
	Type and extent of the authority of management and supervision **)		
	Degree of responsibility of the directly subordinated collaborator **)		
Degree of the expenditure of energy	Intensity of the expenditure of energy	4	5
	Duration of the expenditure of energy		
Degree of preparatory training and education	Qualification	4	22
Degree of experience	Extensiveness and thoroughness of additional knowledge and abilities	4	8

*) Additional designation grading, if the actual performance of work mainly requires that signs, situation or processes communicated have to be indirectly turned into action.

**) Alternative designation grading, if the actual performance of the work mainly requires that the holder of the position has for a specific field to:

- manage the fulfilment of duties through others,
- be in charge of legal supervision, technical supervision organic supervision.

3.3. ASSESSMENT TABLE

Assessment grade	Description of grades	Examples of positions	Value grading figure
1	The fulfilment of duties is pre-determined in detal, at the most there is the possibility of the timely distribution thereof.	Clerk competent for the pay of servicemen at the WBGA Pay-roll accounting clerk at the garrison administrative headquarters Booking clerk for foreign tickets at the German Railways Engineer (TEE, IC and D trains) Signal attendant at the German Railways Gateman (high traffic density) Attendant to emptying mail boxes (with motor-car) Income tax assistant at the Revenue Board Inspector of works at a municipality Adjunct (simple assignments)	10
2	The fulfilment of duties is pre-determined enabling in part limited possibilities of choice.	Main cashier (average main cash-desk)at a post office Supervision of the dispatch of mail (big post) at a post office Regional official in a restricted area in a regional police authority Departmental expert for inspection of works (grade class 3) in a province Departmental expert for administrative budget in a municipality	22

Assessment grade		Description of grades	Examples of poisitions	value grading figure
3		The fulfilment of duties is pre-determined enabling in part great possibilities of choice.	Industrial auditor for difficult average and big firms at a Revenue Board Head of an Income Tax section at a Revenue Board Head of the Department of Inspection of Works at a Municipality	37
4	a)	The fulfilment of duties is only partially determined by the general objectives designated.	Head of a Section for industrial auditing at a Revenue Board Head of a criminal dept. in a regional police authority	55
	b)	Within the framework of the latitude of action described under grade 3, particularly often own decisions have to be taken.		
5	a)	The fulfilment of duties is predominantly determined by the general objectives designated alone and enables a wide-reaching freedom of disposition.	Departmental expert for professional promotion (with development duties) at an administrative section of the Armed Forces (with at least six fields of specialization)	76
	b)	There is a latitude of action according to grade 4 point a). Within the framework of the fulfilment of duties in the whole, particularly often own decisions have to be taken.	Director of a Revenue Board (whith at least six fields of specialization) Director of a Mining Office Head of a criminal section at a regional police authority Head of a public health office (average district of jurisdiction)	
6		There is a latitude of action according to grade 5 point a). Within the framework of the fulfilment of the duties in the whole, particularly often, own decisions have to be taken.		100

Should the actual performance of the work mainly require that duties have to be taken up at own initiative and processes have to be set going, the following higher value grading figure is to be given; the value grading figure of grade 6 cannot however be exceeded.

SHIELDS, LCdr. William S., Canadian Forces Personnel Applied Research Unit, Toronto, Ontario.

CROSS-VALIDATION OF A FOUR-PARAMETER ITEM CHARACTERISTIC CURVE TEST SCORING MODEL (Wed A.M.)

This paper reports experience with a new four-parameter Item Characteristic Curve model in scoring multiple-choice tests of English grammar. Two parameters are the total-score means of candidates who answered an item correctly and of those who answered incorrectly. A third parameter is the score variances of these groups (assumed to be equal) and the fourth is the ratio between the group sizes. Alternatively, using a partitioning of total variance to estimate group variance, the third and fourth parameters can be defined simply as the relative sizes of the two groups. Parameter estimation is very uncomplicated, and no instances of nonconvergence of this model have been encountered. Scoring of new candidates normally requires a maximum of six iterations using the number-right score as a starting value. Both validation and cross-validation demonstrated the superiority of the model over the number-right scoring method.

CROSS-VALIDATION OF A FOUR-PARAMETER ITEM CHARACTERISTIC CURVE TEST SCORING MODEL

Lieutenant-Commander W.S. Shields[1]

Canadian Forces Personnel Applied Research Unit
4900 Yonge St., Willowdale, Ontario
M2N 6B7

1. INTRODUCTION

An appealing alternative to the "number-right" test scoring method is a class of procedures known variously as "latent trait theory", "item response theory", or "item characteristic curve (ICC) test scoring". Although the theory has been developing in the literature for nearly forty years, relatively few practical applications have been reported. This paper reports an application using two multiple-choice tests of English grammar and a new four-parameter ICC model developed by the author. Both preliminary and cross-validation studies provide evidence of superiority of the model over the number-right method.

2. BACKGROUND

The introduction of ICC theory is usually attributed to Lawley (1943), although ingredients of the method are apparent in the work of both Guilford (1936) and Richardson (1936). Of proponents of the theory, Frederic M. Lord was perhaps most instrumental in keeping the subject alive until developments in computers could make them practical. Lord wrote in one of his 1953 papers, "... in view of the heavy (but not insuperable) computational difficulties in the way of any practical application, the present discussion is directed chiefly towards determining what conclusions of general theoretical significance can be drawn" (1953 a, p 57). He went on to develop what has been called the "normal ogive" model.

Very few practical applications appeared in the literature until Lord's work with the Verbal Scholastic Aptitude Test (1968). In the same year, Lord and Novick published "Statistical Theories of Mental Test Scores" which included a unified development of the normal ogive model (Chap 16) and a contribution by Allan Birnbaum describing his "logistic" model (Chap 17). This book, together with work by Fumiko Samejima which began to appear in the English language literature at about the same time, sparked a renewal of interest in ICC procedures.

In 1973, Samejima pointed out certain difficulties with Birnbaum's logistic model. In 1975, Owen described a Bayesian procedure. In the same year, Lord reported a further problem with the logistic model and prescribed a remedy. Samejima published two articles in 1977, one of which prescribed maximum-likelihood procedures for estimating ICC parameters (1977a). In the same year, a special edition of the journal of Educational Measurement was devoted to ICC theory, and contained several excellent

---

[1]The views and opinions expressed in this paper are those of the author and not necessarily those of the Department of National Defence.

(851)

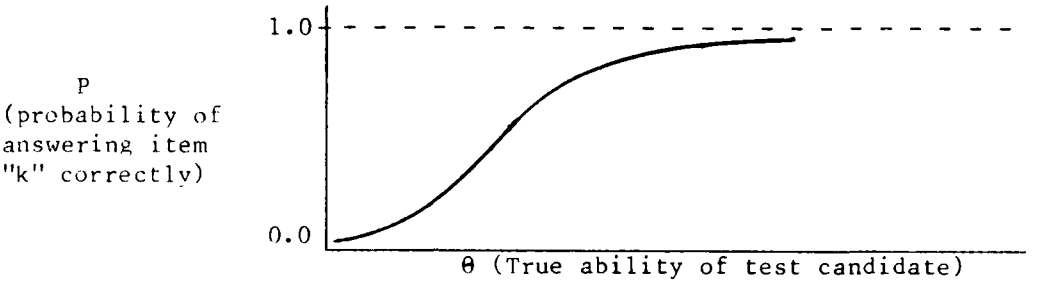
articles. One of these (Hambleton and Cook, 1977) gives a lucid description of ICC procedures which is sufficiently non-mathematical to be understood by any test practitioner. Warm (1978) has produced an excellent easy-to-read summary which is technically complete from an applications standpoint.

3. THE NATURE OF ICC THEORY

Although general ICC theory deals with a number of traits which may be measured simultaneously by a test, this discussion will be restricted to a test which is intended to measure a single trait which will be called "ability" θ. It is hypothesized that a "true" ability scale exists and that a given candidate has a unique "true" position on that scale. ICC theory then evolves from the graph shown in Figure 1.

Figure 1

The Item Characteristic Curve



ICC models differ primarily in their assumptions about the shape of the curve of Figure 1 and in the number of parameters used to express it. Because the positions of test candidates on the scale θ are outputs from rather than inputs to each model, the assumptions about the shape of the ICC and the methods used to estimate its parameters for each item are the essential determinants of the nature of the scale θ.

Probably the simplest ICC model is the Rasch model (1960). Its only parameter is item difficulty d. It assumes that the relationship between θ and p is given by

$$p = \frac{e^{\theta-d}}{1+e^{\theta-d}} \qquad (3.1)$$

It can be seen from 3.1 that if difficulty d is very high in relation to the candidate's ability θ (note that θ and d are both measured on the same scale), $e^{\theta-d}$ will be a very small number, as will probability p that the candidate will answer correctly. If, on the other hand, θ is very high or d is very low, p will be large, approaching 1.0 as an upper bound. The main objection of critics of the Rasch model is that it totally ignores the discriminating power of a test item, while the main claim of its proponents is that it is the only model that is totally compatible with number-right scoring.

(852)

Lord and Novick (1968) have provided a detailed development of their two-parameter "normal ogive" model. The model assumes that the shape of the curve of Figure 1 is that of the cumulative normal probability density function. They have given a nonrigorous but mildly appealing justification for this assumption (op cit, pp 370-71). The two parameters are difficulty, as in the Rasch model, which determines the lateral position of the curve, and discrimination, which is proportional to the maximum slope attained by the curve. Unfortunately, no algebraic equation is available to express the cumulative normal curve.

Birnbaum's logistic model is based on the (cumulative) logistic curve whose equation is strikingly similar to the Rasch formula:

$$p = \frac{e^{1.7\theta}}{1 + e^{1.7\theta}} = \frac{1}{1 + e^{-1.7\theta}}$$

Its primary justification is that it differs from the cumulative normal curve by less than .01 for all θ. It is, however, much easier to calculate than the cumulative normal curve. It also easily accommodates Birnbaum's three parameters:

a	item discrimination
b	item difficulty, and
c	guessing parameter.

The formula for p becomes

$$p = c + \frac{1 - c}{1+e^{-1.7a(\theta-b)}} \qquad (3.2)$$

As an example of the use of c, if a multiple-choice test has n choices per item, c would very nearly equal 1/n, because if θ were infinitely low the second term would be negligibly small, and chance alone would determine the probability of responding correctly.

Customarily, ICC models make an assumption of "local independence", which means that item responses are intercorrelated only through trait θ. The local independence assumption permits calculation of the joint probability of a test candidate's entire response pattern given that his true ability equals θ. It is simply the product over all items of the probabilities of the responses, both correct and incorrect, that the candidate made. This joint probability is proportional to the likelihood of any hypothesis about the value of θ for the candidate. Customarily, the value $\hat{\theta}$ of θ is chosen which makes the joint probability, and thus the likelihood of the hypothesis that $\theta = \hat{\theta}$, a maximum.

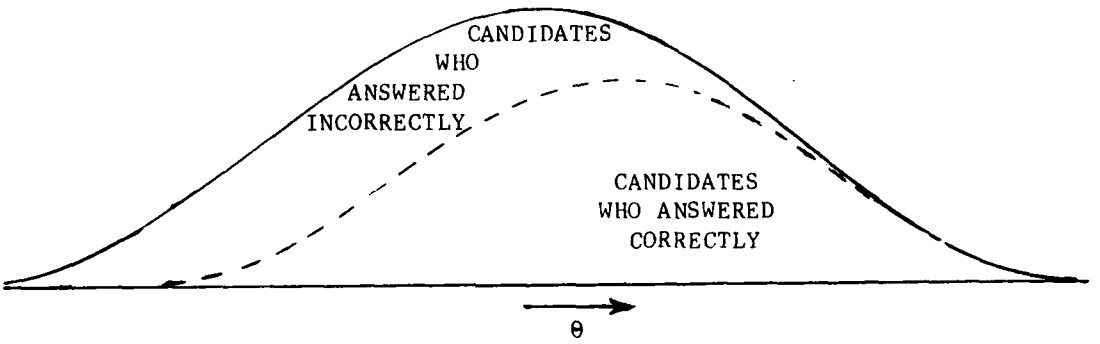
One of the very crucial steps in the ICC procedure is the estimation of the ICC parameters. Usually an iterative process is used with a sample of data, so that starting values of $\hat{\theta}$ are fed into a model which finds parameter values which fit these values of $\hat{\theta}$. Then new values of $\hat{\theta}$ are

calculated from the IC curves, from which a new set of values of parameters is obtained, etc. If convergence results, a solution for all values of $\hat{\theta}$ is obtained, as is a set of values of the ICC parameters which can then be used in scoring a new set of candidate responses.

4. A FOUR-PARAMETER MODEL

Every population has some distribution of ability θ as shown in figure 2. For any given test item k, the population can be divided into two groups by the dotted line shown, those who answered k correctly and those who answered incorrectly. The issue as to what shape k's item characteristic curve should have is equivalent to that of what path the dotted line should take through the population.

Figure 2



Suppose, for the moment, that the distribution A_k of θ for those who answered correctly is normal, i.e., that

$$A_k = \frac{n_{Ak}}{\sigma_{Ak}\sqrt{2\pi}} e^{-1/2\left(\frac{\theta-\mu_{Ak}}{\sigma_{Ak}}\right)^2}$$

where n_{Ak} is the number who answered correctly, and μ_{Ak} and σ_{Ak} are the usual distribution parameters. Suppose also, that the distribution B_k of those who answered incorrrectly is normal, i.e., that

$$B_k = \frac{n_{Bk}}{\sigma_{Bk}\sqrt{2\pi}} e^{-1/2\left(\frac{\theta-\mu_{Bk}}{\sigma_{Bk}}\right)^2}$$

Unless k has an extraordinarily high discriminating power, the distribution A_k+B_k of the whole population will be very nearly normal also, because the sum of two normal distributions of similar variance but slightly different means differs only slightly from normality. This can be demonstrated by the fact that if a row in Pascal's triangle of binomial coefficients is displaced one position to the right or left and added to its original self, the next row of the triangle is obtained. Of course, the rows of the triangle are asymptotically normal.

The model described in this paper does not assume that any distribitions are normal. However, it does assume that the dotted line in figure 2 has such a path that, if A'_k is the distribution of those answering correctly, and B'_k is the distribution of those answering incorrectly, then

$$A'_k/(A'_k+B'_k)=A_k/(A_k+B_k)$$

Implicit in this assumption is that $\mu'_{Ak} = \mu_{Ak}$ and $\mu'_{Bk} = \mu_{Bk}$.

Let V_k be a response vector of a particular person, and let V_k operate in such a way as to substitute B for A and A for B, subscripts included, whenever this person answered item k incorrectly. If this person answered a particular item k correctly, the likelihood of a hypothesis that his ability is θ is $A_k/(A_k+ B_k)$.

If he answered it incorrectly it is $B_k/(A_k+B_k)$. Assuming local independence, the joint likelihood of this hypothesis for all items is proportional to

$$\prod^{k} V_k \cdot \frac{A_k}{A_k + B_k} \qquad (4.1)$$

Because the logarithm of the above likelihood is a monotonic function of the likelihood, it will reach a maximum at the same value of θ for which the likelihood itself is maximum. The log-likelihood equals

$$\sum^{k} V_k \cdot \ln \frac{A_k}{A_k + B_k}$$

Its maximum will occur when its first derivative with respect to θ equals zero. Performing the differentiation, and equating the result to zero, results in the equation:

$$\theta = \frac{\sum^{k} V_k \cdot \left[\frac{B_k}{A_k + B_k}\left(\frac{\mu_{Ak}}{\sigma_{Ak}^2} - \frac{\mu_{Bk}}{\sigma_{Bk}^2}\right)\right]}{\sum^{k} V_k \cdot \left[\frac{B_k}{A_k + B_k}\left(\frac{1}{\sigma_{Ak}^2} - \frac{1}{\sigma_{Bk}^2}\right)\right]} \qquad (4.2)$$

The condition that this be a maximum is that the second derivative be negative, which can be shown to happen when

$$\sum^{k} V_k \cdot \left[\frac{B_k}{A_k + B_k}\left(\frac{1}{\sigma_{Bk}^2} - \frac{1}{\sigma_{Ak}^2}\right)\right] < 0$$

F55

Obviously, response vectors are possible for which this will not be true. Therefore, any procedure to estimate θ for an individual will have to confirm the negativity of the second derivative.

The probability that the second derivative will be negative can be increased by converting the maximum likelihood solution into a Bayesian solution. The Bayesian solution merely replaces the likelihood function by a posterior probability function, which equals the product of the prior probability and the likelihood. The prior distribution can be taken, for example, as a normal distribution having the same mean and variance as a set of scores using the number-right scoring method. Use of such a prior makes item characteristic scores easy to compare with number-right scores.

Use of a normal prior involves multiplying 4.1 by

$$\frac{n_{Ak} + n_{Bk}}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{\theta - \mu}{\sigma}\right)^2}$$

where μ and σ can be, if desired, derived from number-right scores. The net effect on 4.2 is to add the quantity μ/σ^2 to the numerator and $1/\sigma^2$ to the denominator. The effect on the second derivative is to subtract the quantity $1/\sigma^2$, thus improving the chances that the second derivative is negative.

Actually, this Bayesian model can be designed in such a way as to guarantee negativity of the second derivative. This can be done by assuming that, for each item k, $\sigma_{Ak} = \sigma_{Bk}$. This is not an unreasonable assumption. In fact, experience indicates that the usual estimates of σ_{Ak} and σ_{Bk} are very unstable, especially for very difficult or very easy questions, and that their use can prevent convergence of an iterative procedure used to calculate $\hat{\theta}$ for all candidates. The assumption of homoscedasticity is by no means new; it is used almost universally in regression analysis.

Letting $\sigma_{Ak} = \sigma_{Bk}$ reduces the model to a logistic form having three constraints imposed upon it. Referring to formula 3.2, the constraints are $a = (\mu_{Ak} - \mu_{Bk})/(1.7\ \sigma_{Ak}^2)$, $b = (\mu_{Ak} + \mu_{Bk})/2 + [\sigma_{Ak}^2/(\mu_{Ak} - \mu_{Bk})]\ln(n_{Bk}/n_{Ak})$, and $c = 0$. These constraints undoubtedly reduce by several orders of magnitude the volume of the space which trial solutions can occupy during convergence, and thus should, theoretically, tend to reduce the length of the convergence path.

Letting $\sigma_{Ak} = \sigma_{Bk}$ reduces 4.2 from six parameters to five. It will later be shown that σ_{Ak} can be written as a function of n_{Ak}, n_{Bk}, μ_{Ak} and μ_{Bk}; thus the model, in the form in which it was applied by the author, has only these four parameters. Letting the standard deviations be equal can only be done with a Bayesian model, because the denominator of 4.2 vanishes, its only remnant being whatever constant was added to it through imposition of the prior distribution. Adjusting the size of the constants added to the numerator and denominator of 4.2 can provide an easy method of controlling the mean and variance of the ICC scores produced. It has been found experimentally that the size of the constants chosen has virtually no effect on relative scores.

It is appropriate that 4.2 should be indeterminant without the imposition of a prior distribution, because the moments of the distributions are all functions of θ, so 4.2 as it stands contains no provision for estimating the sign or order of magnitude of the test scores. If the constant added to the denominator is made equal to 1/c, where c is a constant used to regulate the variance of $\hat{\theta}$, the constant added to the numerator will be μ/c, where μ is the mean of the prior distribution (which will determine almost exactly the mean of the scores produced by the model). Equation 4.2 then becomes:

$$\hat{\theta} = \mu + c\sum^{k} V_k \cdot \left[\frac{B_k}{A_k + B_k} \left(\frac{\mu_{Ak} - \mu_{Bk}}{\sigma_{Ak}^2} \right) \right] \qquad (4.3)$$

Thus, if no items have yet been answered by a candidate, the formula would simply estimate his ability $\hat{\theta}$ as μ, the mean of the prior distribution. It has been found that if the constants μ and c are chosen so that 4.3 gives a score of zero to a candidate who got all the items wrong, and a perfect score to one who got all the items right, the mean and variance of $\hat{\theta}$ closely approximate those obtained using number-right scoring.

One must avoid at this point the temptation to argue that if a candidate has the quantity

$$\frac{B_k}{A_k + B_k} \left(\frac{\mu_{Ak} - \mu_{Bk}}{\sigma_{Ak}^2} \right)$$

added if he answered item k correctly, and the quantity

$$\frac{A_k}{A_k + B_k} \left(\frac{\mu_{Ak} - \mu_{Bk}}{\sigma_{Ak}^2} \right)$$

subtracted if he answered it incorrectly, the difference for the candidate between these two is simply

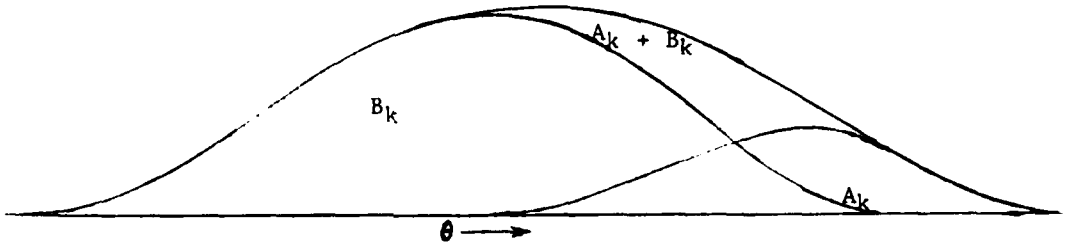
$$\frac{\mu_{Ak} - \mu_{Bk}}{\sigma_{Ak}^2}$$

which should be added for each item answered correctly and no marks given for those answered incorrectly. Unfortunately, such differential analysis is valid only in a system in which a given item is "worth" the same amount to each candidate. This is not the case, as inspection of 4.3 clearly reveals.

Figure 2 is a typical response distribution of an "easy" question. Looking at 4.3, candidates with high ability receive a very small contribution to their score by answering this question correctly. Those of them who answer it incorrectly, however, receive a very large penalty, because the operator V_k will substitute A_k for B_k in the numerator of 4.3. Candidates of low ability will receive a modest score contribution if they answer this question correctly, but only a small penalty for answering it incorrectly.

Figure 3 shows the distribution of θ for a "hard" question. Note first from 4.3 that practically all candidates receive a higher score contribution

Figure 3



for hard questions than for easy ones, but particularly candidates of low ability. All candidates suffer a low penalty for getting a hard question wrong, but again this is particularly true of candidates of low ability.

Actually, if all candidates answer all questions, the item weight $(\mu_{Ak} - \mu_{Bk})/\sigma^2_{Ak}$ applied to a linear aggregate of all questions correctly answered does provide a "sufficient statistic", i.e., will rank order candidates identically to the use of 4.3. This is because to the extent that the questions a candidate answered correctly tended to be hard ones (therefore bearing a larger bonus), those that he answered incorrectly would tend to be easy ones (therefore bearing a larger penalty). However, 4.3 is more appropriate if there were any unreached questions and probably also when any questions were omitted. It also has obvious advantages in tailored testing.

5. TEST OF THE MODEL

The model of 4.3 was compared with the number-right scoring method using data from an ongoing testing situation. Two parallel 50-item four-choice multiple-choice tests of English grammar and style are routinely administered to officers of Captain rank in the Canadian Armed Forces while they attend a ten-week professional development course at The Canadian Forces Staff School in Toronto. The first test is administered to all English language students on the second day of the course. Those whose number-right score is less than 66% (normally about 60% of test candidates) are required to take the second test five weeks later. Neither test is time limited.

During the time between tests two voluntary remedial programmes are offered. One consists of individual study using English 3200 (Blumenthal, 1972). The other consists of a half-hour per day of classroom instruction outside normal programme hours. In addition, students engage in a number of writing and speaking exercises, which are critiqued for grammar and style (among other criteria).

Thus, there are several reasons why a student's score is expected to be higher in the second test. An added reason, unrelated to any increase in proficiency, is the excusing from the second test of about the top 40% of candidates. Especially if the tests were unreliable, scores would tend to increase under such a system. Practice effects of the first test on the second would also tend to increase scores.

However, no doubt partly because the remedial programmes are voluntary, about 20% of test candidates have number-right scores that go down on the second test. Of course, this phenomenon is probably not due entirely to test error. The candidate may be fatigued or even ill while he is taking the second test. Although a separate test administration is available to candidates who report illness, it is likely that some such problems are unreported.

Nevertheless, it was assumed that test error was an important cause of any score decline, and that therefore a more accurate scoring method would reduce, if not the incidence, at least the extent of such decline. Therefore the total number of points of score decline was chosen as the criterion by which to compare competing scoring methods. It is recognized that this criterion addresses directly only the reliability component of test validity, but it has the advantages of simplicity and objectivity. Because neither the formal remedial instruction nor the English 3200 programme is aimed directly at the content of the second test, gains from the first test to the second constitute good evidence of improvement in the correct use of English language.

A primary requirement in the comparision was to balance the two tests for difficulty. The tests were parallel for subject matter, e.g. item 3 on both tests dealt with subject/verb agreement, item 4 with word choice, etc. Using a percent-wrong criterion of item difficulty, five of the first ten items of the first test were exchanged with the five corresponding items of the second test in such a way that the average difficulty of the five questions moved to the second test differed as little as possible from the average difficulty of the five that remained behind, and similarly for the five questions moved from the second test to the first. Of the 252 possible ways of picking five items out of ten, the unique optimal exchange was found and made.

This process was repeated for items 11-20, 21-30, etc. until exactly half of the items of the first test had been exchanged with their counterparts in the second test. A number of final passes were now made to look for an exchange involving two items from each test that would improve total test balance. On the third such pass, no lucrative exchanges were found, and overall test difficulty differed by less than 0.1 marks (out of 50).

The elaborate balancing procedure described was necessary because the item difficulty measures in the first test could not be compared with similar measures in the second, because both instruction and truncation of the sample had occurred in between.

The procedure also produced some balancing for item discrimination, because at an earlier stage in test design items having the highest discrimination indexes (point-biserial correlations with total test scores) had all been placed in the first test while revision of some of the choices and stems of items in the second test was carried out. After this revision, but before the balancing procedure, all items of both tests met the following criteria, when item-analysed using samples of 324 (first test) and 217 (second test):

a. all keyed choices and distractors functioning,
b. all keyed choices discriminating positively, and
c. all distractors discriminating negatively.

Thus, although no conscious attempt was made to balance for discriminating power, half of the most highly discriminating items ended up in each test.

Four courses embracing an instructional period of one year, including 340 test candidates for the first test and 204 for the second, were chosen as the sample for which to compare ICC with number-right scoring.

ICC scores were not computed using any "curves", but rather by substitution of parameter values into formula 4.3. Using number-right scores as starting score values, parameters were estimated, then a new set of candidate scores, then new parameter values, etc. Although θ appears on both sides of 4.3, because both A_k and B_k are functions of θ, it was found that successive iterations, as described above, resulted in convergence. Although traditional estimates were used for μ, μ_{Ak} and μ_{Bk}, and observed numbers for n_{Ak} and n_{Bk}, σ_{Ak} was estimated through an analysis of variance in which the total sum of squares was segregated between "sum of squares among" and "sum of squares between". This produced the formula:

$$\sigma_{Ak}^2 = \sigma_{Bk}^2 = \sigma^2 - \frac{n_{Ak}}{N}(\mu - \mu_{Ak})^2 - \frac{n_{Bk}}{N}(\mu - \mu_{Bk})^2$$

where $N = n_{Ak} + n_{Bk}$. More direct methods of estimating σ_{Ak} proved to be unstable to the point where they regularly prevented convergence.

It was found through experimentation that convergence was accelerated by putting some "damping" into the system. The damping was introduced by applying only a percentage of the indicated change to the values of $\hat{\theta}$ at each iteration. A figure of 30% was found to work quite well.

Convergence was deemed to have occurred if no candidate's score had an indicated change of more than one tenth of a mark (out of 50) between two consecutive iterations. Convergence of the first grammar test required six iterations; that of the second test required nine. Central Processing Unit time (on an IBM 370 using FORTRAN programming) was 13.6 seconds for the first test and 11.7 seconds for the second.

A summary of the comparison between number-right scoring and the ICC model is shown in Table 1. Because no students had the same ICC scores for both tests, no column was provided for that category. It is apparent from the table that the ICC scoring produced no noteworthy reduction in the incidence of score decline, but a palpable reduction in its extent. A two-way analysis of variance revealed that the reduction in extent is significant at $p < .10$ but not at $p < .05$.

It is possible to make a crude estimate of test standard error of measurement by assuming that the 41 candidates who failed to show improvement under either scoring system actually had no appreciable change in ability between tests. Under this assumption, the mean square score decline of these 41 candidates would equal the sum of the error variances of the two tests. If the tests are then assumed to share this error variance equally, test standard error can be estimated. This procedure produced a standard error of 3.360 marks for the number-right method compared with 2.590 marks for ICC scoring. Thus, ICC scoring resulted in a 40% reduction in error variance, a very worthwhile improvement.

Table 1

Staff School Course No.	Number-Right Scoring			ICC Scoring	
	No of Students Whose Marks Declined	No Whose Marks Were the same	Total Decline (Marks out of 50)	No of Students Whose Marks Declined	Total Decline (Marks out (of 50)
34	12	5	58	13	33.6
35	9	5	43	13	35.6
36	8	2	28	10	24.9
37	12	4	33	15	33.1
TOTAL:	41	16	162	51	127.2

6. CROSS-VALIDATION

Parameter values for all 100 items were saved on a computer disk file. This enabled scoring of subsequent test candidates with a minimum of computation. Some iteration, one candidate at a time, was necessary because of θ's occurrence on both sides of 4.3. The damping factor of 30% was retained, and found to be necessary.

Cross-validation was performed using the next three Staff School student intakes (Staff School Courses 38, 39 and 40). The number of iterations required to obtain the ICC score for each student in course 38 is shown in Table 2. Convergence was achieved for all candidates in all intakes and in no instance required more than six iterations. Thus, it would appear that formula 4.3 has reliable convergence properties when used with an appropriate damping factor.

Table 2

Iterations Before Convergence	Number of Students	
	First Test	Second Test
1	0	2
2	5	28
3	20	16
4	39	9
5	15	1
6	4	0

The results of the cross-validation are shown in Table 3. The total decline in marks was smaller using ICC scoring for two of the three samples and the same for the remaining sample. These results are in agreement with those obtained in the validation phase. The 22 students who failed to show any improvement under either scoring method experienced a mean square decline of 13.23 under number-right scoring compared with a mean square decline of 9.25 under ICC scoring. This gives standard error measures of 2.57 marks for number-right scoring compared with 2.15 marks for ICC scoring, a 30.1% reduction in error variance, in good agreement with similar figures obtained using the primary validation sample.

Although the superiority of ICC scoring, using extent of score decline as the criterion, fell short of significance in both the validation and cross-validation samples, the difference between the ICC and number-right methods for the combined samples produced a t-value of 2.975, with six degrees of freedom, which is significant at $p < .05$.

Table 3

CROSS-VALIDATION

Staff School Course No.	Number-Right Scoring			ICC Scoring	
	No of Students Whose Marks Declined	No Whose Marks Were the same	Total Decline (Marks out of 50)	No of Students Whose Marks Declined	Total Decline (Marks out of 50)
38	8	3	30	13	30.0
39	11	2	37	7	24.6
40	6	4	19	8	16.9
TOTAL	25	9	86	28	71.5

7. DISCUSSION AND CONCLUSIONS

ICC scoring, using the model developed in this paper, appears to result in a small but very worthwhile improvement in the precision of measuring ability using multiple-choice tests. Although the word "ability" was used in the development of the model, nothing in the logic or its formulation assumed that θ was any particular kind of trait. The concept of item "difficulty" has no fundamental difference from similar "p-value" measures in other kinds of psychological tests, and merely relates to differences in **proportions** of people who answered a particular item in different ways. Thus ICC theory is perfectly extendable to other psychological instruments (personality tests, structured interviews, rating scales etc) in which a reasonably limited range of responses is possible.

The ICC scoring procedure presumes that item validation has already taken place. However, it can easily be designed to automatically eliminate items that lack statistical validity, i.e. those for which $\mu_{Ak} < \mu_{Bk}$. This was, in fact, done with the grammar tests used here, with the result that one item was eliminated from the first test and three from the second. The only alternative to eliminating these items would be to decrease a candidate's score for any of them answered correctly and increase it for any answered incorrectly.

The ICC model of formula 4.3 is open to at least two criticisms, namely the assumption that $\sigma_{Ak} = \sigma_{Bk}$, and the absence of a guessing parameter. It can be argued that σ_{Bk} ought to be larger than σ_{Ak}, because the people who answered a question correctly made only one response, whereas those who answered incorrectly made a variety of responses. Alternatively, one could argue that σ_{Ak} ought to be larger than σ_{Bk}, because those who answered correctly constitute a mixture of "those who knew" and "those who guessed".

The author postulates that very little pure guessing goes on during tests. Probably there is considerable narrowing of the possibilities while a candidate rejects choices that be believes to be wrong. He receives, as he should, some credit for this narrowing because his probability of a correct response is increased. However, his final choice is probably very rarely a matter of pure chance. As evidence in support of this postulate, the first grammar test contained three items of such high difficulty that considerably fewer than 25% of candidates got them right. One would expect 85 of the 340 candidates in the validation sample to get these questions right by pure guess-work. However, only 66, 69 and 46 people answered them correctly. Nevertheless, all three items had good discriminating power and made an important contribution to the test. To place a 0.25 floor on the probability that any four-choice multiple-choice question will be answered correctly, unless examinees are totally ignorant, is a practice not well supported by available evidence.

A conclusion often reached in the face of results like those reported here is that the improvement in scoring efficiency of ICC procedures is small, and therefore so also will be the benefits reaped from adopting them. That this conclusion is fallacious has been very ably demonstrated by Lord (1968, p 1007). He points out that if an improved scoring method is successful in increasing the predictive validity of a test from 0.50 to 0.51, one could randomly discard 30% of the test items in the process of converting to the new scoring method, with no loss in predictive validity. If one discards, instead, the items of lowest individual predictive validity, test length could in most cases be reduced by at least half. Considering the cost of the test candidate's time, such a gain would be very substantial.

REFERENCES

Bejar, Issac [illegible]nd Weiss, David J. Computer Programs for Scoring Test Dat[illegible] Item Characteristic Curve Models. Minneapolis: Research Rep[illegible]-1, Department of Psychology, University of Minnesota, February, 1970.

Blumenthal, Joseph C. English 3200. New York: Harcourt Brace Javanovich, 1972.

Guilford, J.P. Psychometric Methods. New York: McGraw-Hill, 1936.

Hambleton, Ronald K., and Cook, Linda L. "Latent Trait Models and their Use in the Analysis of Educational Test Data", Journal of Educational Measurement, Vol 14, No 2 (Summer, 1977), 75.

Lawley, D.N. "On Problems Connected with Item Selection and Test Construction. Proceedings of the Royal Society of Edinburgh, Vol 61 (1943), 273-87.

Lord, Frederic M. "An Application of Confidence Intervals and of Maximum Likelihood to the Estimation of an Examinee's Ability". Psychometrika Vol 18, No 1 (March, 1953), 57-76. (a)

_____ .The Relation of Test Score to the Trait Underlying the Test". Educational and Psychological Measurement, Vol 13 (1953), 517-548. (b)

_____ ."An Analysis of the Verbal Scholastic Aptitude Test Using Birnbaum's Three-Parameter Logistic Model" Educational and Psychological Measurement, Vol 28 (1968), 989-1020.

_____."The 'Ability' Scale in Item Characteristic Curve Theory". Psychometrika, Vol 40, No 2 (June, 1975), 205.

_____."Practical Applications of Item Characteristic Curve Theory". Journal of Educational Measurement, Vol 14, No 2 (Summer, 1977), 117.

_____ . and Novick, Melvin R., Statistical Theories of Mental Test Scores. Don Mills, Ontario: Addison-Wesley, 1968.

Marco, Gary L. "Item Characteristic Curve Solutions to Three Intractable Testing Problems". Journal of Educational Measurement, Vol 14, No 2 (Summer, 1977), 139.

Owen, Roger J., "A Bayesian Sequential Procedure for Quantal Response in the context of Adaptive Mental Testing". Journal of the American Statistical Association Vol 70, No 350 (June, 1975), 251-6.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Nielson and Lydiche (for Danmarks Paedagogiske Institute), 1960.

Richardson, M.W. "Relation Between the Difficulty and the Differential Validity of a Test". Psychometrika Vol 1, No 2 (1936), 33-49.

Samejima, Fumiko. "Estimation of Latent Ability Using a Response Pattern of Graded Scores." Psychometrika, Monograph Supplement No 17. 1969.

_____. "A Comment on Birnbaum's Three-Parameter Logistic Model in the Latent Trait Theory". Psychometrika, Vol 38, No 2 (June, 1973), 221-33.

_____. "A Method of Estimating Item Characteristic Functions Using the Maximum Likelihood Estimate of Ability". Psychometrika, Vol 42, No 2 (June, 1977), 163-91. (a)

_____. "Weakly Parallel Tests in Latent Trait Theory With Some Criticisms of Classical Test Theory." Psychometrika, Vol 42, No 2 (June, 1977) 193. (b)

Swaminathan, H. and Gifford, J.A. Bayesian Estimation in the One-Parameter Latent Trait Model. Research Report 80-1, Laboratory of Psychometric and Evaluative Research, University of Massachusetts, March, 1980.

Warm, T.A. A Primer of Item Response Theory Technical Report 94108, U.S. Coast Guard Institute, Oklahoma City, Oklahoma, October, 1978.

_____, Edge, G.J. and Pastene, C.R. "The First Military Operational Application of Item Response Theory", Proceedings of the Military Testing Association, Toronto, October, 1980.

Wright, Benjamin D. "Solving Measurement Problems With the Rasch Model" Journal of Educational Measurement, Vol 14, No 2 (Summer, 1977), 97-115.

SELF-FULFILLING PROPHECIES
A MODEL FOR THE PERFORMANCE APPRAISAL OF WOMEN IN THE CANADIAN FORCES

Captain S.P. Simpson

Abstract submitted for the 22nd Annual
Conference of the Military Testing
Association - 27 to 31 October 1980

This paper describes a study scheduled for completion by Spring 1981 which will analyze the relationship between performance appraisals men and women receive in the Canadian Forces and the supervisors' attitudes towards women. Consistent with other research it is predicted that supervisors who hold more traditional attitudes towards women will rate women lower, and will have lower expectations for the performance of women than supervisors with more egalitarian attitudes. In addition, a model of self-fulfilling prophecy has been hypothesized. According to this model, it is predicted that supervisors who have lower expectations for women will act in a manner such as to reinforce the behaviour of their female subordinates and set up their work environment so that the women's performance at work will be in keeping with their supervisors' expectations. The women's expectations regarding their own performance ability will eventually come to meet those of their supervisors. The measures which will be gathered to test this model are discussed.

SELF-FULFILLING PROPHECIES A MODEL FOR THE PERFORMANCE APPRAISAL OF WOMEN IN THE CANADIAN FORCES

Captain S.P. Simpson

Submitted for the 22nd Annual
Conference of the Military Testing
Association - 27 to 31 October 1980

Research examining evaluation procedures, and other personnel decisions affecting careers of men and women in the workplace has consistently demonstrated bias against women, particularly in professions and occupations not traditionally held by women. Women are directed towards lower paying jobs, requiring less education, and are selected less frequently than men for management positions (Donahue and Costar, 1977; Fidell, 1970; Heneman, 1977; Rosen and Jerdee, 1974b; Shaw, 1972). Once on the job women are given fewer professional development opportunities, are promoted less frequently and may be evaluated on different dimensions of an attribute (e.g. leadership) than men; (Bartol and Butterfield, 1976; Rosen and Jerdee, 1973; Rosen and Jerdee, 1974c; Terberg and Ilgen, 1975). On occasion bias may work in favour of women in that employers are probably more lenient with women as far as family demands are concerned (Rosen and Jerdee, 1974a). Sex role stereotyping is identified by most as the mechanism which operates against women achieving at work, especially in those jobs or professions not conventionally held by women (Terberg, 1977).

Almost all of the studies have simulated evaluation and selection procedures used in the workplace. It could be argued that under these circumstances conditions are particularly conducive for finding differences in the evaluations of men and women. In simulations there are likely to be gaps or ambiguities in the information given to the evaluator about those being evaluated. Subjects are likely to "fill in" or interpret information given to them to form a good "gestalt" consistent with commonly held beliefs about the attributes and appropriate roles for men and women. Hence, one of the aims of this study is to determine what happens in actual evaluations of men and women at work, under conditions where the evaluations have significant impact on the workers career, and where the evaluators have a great deal of knowledge about the performance of their subordinates on the job. In addition, we wish to clarify some of the mechanisms involved in superior and subordinate relationships which would create lower evaluations, in this case the lower evaluation of women.

969

Jones (1977) argues in his book called Self-Fulfilling Prophecies that stereotypes are regions within one's implicit theory of personality. The act of assigning a label to an individual, or assigning the individual to a distinct social group, results in certain expectancies for the individual consistent with the label; in this case, for the women to behave in a way consistent with the stereotype for their sex. By virtue of having these expectancies individuals in high status positions can set up the environment, attend to behaviour and establish reinforcement contingencies so that it is difficult for the stereotyped individual to act in a manner inconsistent with expectancies. This is very similar to a concept called role entrapment advanced by Kanter in 1976. According to Kanter, assumptions and mistaken attributions made about token groups tend to force these groups into playing limited and caricatured roles. She says that typically, men respond to women in a work group, when the women are few in number, in ways which preserve the familiar form of interaction that they have had with women in the past. Often, Kanter argues, it is easier for the tokens to accept stereotyped roles than to fight them, even if their acceptance means limiting a token's range of expressions or demonstrations of task competence.

That one individual's expectancies may indeed operate to cause the behaviour of another to be in keeping with these expectancies is demonstrated in an experiment by Snyder, Tanke, and Berscheid (1977). In this experiment college men and women were informed that they would be participating in an investigation of the acquaintance process. They were asked to complete a biographical questionnaire on themselves to be given to another subject of the oppostive sex with whom they would engage in a telephone conversation which was tape recorded. The men were told that each partner of the dyad would receive a snapshot of the other and their pictures were taken. The women were not told this, nor were their pictures taken. Only the males in the dyad received a picture of a woman pre-rated by other college men to be particularly attractive or unattractive (not the female of the dyad). Tape recordings of each participant's conversational behaviour were analyzed by naive observers. These analyses revealed that women who were believed by their male counterparts to be physically attractive came to behave in a more friendly, likeable, and sociable manner in comparison with women who were presumed by their male counterparts to be unattractive. The researchers concluded that behavioural confirmation of the stereotypes involving physical attractiveness (i.e. "beautiful people are good people") had been obtained. Furthermore, if an individual holds a particular stereotype that individual's actions based on stereotyped attributions may cause the behaviour of the target person to conform to that stereotype.

Jones (1977) argues that not only can the behaviour of the stereotyped individuals be modified but also the expectancies they hold for themselves. With respect to performance evaluations at work the results could be that lower performance evaluations given to women would be consistent with their actual performance, but reinforcement and environmental contingencies established by the supervisor may have predetermined this to be the case.

The model outlined in Figure 1 summarizs these processes as they apply to women in the employment setting - specifically, as they apply to women in the Canadian Forces. The model will be discussed by starting at Box A and following the arrows. The supervisor throughout will be referred to as him, becauseonly one percent of the Sergeant to Chief Warrant Officer strength in the Canadian Forces is currently female. This is not necessarily because women have been systematically kept from moving up through the ranks, but rather because the largest portion of women in the Canadian Forces have entered within the last five to six years, when the number of trades in which women could be employed and the limits on their numbers in trades in which they were already employed were increased.

Boxes A, B and C have to do with the supervisor's views on the rights and roles of women in society, his expectations for the performance of women in the Canadian Forces, and his expectations for the particular woman under supervision. If, for instance, a particular supervisor has relatively egalitarian views on the rights and roles of women in society (Box A), he will probably expect women in the Canadian Forces to perform as well as men (Box B) - and all other things being equal, will expect that the particular female subordinate in question to perform as effectively as a man (Box C). On the other hand, if the supervisor holds more traditional views about women, it is less likely that he will expect women to perform as effectively in the Canadian Forces, and in particular will probably not expect the woman subordinate under supervision to do as well as man.

The expectations a supervisor holds for female subordinates will influence the way he treats the particular subordinate on the job (Box D). If he expects that women in the Canadian Forces are less able to carry out the more difficult tasks of their trades, he will be less likely to give them the same range and difficulty of tasks that he would give to male subordiantes. Furthermore, the superior's reinforcement behaviour towards his female subordinate is likely to be in keeping with his expectations. For supervisors with more traditional views this behaviour would include being more lenient with women displaying substandard performance - because this the expected performance from women - and, not being as rewarding towards good performance as they would be with men, viewing the women's performance as perhaps an exception to the rule. On the other hand, if the supervisor holds more egalitarian views he is likely to give women subordinates the same range and difficulty of tasks as male subordinates, and his reinforcement behaviour will not be much different for his male and female subordinates.

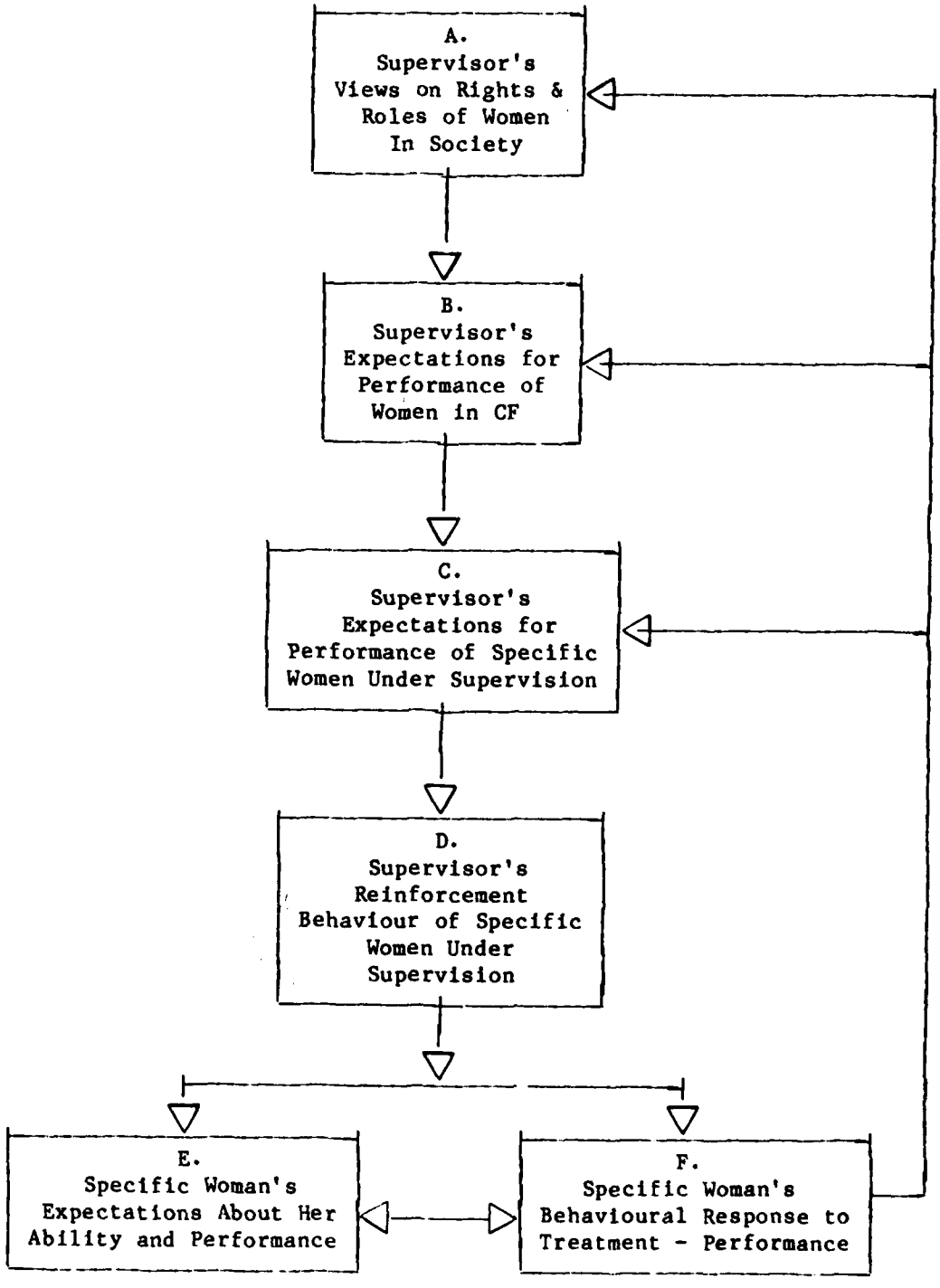


Figure 1. Schematic drawing of the impact of supervisor's expectancies for female subordinates' performance.

(872)

The expectations that the supervisor holds for the subordinate will mold to some extent the subordinate's expectations for herself (Box E). This is particularly so in the first few formative years of the subordinate's career. The subordinate gathers information about his or her capabilities at work, in part, from the way in which the supervisor treats substandard or good performance, and from the degree of difficulty and scope of tasks that the supervisor allows the subordinate to do. Because of this, the subordinate's own expectations will come to be in keeping with those of the superior. Granted, it is a two way street. Subordinates, by performing at a level which is not in keeping with the supervisors expectations, can modify the supervisors opinions of the subordinates; however, as Kanter (1976) has advanced, it is often easier for the supervisor to dismiss very good performance on the part of a woman as being an exception to the rule, than to change a basic belief.

Following this line of reasoning if the supervisor has low expectations for the subordinate in question the subordinate will eventually also come to have low expectations for herself (Box E). The subordinate's expectations will be reflected in the standard of performance she displays (Box F) and her performance as observed by her supervisor will feedback through the system. Her performance will impact on his views about the rights and roles of women in society as a whole, his expectations for the performance of women in the Canadian Forces, his expectations for the subordinates in question, and the cycle continues.

The model has been developed with women in the Canadian Forces in mind, but is applicable for any hierarchical organization in which women have been newly integrated, and in which women are competing with their male peers by means of a performance appraisal system for positions of higher rank and responsibility.

Whether a system such as this is at work in the Canadian Forces will be determined in a study scheduled for completion by Spring 1981. The supervisors views on the rights and roles of women in society (Box A) will be monitored by administering the Attitude Toward Women scale (AWS), a scale designed to assess attitudes varying from highly traditional and conservative to more egalitarian. The expectations that supervisors hold for their subordinates (Box C) will also be tapped through a survey which includes questions about the leadership potential of the subordinate, his or her ability to advance through the ranks, his or her potential to carry out all of the difficult tasks required in the subordinate's trade, as well as others. The survey also includes questions about how the supervisor would respond to outstanding performance or performance just below acceptable standard on the part of the subordinate, and the range of difficult tasks in the subordinate's trade the supervisor is willing to allow the subordinate to undertake (Box D). Finally, the same questionnaire will be given to the subordinates but worded from their perspective. Through this questionnaire the subordinates' expectations for their own careers in the Canadian Forces will be assessed.

Study Methodology

The sample for the study will include a representative group of women and a comparative group of men of non-officer status matched for rank and trade. The supervisors for this study are those who were the supervisors of the selected subordinates for their 1979 Performance Evaluation Report (PER).

TABLE I

SUPERVISORS' ATTITUDES TOWARD WOMEN

SUBORDINATES' SEX	TRADITIONAL/ CONSERVATIVE		LIBERAL/ EGALITARIAN
FEMALE			
MALE			

The design for this study is shown Table I. The independent variables are the sex of the subordinate and the supervisors' attitudes toward women as measured by the Attitude Toward Women Scale (AWS). The AWS is normally used as a continuous variable, but for the purposes of analysis, the group of superiors will be divided into thirds. The bottom 33 1/3 percent will be designated as those who express traditional attitudes, the top 33 1/3 percent as those who express more egalitarian views, and the middle 33 1/3 percent as those somewhere in between[1]. My principal interest, is in the two extreme groups.

1. This classification of supervisors may prove inappropriate upon examination of the supervisors' scores. For example, in the military supervisors may demonstrate a very restricted range of scores relative to other populations in society. If this is the case, alternate strategies for classifying the supervisors will be considered.

(814)

Firstly, the seventeen performance measures of the other ranks PER will be examined in relation to the supervisors' attitudes toward women and the sex of the subordinate. The first hypothesis is that the women subordinates will receive lower performance evaluations than men when both are rated by supervisors who adhere to traditional views about the roles of women in society. It is also hypothesized that there will be no differences in performance evaluations given to male and female subordinates by supervisors who do not adhere to traditional attitudes towards the roles of women. Finally, it is postulated that women who work for traditional supervisors will receive lower performance evaluations than women who work for supervisors who are more liberal in their views.

The second major set of dependent variables in this study is the results from the superiors' and subordinates' Expectancy Questionnaires. Insofar as the hypotheses are concerned the principal interest is in the relationships between female subordinates and their supervisors. Firstly, according to the model, supervisors who adhere to traditional attitudes towards women will have lower expectations for the performance of women at work than will the supervisors who have less traditional views. Secondly, supervisors who hold more traditional views will report less encouragement and reward of good performance and more leniency towards poor performance of their female subordinates than will supervisors with more egalitarian views. Finally, with respect to the women, it is hypothesized that women who work for traditional supervisors will have lower expectations for their own performance and career in the Canadian Forces than will women who work for more egalitarian supervisors.

The questionnaire information will be gathered through the Canadian Forces Personnel Selection Officer (PSO) network. Approximately 1500 questionniares will be sent out to the PSOs' offices for completion. They will get in touch with the pre-identified superiors and subordinates to have them complete the questionnaires. The questionnaires have been set up so that they are fairly self-explanatory, and so that PSO administrative clerks or other assistants can administer them in a classroom setting, if necessary.

Why Do It?

The value of this project can be measured along two dimensions (i) the degree to which it will increase our knowledge about the dynamics of superior/subordinate relationships; and (ii) it's immediate value to the Canadian Forces in it's day-to-day workings. In the first case, there is limited knowledge about the impact of supervisors' expectations on the performance or reported performance of their subordinates. In this study the degree of consistency between attitudes towards women, the expectations that specific superiors hold for specific subordinates, the superiors' reports about how they would respond to different standards of performance on the part of their subordinates the

subordinates expectations for their own careers and performance in the Canadian Forces, and the performance evaluations given to the subordinates will be determined. Should all of the hypothesized results be obtained there will be compelling evidence that the proposed model may be a useful description of the underlying processes. A definitive conclusion, however, can not be drawn about the cause and effect relationships hypothesized in the model. While the evidence may be very supportive, future research artifically inducing differing levels of expectations on the part of supervisors, will be required. The supervisors' consequent treatment of subordinates and their performance on the job would then have to be monitored by independent objective observers. In addition, the expectations the subordinates hold for themselves would have to be monitored. The commitment of research resources in the Canadian Forces to such an undertaking at this point would be premature.

While a definitive answer about all of the relationships involved may not be forthcoming, if the hypotheses are supported there would be sufficient evidence to suggest that there is a relationship between attitudes towards the rights and roles of women on the part of supervisors, and the kinds of performance evaluations women in the Forces receive. In addition, we may catch a glimpse as to why this is so. This would be sufficient reason for the Canadian Forces to introduce programs in their leadership training establishments which will help offset the effects of leaders' attitudes towards women, thereby helping to utilize women to their full potential.

(872)

REFERENCES

Bartol, K.M. & Butterfield, D.A. Sex effects in evaluating leaders. Journal of Applied Psychology, 1976, 61, 446-454.

Donahue, T.J. & Costar, J.W. Counselor discrimination against young women in career selection. Journal of Counselling Psychology, 1977, 24(6), 481-486.

Fidell, L.S. Empirical verification of sex discrimination in hiring practices in psychology. American Psychologist, 1970, 25, 1094-1097.

Heneman, H.G. Impact of test information and applicant sex on applicant evaluations in a selection simulation. Journal of Applied Psychology, 1977, 62(4), 524-526.

Jones, R.A. Self-fulfilling Prophecies. New York: John Wiley & Sons, 1977.

Kanter, R.M. Some effects of proportions on group life. Skewed sex ratios and responses to token women. American Journal of Sociology, 1976, 82(5), 995-990.

Rosen, B. & Jerdee, T.H. Influence of sex role stereotypes on personnel decisions. Journal of Applied Psychology, 1974a, 59(1), 9-14.

Rosen, B. & Jerdee, T.H. Effects of applicant's sex and difficulty of job on evaluations of candidates for managerial positions. Journal of Applied Psychology, 1974b, 59, 511-512.

Rosen, B. & Jerdee, T.H. Sex stereotyping in the executive suite. Harvard Business Review, 1974c, 52, 45-48.

Rosen, B. & Jerdee, T.H. The influence of sex-role stereotypes on the evaluation of male and female supervisory behavior. Journal of Applied Psychology, 1973, 57, 44-48.

Shaw, E.A. Differential impact of negative stereotypes in employee selection. Personnel Psychology, 1972, 25, 333-338.

Snyder, M., Tanke, E.D., & Berscheid, E. Social perception and interpersonal behaviour: On the self-fulling nature of social stereotypes. Journal of Personality and Social Psychology, 1977, 35(9), 656-666.

Terborg, J.R. Women in management: A research review. Journal of Applied Psychology, 1977, 62(6), 647-664.

Terborg, J.R. & Ilgen, D.R. A theoretical approach to sex discrimination in traditionally male occupations. Organizational Behavior and Human Performance, 1975, 13, 352-376.

SKINNER, Mary J., Air Force Human Resources Laboratory, Manpower and Personnel Division, Brooks AFB, Texas.

EVALUATION OF JOB APTITUDE REQUIREMENT WAIVERS FOR RETRAINED AIRMEN (Wed P.M.)

The viability of the current Air Force policy to waive 10 points of the minimum job entry aptitude requirement for airmen retraining from one occupational specialty to another was evaluated. Optimal performance and aptitude tradeoffs for retrainees were explored through performance comparisons with new recruits (non-retrainees) in technical training. Academic achievement and attrition rates for 19,885 retrainees and 231,317 non-retrainees attending 272 technical schools were examined. The specialities were categorized for analysis into 18 subgroups based on common mandatory aptitude entry requirements. Multiple linear regression analyses examined relationships between performance criteria and retraining status and aptitude predictors. Performance differences for retrainees and non-retrainees at various aptitude levels were then inspected. In general, the results supported the 10-point waiver practice for retrainees. The implications of instituting a more liberal policy are discussed.

EVALUATION OF JOB APTITUDE REQUIREMENT WAIVERS FOR RETRAINED AIRMEN

Mary J. Skinner
Air Force Human Resources Laboratory
Manpower and Personnel Division, Brooks AFB, Texas 78235

INTRODUCTION

Qualifications for entry into Air Force enlisted occupational specialties have for many years been based primarily on the aptitude requirements of the job. Job selection and assignment procedures require that enlistees meet minimum aptitude levels (AFR 39-1, 1977). A current policy which applies to enlisted personnel retraining from one Air Force specialty (AFS) to another is a notable exception to the standard of minimum aptitude performance. A 10-point aptitude waiver may be granted to participants in the Airman Retraining Program who are seeking to qualify for transfer to a second specialty (AFR 39-4, 1979).

The waiver procedure was incorporated in regulatory guidance for the retraining program in the early 1960s. At the time many AFSs had manpower imbalances with too few or too many personnel assigned to accomplish the specialty mission. To encourage retraining from overage to shortage specialties, the waiver practice was introduced. Less stringent job entry requirements were intended to improve job opportunities for enlistees by increasing the number of specialties for which most enlistees qualified. Air Force managers judged that such factors as a retrainee's prior exposure to military life, experience in a military occupation, initiative, motivation, or education would offset any detrimental influence of lower aptitude on his/her job performance in the new specialty (Reese, Note 1). Research on Navy and Air Force personnel suggests that previous military service experience has a positive influence on the performance of retrainees in technical training for second specialties (Booth, McNally, & Berry, 1975; Skinner & Alley, 1980). These studies provide indirect evidence in support of the aptitude waiver practice and underlying rationale, but do not permit an empirical assessment of whether retrainees' performance justifies discounting aptitude requirements and, if so, the magnitude of the allowable discount. Such information would be of value to Air Force managers concerned with the selection and classification of retrainees. The objective of the present study was to determine the optimum trade-off in performance and aptitude levels for retrained airmen.

APPROACH

The conceptual framework for the study rests on job assignment procedures in the Air Force. Mandatory minimum aptitude requirements for specialties are based on the lowest performance level needed to satisfactorily accomplish job duties. The performance of non-prior-service recruits with

minimum aptitudes was selected as the standard against which to compare the retrainees. Aptitude differentials for retrainees achieving performance levels equivalent to the recruit standard could then be inspected for various specialties. The recruits are referred to as non-retrainees, since they lacked prior experience in an AFS. The study methodology provided for the examination of aptitude and performance relationships in the technical training environment. Technical training was suited to the present research interest for several reasons. Performance comparisons between retrainees and recruits were possible, since the majority of retrainees preparing for their second specialty attend the same basic technical schools as recruits. Further, a representative sample of AFSs could be evaluated, because the majority of available specialties require formal skills training. A final important consideration was that achievement levels on the Armed Services Vocational Aptitude Battery (ASVAB), the aptitude assessment instrument used to measure qualifications to enter an AFS, are validated against technical school performance, specifically the final grade earned at course completion. This academic achievement measure was the primary performance criterion of interest, with secondary consideration given to pass/fail ratios.

The approach taken to evaluate the aptitude waiver policy is illustrated in Figure 1, which shows four possible research outcomes. Scores on the selector aptitude measure are plotted on the horizontal axis and technical school performance on the vertical axis. An increasing, curvilinear relationship between aptitude and performance for retrainees and recruits is represented with constant differences between the two groups at all levels of aptitude. In the outcome shown in Figure 1a, retrainees scoring 10 points below the job entry cutoff score perform as well as new accessions with the minimum qualifying score. This finding would support the current 10-point waiver procedure with its attendant advantage of enlarging the pool of enlistees eligible for retraining. An alternative outcome is shown in Figure 1b. Retrainees scoring more than 10 points below the cutoff achieve performance levels comparable to new accessions at the cutoff. A more liberal selection policy for retrainees would be feasible under this condition. In the event of severe career field manpower imbalances, the ability to retrain enlistees with lower aptitude, while maintaining adequate job performance would be valuable to managers. Figure 1c illustrates a third potential research outcome. Retrainees and recruits achieve equivalent performance levels at all aptitudes. Substandard performance by retrainees scoring below the cutoff is reflected. This condition would suggest that entry requirements presently applied to recruits be reinstated for retrainees. Manning flexibility currently realized by retraining managers would likely decrease. A final outcome is shown in Figure 1d. New accessions at all aptitude levels are portrayed as performing better than retrainees. This finding would suggest the need for more stringent qualifications for retrainees as compared with recruits. In this event, the manpower pool of prospective retrainees would be substantially curtailed.

METHOD

Data on 251,202 enlistees attending 272 basic technical schools between July 1973 and December 1977 were extracted for analysis from Air Force

personnel records. The variables used as predictors were retraining status and aptitude. The retraining status of airmen listed on historical retraining files was verified to confirm that their assigned AFS before retraining was different from the technical school specialty attended. Non-retrainees were recruits in training in conjunction with their initial specialty assignment. Individual aptitude scores were derived from the ASVAB.

The two performance criteria were final school grade and pass/fail status. Final grades were available for a subset of the enlistees who graduated from training. To equate performance ratings across schools, grades were standardized to yield a mean equal to 50.0 and a standard deviation of 10.0. The attrition criterion (pass/fail) was based on the type of final disposition from training. School graduates were identified as passes. Enlistees who had terminated training for not meeting performance standards due to academic deficiencies, medical disqualification, death, or other/unknown reasons were coded failures.

The 272 schools were categorized for analysis based on the aptitude prerequisite for entering the specialty. Schools with the same entry score requirement on the Mechanical (M), Administrative (A), General (G), or Electronic (E) index of the ASVAB were grouped together. For example, AFSs with a selector aptitude index (SAI) of 40, 50, or 60 on the Mechanical (M) composite were combined to form M40, M50, or M60 subgroups. The categorization procedures resulted in 18 SAI subgroups.

Descriptive statistics were used to examine sample characteristics on predictor and criterion variables within SAI subgroups. Relationships between the training performance criteria, retraining status and aptitudes were explored using multiple linear regression techniques (Bottenberg & Ward, 1963). The starting regression model contained predictor variables representing retraining status, aptitude (linear and curvilinear), and their interactions. Tests of significance of full and restricted model comparisons were conducted in sequence for retraining status and aptitude effects, and if appropriate, were followed by a test for the interaction of these two basic predictors and for aptitude curvilinearity. The direction and magnitude of performance differences for significant effects were evaluated for each subgroup, and conclusions drawn concerning minimum job aptitude requirement levels for retrained airmen. Two sets of parallel analyses were conducted, one for each of the technical school performance criteria.

RESULTS

Summary descriptive data on predictor and criterion variables for the final school grade analysis sample are shown in Table 1. The total number of school graduates for whom a final score was available was 227,850. Of these approximately 8% (17,502) were retrainees and the remaining 92% (210,348) were non-retrainees. The sample breakdown by SAI subgroup indicated that retrainees typically accounted for less than 20% of the cases in a subgroup. The mean of the standardized final school grades showed that retrainees in all subgroups scored higher than the average grade (50.0)

achieved in technical training. Comparisons of retrainees and non-retrainees revealed that retrainee performance was superior in all subgroups with mean grades ranging from 1 to 12 points higher. In regard to aptitudes, non-retrainees overall performed better than retrainees. Mean aptitude scores for non-retrainees ranged from 1 to 11 points higher in 16 of 18 SAI subgroups.

Training completion and aptitude data for the pass/fail analysis sample are shown in Table 2. The pass/fail sample totalled 251,202 enlistees of whom 8% (19,885) were retrainees and 92% (231,317) were non-retrainees. The percentage of passes generally exceeded 80% in both the retrained and non-retrained groups. In the majority of the subgroups (10 of 18), the percentage of retrainees graduating was higher than non-retrainees. Aptitudes achieved by the non-retrained group were on the average higher than the retrained group in most subgroups.

Results of the statistical analyses summarized in Table 3 were very consistent for the final school grade criterion in each of the 18 SAI subgroups, but somewhat less regularity was noted in the pass/fail analyses. For the final school grade criterion, academic performance levels were found to differ for retrainees and non-retrainees with equivalent aptitude in 17 of 18 SAI subgroups. Retrainees were found to achieve consistently higher grades than non-retrainees at fixed aptitude levels in these subgroups. In all of the subgroups, school performance for both retrained and non-retrained groups varied as a function of aptitude level. Grade levels were observed to improve with increases in aptitudes. Interaction effects for the retraining status and aptitude predictors were not commonly detected. Rather, the differences in grade level between retrainees and non-retrainees were consistent across aptitudes. The typical form of the relationship between the grade criterion and aptitude predictor was curvilinear. A similar trend was found for analyses of the attrition criterion in about one-half of the subgroups. Training completion rates differed for retrainees and non-retrainees, with retrainees having higher probabilities of completing training at fixed aptitudes. A positive, linear relationship between pass/fail rates and aptitudes was observed in the majority of the subgroups. The predominant pattern of results across both criteria was that performance typically improved as aptitude levels increased. Further, at fixed aptitudes higher performance levels were attained by retrainees than recruits.

Further analyses were conducted to specifically address the aptitude waiver issue. The aptitude at which retrainees achieved a performance level equivalent to non-retrainees at the aptitude cutoff was determined for both training criteria. The difference between the retrainee score and the cutoff score was then computed to determine the magnitude of the allowable aptitude discount, if any, for each subgroup. These data are summarized in Table 4 by subgroup for both criteria. The discount value is expressed as greater than, equal to, or less than the current 10-point waiver. In schools where retrainee and recruit performance did not differ or where retrainees performed poorer than recruits, a discount equal to zero is presented. For the final school grade criterion, the waiver data indicate that retrainees with 10 aptitude points below the cutoff achieved final grades comparable to or higher than the recruits in 17 of 18 subgroups.

Schools with an M60 selector level were an exception in that retrainees and recruits at the cutoff achieved equivalent grades. In 14 subgroups the critical aptitude level for retrainees was more than 10 points below the cutoff. On the attrition criterion discounts equal to or larger than 10 points were found to be appropriate in about one-third of the subgroups. However, the overall graduation rates for retrainees and non-retrainees did not differ in nine additional subgroups. In two subgroups (M50 and A40) training completion rates were lower among retrainees than recruits at the cutoff.

DISCUSSION/CONCLUSIONS

The research results support the policy of waiving 10 points of the job entry aptitude requirement for enlisted personnel retraining to new military occupations. Analysis of academic achievement strongly suggested that final school grades of retrainees to whom 10 points of the aptitude requirement are discounted would not be expected to fall below accepted standards. Results of the attrition criterion analyses were less consistent than the academic performance analyses, but provided some evidence in defense of the waiver practice. Training completion rates for retrainees were not adversely impacted by an entry requirement dispensation in some schools. However, other subgroup analyses revealed that higher attrition occurred among retrainees with a waiver than the non-retrainee performance standard at the entry cutoff.

Beyond the support demonstrated for the operational 10-point waiver policy, the present study has important implications for future retrainee management. The results suggest that a more liberal aptitude discount would be defensible for most Air Force specialties in the event that the manpower requirement for retrainees sharply increased. Current analyses indicate that retrainees with a 15-point waiver would be expected to maintain acceptable academic performance levels in most specialties. The final school grade analyses would not, however, support a waiver of more than 10 points in Mechanical 40, Administrative 40, and General 40 schools. If less stringent entry requirements were adopted, recognition would need to be given to the likelihood of increased attrition rates among the retrainees. In most schools, these rates would be expected to exceed the standard set for recruits at the current aptitude cutoff. Collectively, results for both the academic and attrition measures indicate that technical training performance standards would be met by retrainees with a 15-point aptitude waiver, except in those specialties with an aptitude cutoff score equal to 40. In the latter specialties, the present 10-point waiver should be retained.

REFERENCE NOTES

1. Reese, J. (HQ, USAF/MPPPN, Washington, D.C.) Personal Communication, 18 Oct 79.

REFERENCES

Air Force Regulation 39-1. Airman classification regulation. Washington, D.C.: Department of the Air Force, 1 Jun 77.

Air Force Regulation 39-4. Airman retraining program. Washington, D.C.: Department of the Air Force, 28 Nov 79.

Booth, R. F., McNally, M. S., & Berry, N. S. Demographic characteristics, psychosocial perceptions, and performance of "strikers" accepted for Navy paramedical training. Report No. 75-75, AD-A020 317. San Diego, CA. Naval Health Research Center, 1975.

Bottenberg, R. A., & Ward, J. H., Jr. Applied multiple linear regression. PRL-TDR-63-6, AD-413 128. Lackland AFB, TX: Personnel Research Laboratory, Aerospace Medical Division, March 1963.

Skinner, M. J., & Alley, W. E. Performance of retrained airmen in Air Force technical schools. AFHRL-TR-80-7. Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory, April 1980.

(886)

Figure 1. Hypothetical relationships between performance and aptitude for retrainees and recruits

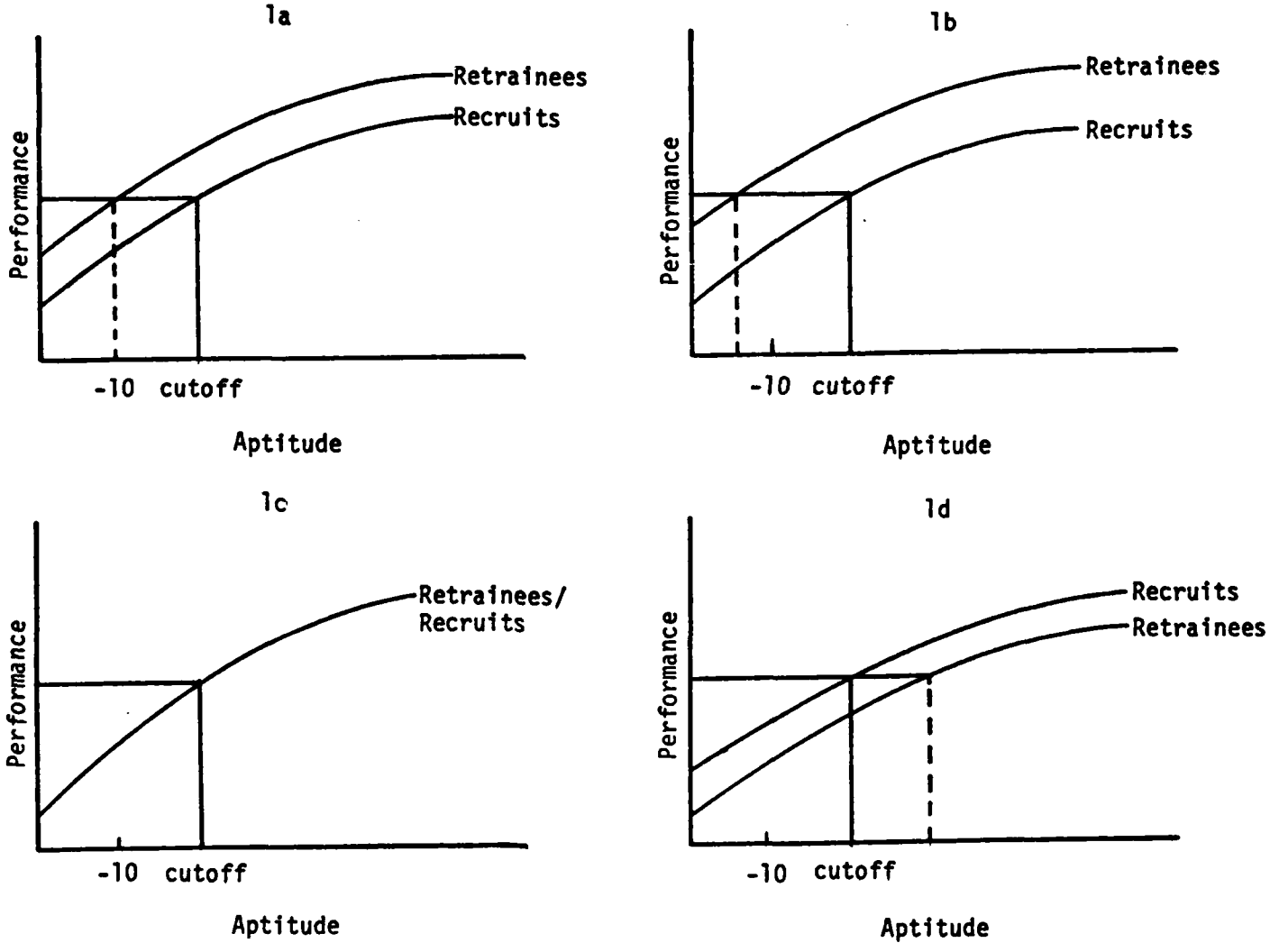


Table 1. Summary Statistics for Final School Grade Criterion Analysis

SAI Subgroup	Retrainee N	Retrainee Final Grade Mean	Retrainee Final Grade SD	Retrainee Aptitude Mean	Retrainee Aptitude SD	Non-Retrainee N	Non-Retrainee Final Grade Mean	Non-Retrainee Final Grade SD	Non-Retrainee Aptitude Mean	Non-Retrainee Aptitude SD
M40	1,351	52.39	10.12	60.25	18.33	28,807	49.89	9.98	61.30	20.06
M50	1,974	53.65	10.44	63.45	16.12	36,705	49.88	9.92	68.35	17.03
M60	160	50.43	9.52	61.94	15.95	12,649	50.07	9.99	73.42	14.04
A40	347	55.42	9.95	61.15	16.60	9,451	49.84	9.94	62.32	15.83
A50	66	58.09	9.56	66.14	15.76	970	49.34	9.77	66.09	13.56
A60	1,614	55.41	9.52	65.91	14.42	15,047	49.47	9.87	69.52	15.08
A70	350	53.29	9.03	75.46	11.06	305	46.52	9.61	79.95	9.60
A80	230	55.28	9.48	82.96	8.63	3,213	49.57	9.93	87.24	6.79
G40	1,036	54.21	9.39	63.31	15.15	34,197	49.97	9.99	64.96	15.32
G50	272	50.94	11.18	65.97	14.85	4,483	47.56	10.26	68.62	13.51
G60	4,956	52.46	9.85	70.51	13.36	22,824	49.95	9.84	76.32	12.40
G65	144	52.54	10.28	78.02	12.76	375	48.91	9.69	84.44	9.49
G70	35	55.76	8.93	73.57	10.25	33	43.90	7.01	84.55	7.00
G80	671	52.22	9.89	81.59	10.87	4,217	49.68	9.91	87.91	6.76
E50	336	54.79	9.34	65.34	14.12	1,790	49.12	9.84	68.62	12.72
E60	489	52.05	10.26	71.61	14.62	3,092	49.65	9.94	71.20	11.31
E70	26	53.10	8.73	80.96	8.66	411	48.10	10.10	81.92	8.72
E80	3,445	53.92	9.93	84.88	7.71	31,779	49.56	9.91	86.53	6.92
Total	17,502					210,348				

Table 2. Summary Statistics for Pass/Fail Criterion Analysis

SAI Subgroup	Retrainee					Non-Retrainee				
		Pass/Fail		Aptitude			Pass/Fail		Aptitude	
	N	% Pass	%Fail	Mean	SD	N	% Pass	% Fail	Mean	SD
M40	1,521	95.60	4.40	60.31	18.53	30,971	95.75	4.25	60.92	20.21
M50	2,102	95.24	4.76	63.16	16.19	38,274	96.48	3.52	67.91	17.24
M60	177	93.22	6.78	61.67	16.10	13,286	95.86	4.14	73.15	14.11
A40	377	93.10	6.90	60.46	16.63	9,898	96.12	3.88	62.15	15.82
A50	66	100.00	0.00	66.14	15.76	998	98.00	2.00	66.10	13.55
A60	1,683	97.33	2.67	65.90	14.41	16,970	94.87	5.13	69.91	14.91
A70	353	99.43	.57	75.42	11.05	320	97.19	2.81	79.78	9.53
A80	239	96.23	3.77	83.03	8.61	3,402	95.36	4.64	87.10	7.03
G40	1,101	96.46	3.54	63.23	15.26	35,469	97.42	2.58	64.90	15.29
G50	302	99.34	.66	65.71	14.68	4,635	96.74	3.26	68.52	13.53
G60	6,010	95.12	4.88	70.07	13.32	27,972	94.35	5.65	75.69	12.32
G65	159	91.19	8.81	76.86	13.07	407	94.10	5.90	86.86	9.68
G70	73	90.41	9.59	74.73	10.33	48	93.75	6.25	85.52	7.23
G80	841	91.32	8.68	81.46	10.85	5,301	81.04	18.96	87.73	6.89
E50	367	92.37	7.63	64.97	14.26	2,077	86.81	13.19	67.74	12.76
E60	593	86.34	13.66	71.57	14.46	3,721	84.31	15.69	70.47	11.36
E70	27	96.30	3.70	80.74	8.57	427	97.19	2.81	81.80	8.72
E80	3,894	91.88	8.12	84.57	7.82	37,141	87.81	12.19	85.96	7.13
Total	19,885					231,317				

Table 3. Summary of Statistical Findings

Source of Effect	SAI Subgroup																		
	Mechanical			Administrative					General						Electronic				
	40	50	60	40	50	60	70	80	40	50	60	65	70	80	50	60	70	80	
Final School Grade																			
Retraining Status (R)	*	*	ns	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
Aptitude (A)	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
R X A Interaction	*	*	-	ns	ns	ns	ns	ns	*	ns	*	ns	ns	*	ns	ns	ns	*	
Curvilinear Aptitude	*	*	*	ns	ns	*	ns	*	*	*	*	*	ns	*	*	*	ns	*	
Pass/Fail																			
Retraining Status (R)	ns	*	ns	*	ns	*	ns	ns	ns	*	*	ns	ns	*	*	*	ns	*	
Aptitude (A)	*	*	*	*	ns	ns	ns	*	*	*	*	*	ns	*	*	*	ns	*	
R X A Interaction	-	*	-	*	-	-	-	-	-	ns	ns	-	-	ns	ns	*	-	*	
Curvilinear Aptitude	*	*	ns	ns	-	-	-	ns	ns	ns	*	ns	-	ns	ns	ns	-	*	

Note. An asterisk(*) in the table indicates statistical significance ($p < .05$) for a predictor. The designation ns specifies a non-significant predictor. A dash (-) indicates F-test was inappropriate and assumed to be non-significant.

Table 4. Aptitude Differentials for Retrainees by SAI Subgroup on Final School Grade and Pass/Fail Analyses

SAI Subgroup	Final School Grade	Pass/Fail
M40	10	0
M50	>10	0
M60	0	0
A40	>10	0
A50	>10	0
A60	>10	>10
A70	>10	0
A80	>10	0
G40	10	0
G50	>10	>10
G60	>10	<10
G65	>10	0
G70	>10	0
G80	>10	>10
E50	>10	>10
E60	>10	>10
E70	>10	0
E80	>10	>10

SLIMMAN, LCol D.J., Director Personnel Development Studies, National Defence Headquarters, Ottawa, Ontario.

LATERAL SKILL PROGRESSION. (Tue P.M.)

Although the personal skills and motivation required for leadership/supervision/management are different from those required for technical or trade proficiency, the Canadian Forces uses advancement in rank as its only means of reward for superior performance. This use of rank, which is essentially an instrument and symbol of authority, may be dysfunctional as a means for rewarding those whose skills lie in their technical, and not in their supervisory expertise. Lateral Skill Progression (LSP) is therefore proposed as a career advancement system for measuring, recognizing and rewarding technical/trade proficiency concurrently with, but independently of, supervisory/leadership skills. Studies are currently underway to determine the most effective means, applicable to all Military Occupational Classifications, of measuring trade skills separately from supervisory/leadership abilities.

LATERAL SKILL PROGRESSION AND PERFORMANCE EVALUATION

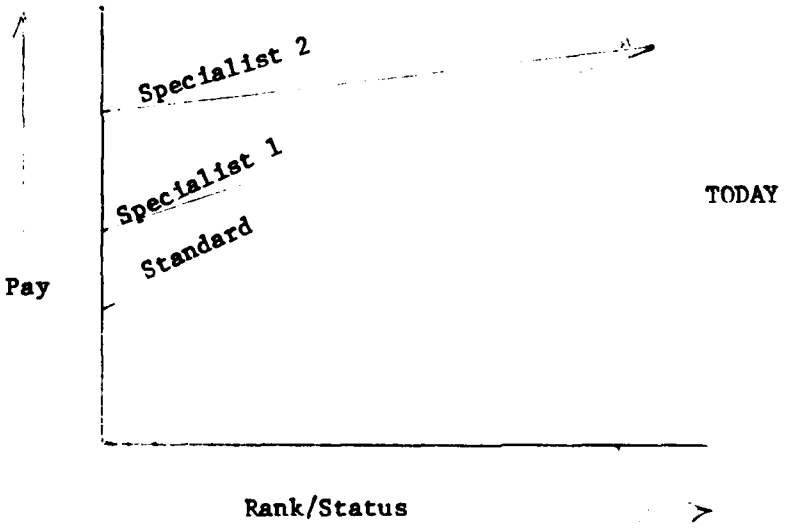
LIEUTENANT-COLONEL DONALD J. SLIMMAN, CANADIAN FORCES

DIRECTOR PERSONNEL DEVELOPMENT STUDIES

BACKGROUND

1. Like most other military organizations, the Canadian Forces use promotion in rank, with its attendant rewards in compensation, social status and perquisites, as the sole means of recognizing superior performance and marking the visible steps in career advancement.(Figure 1).

FIGURE 1 (not to scale)



NOTE: Each trade is assigned to one of three trade groups - Standard, Specialist 1 or Specialist 2 - depending on its technical complexity

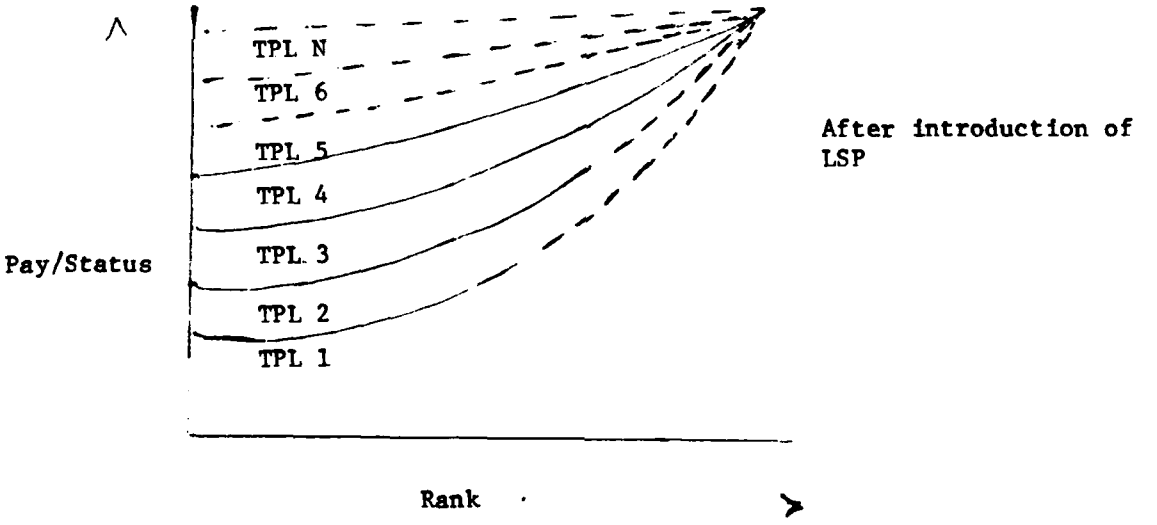
If, however, rank is seen as an instrument and symbol of authority, as it traditionally functioned in the pre-technical armed forces throughout preceding centuries, its use as a reward for superior performance in a purely technical/trade sense may be inappropriate and even dysfunctional. Although the personal skills required for leadership/supervision/management are different from those required for technical or trade proficiency, the CF have no other means available for motivating servicemen and women to

increase their technical knowledge and performance levels and to reward them when they do. From our many discussions with servicemembers, we found that a substantial proportion of men promoted to higher rank solely on the basis of their technical skills turn out to be poor leaders and supervisors. In cases such as these, the CF can be said to have lost a good tradesman and gained a poor leader.

LATERAL SKILL PROGRESSION

2. The concept of Lateral Skill Progression (LSP) is currently being studied by the CF in order to move beyond a unidimensional reward system. LSP can be defined as a career advancement system for measuring, recognizing and rewarding technical/trade proficiency concurrently with, but independently from, supervisory/leadership skills. (Figure 2).

FIGURE 2 (not to scale)



NOTES: 1. TPL - Trade Progression Level

2. In most, if not in all, trades supervisors would be required to be at least "competent technicians", with TPLs in excess of the basic levels.

3. The maximum number of TPLs remains to be determined, and might vary between trades.

(896)

In my capacity as Director Personnel Development Studies, I am the Project Officer of a twelve-man team tasked with determining the feasibility of applying LSP to each of the 100 other ranks trades, known as Military Occupation Classifications (MOCs). The task is formidable. Five MOCs, selected to form a representative sample across environments, technical complexity, leadership requirements, homogeneity and diversity of employments, will be examined in the initial year-long stage of the study. Each establishment position within the selected MOCs will be examined to determine its degree of trade complexity and allotted a Trade Progression Level (TPL). Simultaneously, each position will be examined to determine its minimum rank level consistent with effective command and control. Rank, let us repeat, will no longer be used to represent trade skill, or longevity, but only to serve as a symbol and instrument of authority. Rank will therefore be allotted "from the bottom up". In other words, when a certain number of workers are together in a specific group, a supervisor (or worker/supervisor) with one rank higher will be needed. The number will be determined primarily by the appropriate span of control and may differ from one type of employment to another, depending on circumstances. This will represent a substantial change from current CF trends, where some trades show a rank to rank ratio of 1:1 at certain levels. In colloquial terms, we have too many chiefs and not enough Indians. Moreover, many of these chiefs are in fact highly skilled Indians and what we propose to do is to remove their headdress and give them instead different insignia to mark their technical skills, different in nature to the interpersonal skills required by the true chief but equally essential to the efficient functioning of the organization. Removing headdress is a touchy business, however; and an important part of the LSP study will be to discover whether personnel can be motivated to seek the rewards and status of the master

craftsman as eagerly as they now seek those of increasing rank. This is of course one of the many areas in which we shall be seeking the assistance of behavioural science from our colleagues at CFPARU, and through them, of research conducted in the USA.

TRADE PROGRESSION LEVELS

Delineation of the various TPLs will rely to a great extent on occupational analysis (OA) techniques developed for CF use by our colleagues in the Directorate of Military Occupational Structure, two of whose officers delivered a paper on OA to this conference last year. The process of sorting the entire range of skills and knowledge required to perform the various tasks called for within each MOC into a hierarchical TPL structure will be complicated and time-consuming. Needless to say, the end results, both within and between trades, must be seen to be demonstrably fair, in particular within the context of the wide range of occupations represented in our unified services. As an example of one of the problems to be faced is the question of whether the number of TPLs should be the same for each MOC. (Figure 2, Note 3). Representatives of the technical trades tend to argue that the number of TPLs should reflect the technical complexity of each trade, so that the Radar Systems Technician trade might, for example, have double the number of levels as the Infantryman MOC, where the leadership component predominates over the technical. Spokesman for the Combat Arms, however, would argue that in the same way that each private can aspire to the rank of Chief Warrant Officer, he should also be able, theoretically at least, to attain the highest TPL in the CF. The number of TPLs, they maintain, should be the same for each trade, with a greater proportion of positions in the technical trades being established for higher TPLs than in the less technical trades. Since substantial pay increases are envisioned for each TPL increase, the question is not academic.

898

SELECTION FOR CAREER ADVANCEMENT

As stated earlier, career progression is currently marked and rewarded solely in terms of rank, with promotion selections being made annually by centrally-controlled merit boards. These merit boards use as their primary tool the annual Personnel Evaluation Report (PER), similar to the US efficiency report. The PER contains a narrative report and some 17 performance measures, each scoreable on a 1-7 point scale. Merit board members, who are given access to all PERs written on each servicemen, list each qualified member, by trade and by rank, in a merit order from which promotions are made to fill vacancies as they occur. To be qualified, a serviceman must have been recommended by his commanding officer, be medically fit, have completed a certain minimum period in his current rank, and have completed certain trade qualifications.

If, then, we are to offer career advancement in two dimensions (rank and trade skill) vis-à-vis one as at present, we must develop two sets of criteria for such advancement. The current PER described above has seventeen performance measures, as follows:

1. Preparation and planning
2. Delegation
3. Performance under stress/pressure
4. Cooperation
5. Command and self-assertion
6. Support of subordinates
7. Briefing Others
8. Knowledge of trade/knowledge of job when employed out of trade
9. Ability to apply his/her knowledge
10. Adaptability

(899)

11. Initiation
12. Appearance and bearing
13. Supervision
14. Ensuring understanding of assignments
15. Responsibility
16. Conduct
17. Learning from experience

It is not difficult to associate some of these measures with the interpersonal skills required for leadership/supervisory abilitities, and others with the individual qualities that might contribute to superior technical/trade/job performance. Indeed, analysis conducted by Bain, Skinner and Rampton in CFPARU Research Report 80-6 shows that two major factors emerge from the PER form. The critical requirements of Delegation, Command and Self-Assertion, Support of Subordinates, Briefing Others, Supervision, and Ensuring Understanding of assignments, load heavily on the first factor, referred to as Influencing. This factor, it would appear, would continue to be applicable for merit boards in considering potential for promotion in rank, where such rank was associated solely with leadership/supervisory potential. The remaining PER performance measures, with the exception of Conduct and of Appearance and Bearing, load heavily on a second factor described by Bain and group as Individual Effectiveness. This factor may be useful should merit boards be asked to select candidates for TPL advancement from among those meeting the necessary qualifications - such as, it is proposed, time in current TPL, trade exams passed, specialist qualifications, hands-on testing by supervisors and a commanding officer's recommendation.

(900)

That these two factors do fall out of the current PER demonstrates that the LSP concept is compatible with underlying regularities which are apparent in the current appraisal system. At the same time, further work, presumably using the resources of CFPARU, will be required before a definitive system can be introduced.

This paper has so far looked at LSP solely in the context of an established system. Despite the fact that the CF will be going further than any other military force in creating a career advancement system with independent, and yet concurrent, dimensions of recognition and reward, there appears to be general acceptance that LSP will represent a substantial improvement over the current unidimensional system. The introduction of such a radically new system creates its own problems, particularly in the initial transition phases, and the study team will have to demonstrate that these problems will not outweight the apparent advantages of the LSP concept. What may be required is a balancing act; bringing LSP into being quickly enough to overcome the problems inherent in the present system, and yet gradually enough to permit individuals to adjust to a new set of values and to a new system of career progression.

SMITH, Brandon B., Minnesota Research and Development Center for Vocational Education, University of Minnesota, Minnesota, Minneapolis.

COGNITIVE STRUCTURE OF TECHNICAL KNOWLEDGE A FREE ASSOCIATION TECHNIQUE (Wed P.M.)

The purpose of this paper is to (a) discuss the results of research concerning the free association methodology as a valid way to produce a hierarchical associative conceptual structure of a technical subject matter and to (b) review other research using the associative methodology to teach for conceptual structive while accounting for individual differences in learners.

Three studies were originally conducted to assess the associative conceptual structure of radio and television repair persons (Moss, 1968; Smith, 1968) and for mobile radio communications repair persons (Pratzner, 1969). Two sets of four flexible and inflexible workers in each occupation were identified as performing their job in the most and least effective manner respectively. Each flexible and inflexible worker was given one minute to respond to each of about 200 technical electronic terms with as many technically relevant terms as possible. The free association responses were pooled for the four flexible and inflexible workers respectively in each of the two occupations. Data were keypunched and subjected to a hierarchical factor analysis program using the Relatedness Coefficient (RC Coefficient) [Garshoff and Houston, 1963]. Factors were labelled by six experts in the field of electronics.

The results for the flexible and inflexible workers in the two occupations were the same. The flexible workers produced a hierarchical associative conceptual structure where one factor subsumed all lower order factors and were linked to it in a psychologically associative, meaningful way. On the other hand the inflexible workers who responded to the same technical terms with the data analyzed in the same way produced a more disjointed structure with no single subsuming concept for their associative structure. The methodology seems to have implications for the development and evaluation of curriculum and instructional delivery systems for both vocational education and at least some types of military training programs. The procedure produces a pictorial map of a technical associative conceptual structure which seems to have direct implications for curriculum design and for group or individualized instruction. Its major contribution is that it represents an alternative to the task analysis approach for organizing and sequencing instruction.

COGNITIVE STRUCTURE OF TECHNICAL KNOWLEDGE: A FREE ASSOCIATION TECHNIQUE

Brandon B. Smith
Minnesota Research and Development Center
University of Minnesota
Minneapolis, Minnesota

Introduction

Both the military and vocational education make extensive use of the task analysis approach to identifying, organizing and sequencing technical content for their training programs (Christal 1979); (Fryklund 1933); (Ellis, Wielfack, Frederick, 1979); (Silverman 1970). Instruction based on a task analysis typically proceeds inductively from a part to whole sequence using a behavioristic model. Instruction begins with the basic, fundamental facts or concepts and systematically builds towards higher orders of behaviors (Gagne 1965) (e.g. concept formation, principle learning and problem solving). The typical task analysis therefore tends to produce an instructional delivery system for either group or individualized instruction which for the most is organized and presented to the learner as part to whole instruction. Task analysis therefore produces a logical structure of behaviors from which inferences must be made about the necessary cognitive knowledges. The procedure does not take into consideration either the state of readiness of the learner nor is it organized or presented in a manner which is most psychologically meaningful or relevant to the learner. The problem has been to identify a procedures by which the conceptual structure of a technical knowledge can be identified thus enabling tasks to acquire a more relevant meaning or content.

Purpose

The purpose of this report is to (1) discuss the rationale and methodology used to conduct two studies utilizing the free association technique to identify the hierarachical associative conceptual structure of technical workers in the field of radio and television repairmen and model radio communications repairmen, and (2) to briefly review other applications of the free association methodology.

Rationale

The rationale used for this series of studies was based on the work of Deese (1962, 1964) at Johns Hopkins; Bertram Garshof and John Houston (1963) at the University of Michigan; Paul Johnson (1964, 1967) at the University of Minnesota; David Ausubel, and Jerome Bruner. All of these individuals are interested in the structure of knowledge and verbal learning. Their work with the concepts of verbal learning and/or the free association technique has made it possible to make the following assumptions with a reasonable degree of assurance.

1. Technical fields or occupations have a body of technical knowledge related to the quality of performance of workers in the field.

2. The technical knowledge of a field possessed by an individual consists of concepts which are the psychological meaning of the tools, processes, units of measures of the field.

3. Individuals organize their technical concepts into an integrated structure dependent upon its functional relationship to other concepts.

4. Technical terms or words are verbal labels for concepts.

5. Experienced workers who are performing in the technical field possess the verbal labels necessary to assess their meanings of various technical concepts.

6. The associative meanings of a concept is defined as the total free association response distribution which a given stimulus word elicits.

7. The first response to a stimulus word is more highly related to the psychological meaning of a stimulus word than later responses.

8. The total stimulus word domain can be identified by repeated administrations of a free association instrument to a group of individuals until no new technical words are elicited as responses.

9. The associative relatedness (meaning) of two concepts is determined by the extent of overlap between the two associative response distributions (meanings) for an individual or group of individuals.

10. The associative structure of technical concepts can be generated for an individual or group of individuals by calculating a matrix of relatedness coefficients (RC between all possible pairs of associative meaning to stimulus words).

11. The graphic pictorial structure of technical concepts can be identified by subjecting the relatedness (RC) matrix to some type of mulidimensional solution, and then systematically redefining the associative meaning of factors and subjecting it to higher order solutions.

In summary, it is believed that it is possible to empirically and objectively identify the technical conceptual structure for any field or occupation by identifying an individual or group of individuals who are performing the work in a manner we wish students to emulate. Once the conceptual structure is produced it is possible to (a) present it in terms of either group or individualized instruction, (b) organized from whole to part instruction, and (c) relate psychomotor skills to the cognitive structure at appropriate, relevant points in the curriculum.

Procedures

The first application of the free association technique conducted in the field of vocational education dealt with electronic appliance repairmen (Moss 1968); (Smith 1968); Pratzner 1969). An attempt was made to test the validity and reliability of an empirical procedure using the free association methodology to identify and compare the conceptual, technical, associative structure of four (4) flexible and inflexible radio and television repair persons and four (4) radio communications repair persons. Pairs of flexible and inflexible workers were identified in each of four (4) major Minneapolis/St. Paul electronic firms employing at least nine other journeymen workers in the same payroll job. Each worker had (a) at least three years of trade experience and (b) had been out of formal training for at

least two years. Flexible and inflexible workers were identified by the immediate supervisor as four pairs of workers who were and who were not capable of performing (a) the greatest variety of appliance repair tasks, and (b) the most novel, unusual or complex repair tasks. From the supervisors nomination two pairs of the most and least flexible workers were identified in the occupation of radio television repair and mobile radio communications repair.

Stimulus Words/Administration

A total of about 450 technical electronic terms were identified from technical manuals and groups of experts for the two appliance repair fields. A random sample of 163 radio and television words and 184 words for mobile radio communications were selected respectively.

The stimulus words were randomly arranged into four different free association test booklets in order to control for possible chaining effect and were administered to the flexible and inflexible workers in two, two-hour test sessions on successive days in each of the two occupations. Each page of the booklet provided a different stimulus word at the top of each page and sufficient space to write twenty-five responses to each stimulus word. Each worker was given only one minute to respond to a stimulus word with as many technically related terms as possible. A tape recorder with a signal tone replicated at one minute intervals was used to control response time. Twenty words were randomly selected and randomly placed in each of the two test booklets to assess the stability of the free response techniques.

Pooling Responses

The free association responses for the four flexible and the four inflexible workers in each of the two occupations were "pooled" respectively. The pooling process eliminated idiosyncratic responses and resulted in a relatively unbiased, salient, agreed upon, rank ordered associative meaning for each stimulus word elicited from a purposive sample of four flexible and inflexible workers in the two appliance repair fields.

A majority of the calculations in the two studies used the relatedness coefficient (RC). The RC is a measure of the extent to which two response distributions overlap or the extent to which the associative meaning of two words are similar or different. The magnitude of the RC is inversely proportional to the psychological distance between two concepts. The RC ranges from .00 to 1.00 where .00 indicates no relationship between two associative concepts and a 1.00 means the associative meaning of the concepts are identical.

RC coefficients were calculated between the pooled associative meaning of the twenty test-retest stimulus words to assess the reliability of the procedure.

The RC statistic was also used to compute the relationship among all possible pairs of the 163 and 184 stimulus word matrics for the radio television and radio communications occupations respectively, thus producing a first order associative structure matrix for the flexible and inflexible workers in each of the two occupations. The 163 and 184 associative structure matrix was than factor

analyzed using the principle component analysis with a varimax rotation. Operationally the factors were statistically defined by those stimulus words which loaded .300 or higher on a factor.

The associative meaning for each factor was defined by pooling all of the original responses for each stimulus word which loaded .300 or higher on the factor. A new second order associative structure matrix was generated and subsequently factor analyzed. The procedure was repeated until it was no longer possible to reduce the factor structure. Associative (pictorial) maps were drawn showing the relationship of lower order to higher order factors and electronic experts in the field were asked to provide labels for each lower order and higher order concept by (a) looking at the response distribution to a factors and (b) looking at the relationship among factors.

Findings

1. The pooled associative meaning for flexible and inflexible workers in the two occupations were reliables. The median RC coefficient for the twenty randomly selected test-retest stimulus words was .80.

2. The associative methodology produces technical conceptual maps with face and content validity for flexible and inflexible workers in each of the two occupations.

3. The technical, conceptual structures (maps) for flexible and inflexible workers for each of the two occupations are different in several highly meaningful ways.

    a. Flexible workers had a larger and different technical vocabulary as measured by (1) the magnitude of the pooled response distribution, (2) the number of different technical terms used and (c) their meanings to 24 and 33 percent of the initial list of stimulus words in the radio television and radio communication occupation respectively.

    b. There were visual differences in the number and organization of higher order concepts between the flexible and inflexible workers in the two occupations. The flexible worker structures were more balanced, integraded and provided greater differentiation among concepts than did the inflexible workers. In both occupations there were four hierarchical levels of most inclusive to less inclusive factor concepts for flexible workers as compared to inflexible workers. Also, the flexible workers had the largest number of integrated factors within their structure and the fewest number of "isolated" factors. While inflexible workers in the two occupations had less integrated factors and the greatest number of isolated factors.

In conclusion, the results of the two studies suggest that the free association procedure is quite reliable and capable of producing conceptual association structures for a technical field showing differences between groups of workers

who are known to be performing their job at different levels of proficiency. In addition the prodedure is capable of producing a conceptual map of technical content which has face and content validity for experts in the field which may therefore be useful as a way in which to develop curriculum and teach technical content "whole to part" in either individualized or group instruction. This would provide learners with a visual picture of the relationship among technical concepts. The conceptual associative maps provide the learner with a graphic "picture" of the "whole" of the learning task while simultaneously showing them the psychological relationship of the various parts to the whole.

Review of Other Research

Since the original studies of (Smith 1968); (Pratzner 1969), other vocational educators and one person in the military have also used the free association technique to identify the structure of knowledge in technical fields and to use the conceptual associative maps to develop curriculum.

Associative Conceptual Structure

Nee (1977) used the free association technique to identify the hierarchical conceptual structure of statistical quality control personnel. He identified 13 practicing quality control workers and asked them to respond to 122 technical terms used in the field of statistical quality control. Using the relatedness coefficient (RC), developed by Garskof and Houston (1963) and following the procedure of pooling response distributions used by Smith and Pratzner, he was able to generate a hierarchical conceptual associative structure consisting of three hierarchical level of factors, the highest most inclusive factor consisted of 30 lower order factors. Two factors were identified at the second higher order level and four factors at first lowest order. The factor structure showed which stimulus words loaded .300 or above on each higher and lower order factor. This provided the basis on which a curriculum for quality control personnel was developed.

Liu (1972) attempted to replicate the study conducted by Pratzner (1969) in the area of mobile radio communications students in a post-secondary vocational program in Minnesota. He identified four high performing and low performing students who were identified by the instructor of the program based on classroom performance. Also, on the basis of a test of creativity, four high creativity students and four low creativity students were identified as a second group of students. Each group of students were asked to respond to eighty-two different stimulus words with as many technical responses as they could provide in a one minute time period. Response distributions for the higher and lower performers and the high and low creative groups were pooled respectively and subjected to a higher order factor analysis procedure in the same way as Smith (1968) and Pratzner (1969). The results of the test-retest procedure for the high and low performers on the 10 randomly selected words was an average median value of .729 for each group respectively and an average median value of .869 and .783 for the high and low creative students respectively.

High performing students generated a hierarchical structure with one all inclusive concept and three other lower level factors. Low performing students produced a

conceptual structure consisting of only two third order concepts (A and B) and 4 and 3 second order factors respectively. It was not a balanced structure and also consisted of many isolated factors which were unrelated to any of the higher order factors. The results for the high and low creative groups of students was almost identical to the high and low performing group.

The general conclusion was that the free association procedure is (a) equally reliable for high and low performing students and for high and low creative students and (b) that the technical conceptual structure for the high performing and highly creative students is technically and structurally different than the low performing and low creative students.

This study demonstrates that the technique can be used for assessing technical structure for students as well as adult workers.

Another study (Ammerman, 1970), while working for Human Resources Research Organization combined the free association methodology with some of the aspects of the functional context method of instruction (Schoemaker, 1960). He identified two groups of twelve students and eight instructors in a 29 week military radar maintenance course to participate in the study. He followed the same free association procedures as those used by Smith (1968) and Pratzner (1969). The three groups of students and instructors were enrolled in three different types of instructional delivery programs.

N = 4	C_1	students in 16th week of conventional electronics course
N = 4	C_2	matched students in conventional electronic course
N = 4	M_3	students in 10th week of functional context radio maintenance course
N = 4	I_1	instructor in convential electronics course - no experience
N = 4	I_2	instructor in conventional course average of 7.5 years of field experience

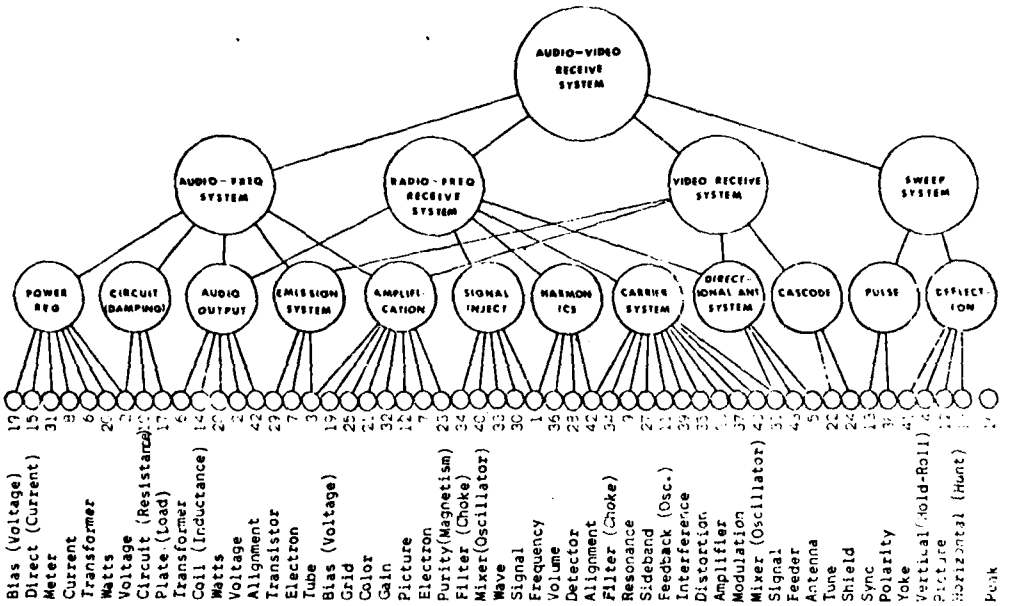
Subjects were asked to respond to only 68 of the most relevant terms used in the Smith study (1960) pertaining to the radio and television repair occupation.

The results show that (a) the students and instructors were able to respond to the stimulus words in a meaningful way, (b) instructors had a larger more extensive technical vocabulary than students, (c) only two levels were identified when student and teacher response distributions were combined and factors analyzed. Because only sixty-eight stimulus words were used and because these were not necessarily radio maintenance terms exclusively, the factor structure was not very meaningful to the students or participating instructors.

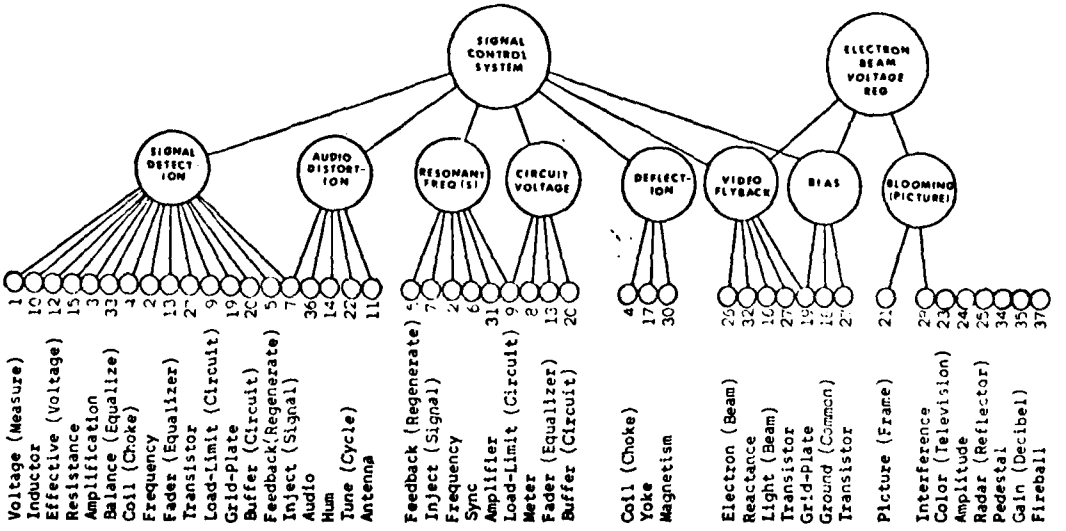
The general conclusion was that the methodology seems to be appropriate for military instructors and students but that a more relevant, comprehensive list of technical terms needs to be generated to determine which more meaningful technical association structure could be generated.

410

THE LABELED TECHNICAL CONCEPTUAL STRUCTURE OF STIMULUS CONCEPTS FOR A GROUP OF *FLEXIBLE* RADIO AND TELEVISION REPAIRMEN



THE LABELED TECHNICAL CONCEPTUAL STRUCTURE OF STIMULUS CONCEPTS FOR A GROUP OF *INFLEXIBLE* RADIO AND TELEVISION REPAIRMEN



(9.1)

FIGURE IX

THE LABELED TECHNICAL CONCEPTUAL STRUCTURE OF STIMULUS CONCEPTS FOR A GROUP OF FLEXIBLE RADIO COMMUNICATIONS REPAIRMEN

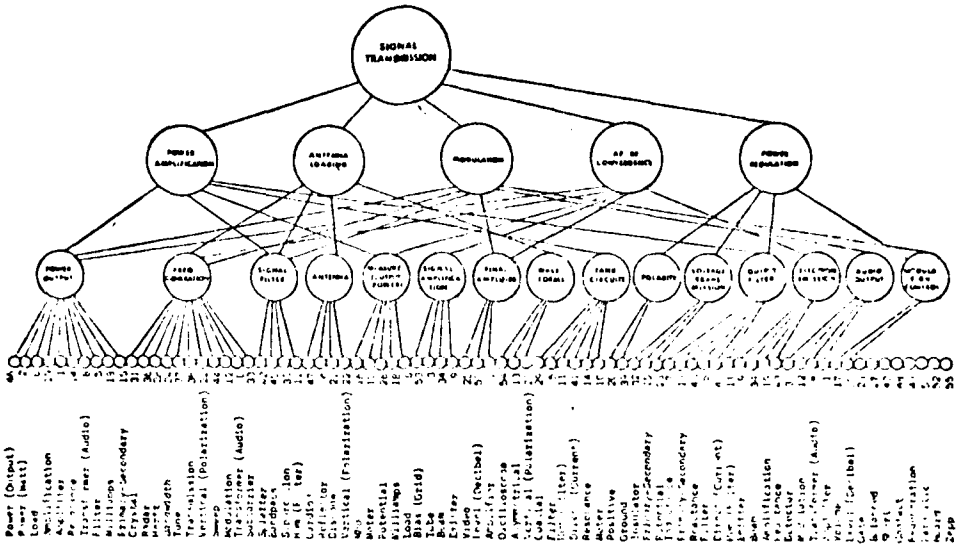
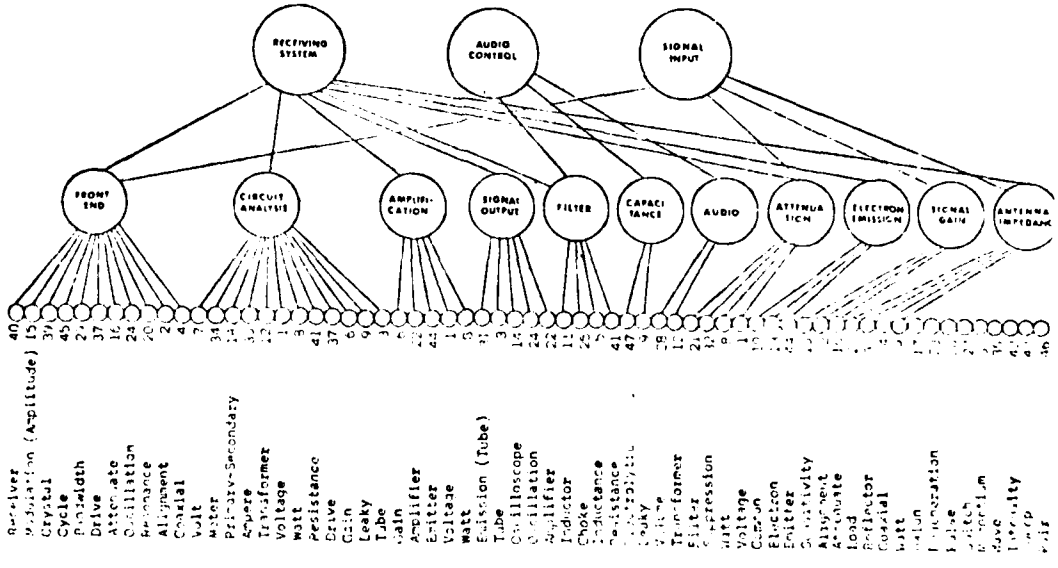


FIGURE X

THE LABELED TECHNICAL CONCEPTUAL STRUCTURE OF STIMULUS CONCEPTS FOR A GROUP OF INFLEXIBLE RADIO COMMUNICATIONS REPAIRMEN



(9/2)

Appendix C

DEFINITION OF (RC) RELATEDNESS COEFFICIENT

The Relatedness Coefficient (RC) is defined as the amount of observed overlap to the maximum possible overlap between two rank order, free association response distributions.

$$RC = \frac{\bar{A} \cdot \bar{B}}{(A \cdot B) - 1}$$

The product (A • B) represents the intersect which would be obtained if every element in A appeared in B (and B in A) and in the same order. A and B are operationally defined by the product sum of the longest rank ordered response distribution of the two stimulus words. Where the two stimulus words are different, a minus one (-1) is used in the denominator to represent an unbiased estimate of maximum overlap. The product $\bar{A} \cdot \bar{B}$ is the product sum of the rank orders of response commonalities observed between two response distributions.

Example

COMPUTATION OF A SAMPLE RC FOR THE STIMULUS WORD YOKE AND DEFLECTION

Associates to "Yoke"	Frequency	Rank	Associates to "Deflection"	Frequency	Rank
Yoke	4	6	Deflection	4	6
Deflection	4	5	Yoke	3	5
Horizontal	3	4	Horizontal	2	4
Vertical	3	3	Vertical	2	3
Ringing	2	2	Circuit	2	2
Trapezoidal	2	1			

A = Rank order of responses to "Yoke". (6, 5, 4, 3, 2, 1)
B = Rank order of responses to "Deflection". (6, 5, 4, 3, 2, 1)
C = Common Responses. (Yoke, Deflection, Horizontal, Vertical)
$\bar{A}$ = (6, 5, 4, 3)
$\bar{B}$ = (5, 6, 4, 3)

$$RC = \frac{\bar{A} \cdot \bar{B}}{(A \cdot B) - 1} = \frac{(6, 5, 4, 3) \cdot (5, 6, 4, 3)}{(6, 5, 4, 3, 2, 1) \cdot (6, 5, 4, 3, 2, 1) - 1} = \frac{85}{(91 - 1)} = .904$$

Selected Bibliography

Ammerman, Harry L., "Systematic Approaches for Identifying and Organizing Content for Training Programs," Professional Paper, _____ Human Resources Research Organization, Alexandria, Virginia, 1970.

Ausubel, David P., The Psychology of Meaningful Verbal Learning, Grune & Stratton, New York, 1963.

Bruner, Jerome S., The Process of Education, Harvard University Press, Cambridge, Massachusetts, 1963.

Deese, James, "On the Structure of Associative Meaning", Psychological Review 69: 161-75, 1962.

Deese, James, The Structure of Associ ions in Language and Thought, Baltimore: The Johns Hopkins Press, 1965.

Garshof, B. D. and Houston, J. P., "Measurement of Verbal Relatedness: An Ideographic Approach," Psychological Review 70: 377-88, 1963.

Johnson, P. E., "Associative Meaning of Concepts in Physics," Journal of Educational Psychology 55: 84-88, 1964.

Johnson, P. E., "Some Psychological Aspects of Subject Matter Structures," Journal of Educational Psychology, 56: 217-224, 1965.

Liu, Cheng C., "Testing an Empirical Procedure for Identifying Technical Associative Structures, Discriminating Between Students Grouped by Performance and by Creative Thinking in a Mobile Communication Technology Program," Unpublished Doctoral Thesis, Minneapolis, Minnesota, University of Minnesota, 1972.

Nee, John G., "The Development of a Technical Conceptual Structure for the Concepts Possessed by Selected Quality Control Specialist," Dept. of Industrial Education Technology, Central Michigan University, Mount Pleasant, Michigan, 1977.

Pratzner, F. C. "Testing an Emperical Procedure for Identifying Technical Associative Conceptual Structure: Discriminating Between Workers Within and Between Two Occupations," Unpublished Doctoral Thesis, Minneapolis, Minnesota, University of Minnesota, 1969.

Schaub, Joseph J., "Structure of the Disciplines: Meaning and Significances, (In) The Structure of Knowledge and the Curriculum. Rand McNally Curriculum Series, Chicago.

Smith, B. S. and Moss, J. (Ed), "Report of a Seminar: Process and Techiques of Vocational Curriculum Development," Minneapolis, Minnesota, Minnesota Research Coordinating Unit for Vocational Education, 1970.

Smith, B. S., "Testing an Emperical Procedure for Identifying Technical Associative Conceptual Structure: Discriminating Between Flexible and Inflexible Radio and Television Repairmen," Unpublished Doctoral Thesis: Minneapolis, Minnesota, University of Minnesota, 1968.

(914)

THEW, Sgt. C., Michael, Air Force Human Resources Laboratory, San Antonio, Texas.

CODAP: AN OVERVIEW OF THE TASK FACTOR TECHNOLOGY (Tue P.M.)

The comprehensive Occupation Data Analysis Program (CODAP) system is a basic tool in the area of occupational analysis. Within this area, task factors have become an essential element in describing the attributes of jobs. The Air Force Human Resources Laboratory has developed a stand-alone Task Factor Package which may also be used with the standard CODAP system. In one project, a list of simulator attributes was presented to a panel of raters evaluating proposed simulation equipment. The use of this package permitted an objective measure of inter-rater reliability which demonstrated a strong consenses of opinion. Another application, related to CODAP, might be to use the existing equipment list, rated for learning difficulty, process the data, and report the results independently from a conventional CODAP analysis. Because the task factor package can be treated as s subsystem of CODAP, any reports generated may be saved and integrated into a CODAP final report. This system still contains the ability to access traditional task factor data and job descriptions.

TRATTNER, Marvin H., U.S. Office of Personnel Management, Washington, D.C.

A SURVEY OF CODAP APPLICATIONS IN FEDERAL CIVILIAN OCCUPATIONS
(Wed A.M.)

The bulk of this paper is devoted to a description of ten instances of the use of CODAP in the Federal civil service. These include almost all of the known applications of CODAP in the Federal sector.

For each application, the agency, occupation(s) and specific use is described. Approximately one-half of the applications deal with training and the others with a variety of products. Even the training applications vary in purpose and only one-half of these involve directly the development of new training courses. Only one of the projects involves CODAP use for more than one purpose.

If CODAP is seen as applicable for many personnel functions why haven't there been more multi-purpose uses? An analysis of the CODAP applications reveals that CODAP is most often employed to meet a new requirement imposed on a personnel activity. CODAP output will usually not be accepted by the bureaucracy if that acceptance necessitates the modification of existing procedures. A means to overcome this inertia is tentatively advanced.

A SURVEY OF CODAP APPLICATIONS WITH FEDERAL CIVILIAN OCCUPATIONS

Marvin H. Trattner
U.S. Office of Personnel Management
Washington, D.C. 20415

The CODAP programs are beginning to have widespread use in Federal civilian personnel and training agencies. Since there is no medium for CODAP information exchange within the Federal civilian service this paper was written to provide a description of these applications. The focus of the paper will be on the specific use to which CODAP is put by the agencies. All known CODAP studies with Federal civilians that have occurred within the last five years are described.

The subject should be of interest to individuals in the military personnel and training agencies who follow CODAP developments since civilian CODAP uses can be expected to vary from military due to the differences in function and legal requirements for the two systems.

This paper will be divided into two parts, first I will describe each use to which CODAP has been put. Then in the second part I will discuss the issues encountered in performing the CODAP studies and utilizing the results. This paper will only deal with Federal civilian government studies. There are a growing number of state and local governments that are using CODAP but these will not be described here.

Description of CODAP Studies

Method

Twelve CODAP studies were found in ten government agencies. These yielded eleven unique uses for CODAP.

In all instances I spoke to the individual who was responsible for doing the study. In about one-half the cases that individual was not the ultimate user of the occupational information. In some of these instances the researcher could provide only a general description of the intended use. Only 60% of the studies had been completed when I interviewed the researcher.

In the interview the focus was on the use to which the CODAP data was to be put. And only the use which was spontaneously given was recorded. When uses were mentioned only as posibilities, which had not yet been fully agreed to within the agency, they are not described in this report.

Results

The table which follows this paper describes the CODAP study by agency, occupation and use or product.

Discussion of Findings

CODAP Uses

About one-half the uses are related to the training function. CODAP is used in training applications either to develop or to determine the appropriateness of existing training programs.

Two-thirds of the applications are either directly or indirectly related to Uniform Guidelines for Employee Selection Procedures requirements that agencies have evidence for the validity of their selection procedures. Federal civilian employers are explicitly required to demonstrate the validity of their selection instruments under the Equal Employment Opportunity Act of 1972.

Problems in CODAP Implementation

a)Bureaucratic resistance. Eleven of the 12 CODAP studies described resulted from a new requirement imposed on an agency. There was only one instance in which CODAP was applied in an existing program. A related fact is the relative scarcity of multiple uses for any single study. Our data show eleven unique uses for CODAP but only three instances in which CODAP was used for two different products in the same study.

One reason for the above appears to be that CODAP, probably like any other innovation, is viewed by the entrenched bureaucracy as a threat to its status. For instance, if a training division changes its procedure and starts developing courses based on CODAP data, it is somehow considered by some to be a negative statement with respect to its previous course development work. Management becomes defensive and CODAP is evaluated not with respect to how it can improve the agency product but rather with what it cannot accomplish. When these criteria are applied CODAP cannot be fully utilized or fairly evaluated.

b)Logistic problems. It is clearly inefficient to continue to perform the Federal studies in the same manner that these twelve were performed. For each study a different data collection form was designed and printed. In some cases these were optically scanned forms. Also each researcher had to find a computer which had or could easily accomodate the CODAP system and where he could run the data.

(920)

c)Educational problems. Many of the researchers knew little about the total CODAP system. They were unaware of the potential of the entire system. In many cases they did not understand all of the output they received. This lead to a very inefficient product and in some few cases to one that had little utility.

Nevertheless about one-half the researchers indicated that their agencies intended to do additional CODAP studies. If CODAP studies are to continue the above two problems point to the need for a centralized CODAP unit devoted exclusively to CODAP work.

Summary Statement

Twelve CODAP studies in Federal civilian agencies are described with the emphasis on the intended use. Unique applications, especially ones designed to meet legal selection instrument validation requirements imposed on civilian agencies are described. Some studies however have given rise to bureaucratic turfophobia, a disease common to individuals who mistakenly believe their turf is being invaded. Nevertheless one-half of the researchers indicated a need to continue CODAP work. Reasons were advanced for the creation of a central CODAP unit to accomplish this.

(923)

CODAP USES IN FEDERAL CIVILIAN AGENCIES

Agency	Occupation(s)	Use
Immigration & Naturalization Service	Border Patrol Officer	New selection procedure developed in which trainees are selected only after successful completion of training. New procedure developed to increase number of minority group selectees. CODAP used to develop training program and to defend selection procedure against legal challenge.
FBI	Special Agent	
Customs Service	Customs Patrol Officer	
Dept. of Interior-Office of Surface Mining	Mine Safety Inspector	Develop training program for the occupation. Incumbents regulate surface mining which is a new Federal function.
Dept. of Interior-Bureau of Land Management	Range Conservation Specialist	Determine degree of appropriateness of present training course(s).
Federal Acquisition Institute	Contract & Procurement Industrial Specialist Purchasing	
Dept. of Navy-Civilian Personnel	Financial Management Career Field	Construct individual training plans for career development program.

Agency	Occupation(s)	Use
Office of Personnel Management Federal Acquisition Institute	Personnel Clerk Dental Hygienist Contract and Procurement Industrial Specialist Purchasing	Aid in development of classification standards. i.e., used to identify subspecialties, prevalance of subspecialties, documents standards. Aid in development of qualifications standards.
U.S. Park Police	Park Police Officer	Construct job information test employed in promotion system. Develop performance rating form.
Dept. of Army-Civilian Personnel	Personnel Career Field	Develop promotion appraisal form for a centralized personnel referral system.
Office of Personnel Management (with State of Utah and several Utah municipalities)	Clerical occupations	Calculate degree of similarity among Federal, state and municipal clerical occupations. Aim is to justify use of Federal selection tests for clerical vacancies in all cooperating Utah jurisdictions.
Office of Personnel Management	Claims Examiner Internal Revenue Officer Customs Inspector	Construct and weight job performance measures and to determine homogeneity of the occupations in a test validation project. Can be used to select research participants if occupations are not homogeneous. Customs Service subsequently used task inventory as basis for work measurement form. Social Security Administration used data to develop Claims Examiner training course.

TUBBS John D., HANSEN, Alan D., BRYANT, James A., EVERETT James E. and
DEASON, Paul J., USATRASANA, White Sands Missile Range, New Mexico.

VULCAN TRAINING SUBSYSTEM EFFECTIVENESS ANLAYSIS (TSEA) (Wed A.M.)

USATRASANA at the request of the US Army Air Defense School, Fort Bliss, Texas has for the past four years been conducting training analyses on fielded air defense systems. The first was REDEYE in which a sample size of 2000 gunners was obtained. The second, to be discussed here was VULCAN, the Army's air defense Gattling gun. The third, CHAPARRAL, is currently under way.

The VULCAN is a a high rate of fire, 6 barrel, 20mm gun with explosive rounds. It is manually controlled with a range radar which allows a computer calculated, automatically inserted lead angle. The gunner must be highly proficient in manually acquiring the target, tracking smoothly, and utilization of the radar directed lead angle. In addition, he must be proficient in boresighting the weapon, visual aircraft recognition, and other related tasks. This study evaluates 891 gunners from units all over the world and points out pertinent problem areas and their solution.

TRAINING SUBSYSTEM EFFECTIVENESS ANALYSIS VULCAN AIR DEFENSE SYSTEM (VADS)

John D. Tubbs
US Army TRADOC Systems Analysis Activity
White Sands Missile Range, New Mexico 88002

1. INTRODUCTION

a. Purpose. The purpose of this Training Subsystem Effectiveness Analysis (TSEA) is to support the United States Army Air Defense School (USAADS) directed program of training evaluation and improvement. The VULCAN Air Defense System (VADS) was evaluated to determine the functional relationship between training and combat effectiveness and to determine the training programs required to optimize this relationship.

b. Background. The WARSAW Pact Air Threat to the US Field Army is highly sophisticated, highly responsive, and massive. The attack force contains a large quantity of low altitude attack helicopters which pose a threat to our front line forces, especially armored units. The VULCAN gun system with the Target Alert Data Display System (TADDS) and the follow-on DIVAD gun offer a potential counter to this threat. VULCAN crew training, both quantity and quality, can substantially impact the effectiveness of the weapon system. This report identifies the critical training factors and relates them to crew proficiency. The analysis also establishes the level of proficiency of VULCAN crewmen in specific tasks at the completion of Advanced Individual Training (AIT) and in active Army units.

c. Problem. The Army has constrained resources for the conduct of individual and unit training and therefore must make optimum use of the resources available to build and maintain Army combat effectiveness. The VADS has been selected by USAADS as a part of the total program to upgrade training while maintaining or reducing current training costs in both the institution and the units. The results of this study, coupled with the previously conducted VULCAN WSTEA, will contribute to the existing data base from which sound decisions can be made regarding training and training resources.

d. Impact of the Problem. The VADS provides air defense at the maneuver unit level. Therefore, the proficiency of the VULCAN gunner directly affects the survival as well as the performance of the unit. It is imperative that VULCAN gunner proficiency levels be defined and evaluated to allow a determination of required proficiency. Training can then be adjusted to achieve this level. This level must be defined to assure that gunners are neither under- nor over-trained. VULCAN gunner proficiency must be maintained at a level sufficient to assure survival of Army assets during a future conflict.

2. OBJECTIVES. The objectives of this study are:

a. To determine if the achieved effectiveness (E_A) is equal to or greater than the design effectiveness (E_D) of the weapon system ($E_A \geq E_D$).

b. To determine which training factors contribute most to proficiency.

c. To analyze various training programs utilized by the school and field air defense units, and determine the most effective programs in terms of both cost and training effectiveness for each.

d. To establish fundamental requirements for ADA personnel as related to technical skills necessary to be a proficient VULCAN gunner.

e. To determine if VULCAN crewmen and REDEYE/STINGER skills are transferable by comparison of course material.

f. To determine the training impact on the DIVAD gun system.

g. To recommend modifications to the VULCAN weapon system as seemed appropriate.

h. To determine ammunition requirements for required/desired training proficiency.

i. Determine a means by which potential Air Defense Artillery (ADA) gunners may be screened to afford the ADA community the best qualified candidates.

3. VULCAN CONCEPT OF OPERATION

a. General. VULCAN is a short range, 6-barrel Gatling, 20mm Air Defense Gun with a high rate of fire.

b. Description. VULCAN is designed to protect forward area combat elements and other critical assets from attack by hostile aircraft operating at low attitudes. VULCANs are also provided to rear area commands to protect air bases, ammunition supply points, and other critical assets. Both the towed self-propelled VULCAN use the same weapon system but have different carriers. The VULCAN system has a maximum rate of fire of 3000 rounds per minute and an effective range of 1200 meters against aerial targets.

c. Engagement. The successful engagement of aircraft by VULCAN requires accomplishment of an ordered sequence of tasks: Detection and Identification, Acquisition and Track, Radar Utilization, and Firing. The proficiency with which these tasks are accomplished depends primarily on the training of the VULCAN crew.

(1) Detection and Identification. To enhance visual detection and identification each VULCAN squad is authorized a TADDS. The battery-operated TADDS is a light weight frequency-modulated receiver used to obtain warning, location, and tentative identification of aerial targets detected by the Forward Area Alerting Radar (FAAR) location in the Battalion Headquarters area.

(2) Acquisition and Tracking. The VULCAN gunner must acquire the previously detected and identified target visually within the reticle of the M61 lead-computing sight. The gunner should acquire the target from the rear, when possible, and establish a smooth track by maintaining the target within the inner (15 mil) reticle pattern. He must track within the 60 mil reticle pattern in order to obtain radar lock.

(928)

(3) Firing. The VULCAN system can be operated in any of four modes: Radar, Manual, External, or Ground. These modes are selected by the gunner who sets the mode switch on his control assembly to the desired operating mode. The radar, manual, and external modes are used against aerial targets. In each of these modes the M61 lead-computing sight sets into the system the predicted lead angle and super elevation necessary to achieve a hit on the target. Two of the modes are discussed below.

(a) Radar Mode. The radar mode provides the most accuracy against aerial targets and is normally used unless the range-only radar is inoperable or cannot be used for tactical reasons. In this mode the target speed and range data are continually being furnished the lead computing sight after the radar lock has occurred. When the gunner has acquired and smoothly tracked the target for approximately two seconds he can radiate with the system radar by depressing the foot switch. When radar locks on to the target, the range and range rate and the turret angular rate are processed by the sight current generator to set the proper lead angle and super elevation into the gun system. During the acquisition time delay (ATD), the lead angle and super elevation are inserted and the READY-TO-FIRE (RTF) light in the sight is lighted. The gunner may fire as soon as the RTF light is lit by squeezing the switch on the turret control hand grip. He is to continue tracking smoothly during firing. The weapon will fire a pre-determined burst duration selected by the gunner, i.e., 10, 30, 60, or 100 rounds/bursts.

(b) Manual mode. In the manual mode, the estimated target range and speed is set into the system by the gunner using switches on the control assembly. This mode is less accurate than the radar mode and is selected when the radar mode cannot be used. In this mode the RTF light is lighted at all times and the gunner fires the weapon after attaining a smooth track and the target is within the range of the gun.

4. MEASURES OF TRAINING EFFECTIVENESS

a. Written Test - The individual scores of 25 questions covering basic knowledge of the VULCAN system hardware. The same test was administered to both AIT students and unit crewmen.

b. Visual Aircraft Recognition (VACR) Skill - The individual scores on the identification of 30 aircraft slides from the Ground Observer Aircraft Recognition (GOAR) kit. This test was administered to all unit crewmen.

c. Questionnaires - These were administered to both AIT and unit crewmen as a survey which covered the background and general attitudes of the individuals. The questionnaire for AIT was structured to also obtain data on the individual assessment of adequacy of training in specific tasks. The questionnaire for the units was modified to obtain data on the amount of task-related training received on a scheduled basis.

d. Operations Test - Hands-on testing of specific critical tasks for unit crewman. The tasks were assigned and observed by a member of the test team. The evaluation was based on the requirements and procedures described in the current operator's manual.

e. Engagement Skills - An evaluation of the gunner's skill in acquisition, acquisition, tracking, radar utilization, and live fire engagement procedures as recorded by video monitoring and recording system. This evaluation was conducted by the test team and assessed the gunner's ability to acquire, track, and properly apply approved procedures to engage a live target. It was specified that the engagement be conducted in radar mode. The 7.62mm minigun was used in firing against both the Radio Controlled Miniature Aerial Target (RCMAT) and the towed sleeve target.

f. Psycho-Motor Skills - Additional testing with a hand/eye coordination device to measure psycho-motor skills applicable to crew skills.

5. RESULTS

a. Tracking. It has been known for many years that VULCAN gunners have great difficulty in maintaining track through lead angle insertion. This fact was brought to the study team's attention early in the study with a request to examine the problem and recommend a solution.

(1) The VULCAN system utilizes a range-only-radar (ROR) to measure target range and rate, and a rate gyro within the turret to measure angular rate. The critical factor in the lead angle calculation is the gunner's ability to smoothly track prior to radar radiate. This smooth track allows the rate gyros to pick up the turret rotational rate which, when coupled with range and range rate allows lead angle computation. The actual sequence of events is as follows:

(a) Referring to figure 1 the gunner must sweep over the target and approach the target from the rear. This allowed the rate gyros to begin picking up the angular rate in the correct direction. The large reticle (60 mil) represents the radar acquisition zone and the inner reticle (15 mil) the tracking zone.

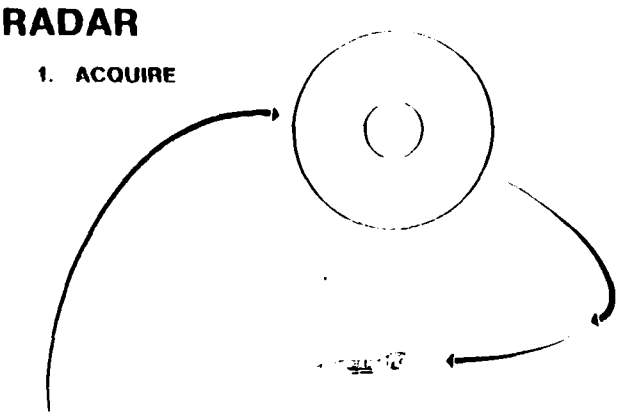


Figure 1. Target Acquisition Procedure

930

(b) The gunner must then track smoothly for a minimum of one to two seconds to allow the gyros to pick up the correct angular rates. This is seen in figure 2.

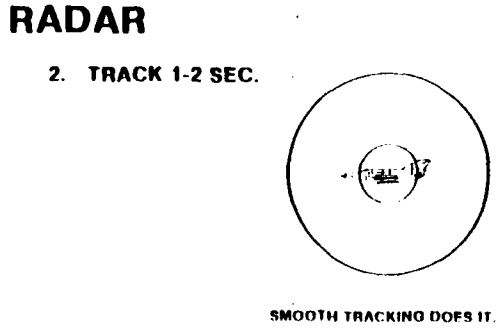


Figure 2. Target Tracking Procedure

(c) After this is accomplished the gunner depresses the radar radiate foot switch which begins the lead angle calculation or ATD phase. Once the radar locks on the target, the range, range rate, azimuth, and elevation rotational rates are fed to the lead angle computer or sight current generator (SCG). At this point the gain on the azimuth and elevation drive motor is increased by a factor of five so that as the barrels are moving to the lead angle position, the sight reticle is driven in the opposite direction at the same rate such that it remains over the target. Since the sight is gyro stabilized, the inherent lag causes the sight to momentarily move behind the target. This is shown in figure 3.

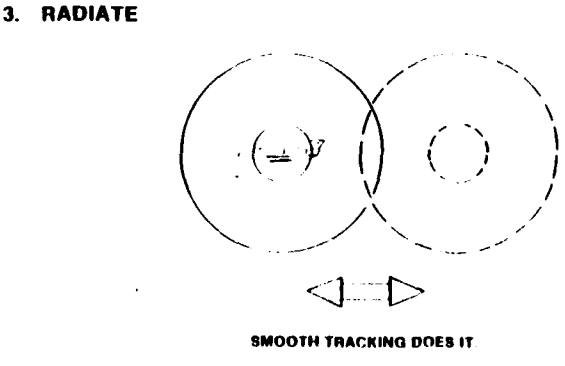


Figure 3. Target Tracking During ATD

The sight reticle will, if left alone, automatically move back over the target as shown in figure 3 within several seconds if the gunner maintained smooth track prior to radiate. The problem observed in the gunner maintaining track through lead angle insertion is related to training rather than hardware.

(2) Since VULCAN was acquired into the air defense force the gunners have been taught both in AIT and units to rotate the gun controls in such manner to move the sight reticle back over the target (see figure 3). In doing this, two problems result:

(a) First the gunner is, through the gun controls, connected to the barrels with a gain of five. Therefore, the barrel motion he has been accustomed to is now amplified by five. Due to this, when he moves the controls the barrels respond in a manner five times the rate that he had prior to lead angle insertion.

(b) Second, the lead angle compute assumes the sight motion to be an increase in target velocity and calculates a new lead angle. The barrels again move forward and the sight falls behind the target putting the gunner back to the same position.

(3) The instruction supplied by General Electric (GE), the system developer, clearly states as do the VULCAN manuals that the momentary sight displacement should not be corrected because this action will result in incorrect lead angles. Referring to figure 4 the fourth bullet clearly states that the tracking rate must not be changed during lead angle insertion. This is further clarified within the enclosed box; however, the two center diagrams are incorrect. The reticle should be shown behind the target and not in front. Consequently, the gunners are incorrectly trained. The Field Manual 44-5 instructions are correct but the diagram is in error. All in all this serves to confuse the gunner, cause him to dislike and distrust the radar mode, and more importantly results in poor gunnery.

(4) It must be stressed that in order to consistently hit targets the radar mode must be utilized. While gunners are able to hit towed banners which are towed in a very consistent pattern and speed (usually very slow, 140 to 250 knots), they will not be able to hit tactical aircraft flying at a wide range of speeds within the VULCAN engagement capabilities. In actuality the hits measured on towed banners are almost always caused by the gunner quickly learning "Kentucky Windage" and leading the target utilizing the manual mode even though he is told to be in the radar mode. This was observed at the firing range at Todendorf, Germany. The gunner in the manual mode must be able to tell, for example, the difference between 175 and 250 knots aircraft speed. Most everyone is incapable of doing this consistently. Therefore, to hit the target consistently the gunner must not only utilize the radar mode, but be correctly trained to use this mode.

b. Visual Aircraft Recognition (VACR).

(1) A VACR test was administered to all units visited during this study. The VACR test consisted of identifying 30 aircraft slides extracted from the GOAR library. Twenty-six of these 30 slides were required for SQT qualification. Only the 26 SQT slides were used as the basis for grading. It should be noted that in Germany all units are required to identify 50-60 aircraft. In OCONUS the mean VACR score was 65.4%, while in CONUS the mean score was 56.0%. SQT requires 90%.

The gunner:

- Traverses mount to acquire target in center of the sight reticle and brings the sight onto the target from the rear.
- Establishes smooth tracking (that is, he establishes a tracking rate that keeps the target centered in the sight reticle; maintains that rate; and adjusts it gradually as required to keep the target centered in the sight reticle).
- Presses foot switch (causing radar to radiate).
- When radar locks on target, maintains a constant tracking rate during the 2-second period that the lead angle is being inserted into the sighting system. The tracking rate must not be changed during lead angle insertion.

The gunner must be prepared for lead-angle insertion as, ***temporarily***, the sight reticle moves away from the target. Any attempt to move the sight reticle back over the target during this period results in incorrect lead angles. Here is what the gunner might see in his sight during lead angle insertion:

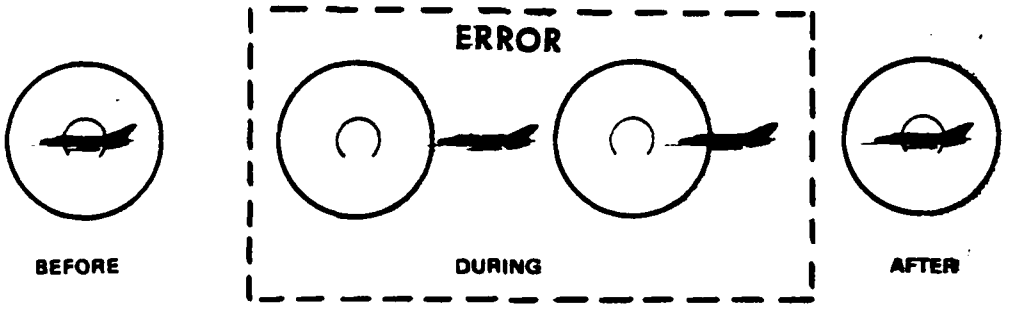


Figure 4. Training Instruction Error

(2) The results indicate the following:

(a) Higher VACR scores were obtained by individuals who had an excess of 12 months within one or more of the following:

- time in battery/battalion
- time in PMOS and/or grade
- time in crew

(b) A rank ordering of VACR performance was directly related to duty position and pay grade. This relationship is depicted below. The scores are weighted mean VACR score percentages and represent all crewmen CONUS and OCONUS.

1.	E6	73.9%
2.	E5	71.1%
3.	Squad leader	68.7%
4.	E4	68.7%
5.	Senior gunner	58.7%
6.	E3	52.2%
7.	Driver	49.3%
8.	E1, E2	45.6%
9.	Assistant gunner	43.5%

(c) VACR training if administered above squad level appears to produce better VACR scores than if the training was conducted at the squad level. This relationship is applicable to CONUS and OCONUS.

(d) Other relationships which tend to compare directly with the higher VACR scores are: higher education level, married, 26-28 year age range, higher usage of Training Extension Course (TEC) tapes, those who state they actively engage in ARTEP tasks and those who anticipate re-enlistment.

(3) Four of the 810 crewmen tested failed to correctly identify any of the aircraft in the VACR test. This was due to either a lack of ability or of interest.

(4) While the 16R Soldier Manual contains 26 aircraft, the units in Germany are required to identify a minimum of 50 and a maximum of 60 aircraft. These must be known by number or NATO name, e.g., MIG 25 or FOXBAT. Since the mean OCONUS score was 65.4% of 26 (i.e., approximately 17 aircraft correctly recognized) and unit records indicate a test result of 22% of 63 (i.e., approximately 14 aircraft correctly recognized), it is obvious that the average soldier will never be able to master 50 to 60 aircraft considering any rational time allocated for training. This number is unrealistic and should be reduced.

c. Unit Live Fire Results

(1) The live fire portions of this test were conducted utilizing live targets flown at realistic threat velocities. The targets were flown by the gun system five times for each gunner tested. Two of these passes were tracking only and the last three firings in which the gunner was allowed to fire as many bursts as he could get off. All results were recorded on video

tape for evaluation. The subsequent evaluation was conducted by VULCAN experts observing TV monitors independent of each other. Based upon the school standards for each of the three engagement phases (i.e., acquire target, smoothly track target, and activate radar mode), the following results were obtained:

	CONUS	OCONUS
Number of Total Engagements	944	146
Number of Engagements in which:		
Acquisition phase was within acceptable standards	27	0
Tracking phase was within acceptable standards	37	1
Radar procedure phase was within acceptable standards	16	0
All three phases were correct within one engagement	2	0

(2) The above results were correlated with gunner mental category and the results indicated no significance at the .05 level. This basically means that VULCAN gunnery is not related to mental category. This is good news since 75% of the VULCAN gunners reside in mental categories IIIB or less.

(3) The reason for the poor gunners is twofold. First, the gunners have been trained incorrectly in firing the weapon for about 13 years. Second, adequate training equipment is not available to the gunners such that they can be trained to proficiency. As a result of this study both of the above have been corrected with new training equipment currently being acquired.

6. STUDY RECOMMENDATIONS

The recommendations of the VULCAN TSEA are summarized in this paragraph.

a. Institutional Training

(1) USAADS publish a correction to Field Manual 44-5 and assure that AIT instructors correctly teach radar utilization.

(2) USAADS continue to only familiarize 16R personnel in engagement techniques.

(3) The reduction or elimination of VIP shows during AIT live firing time.

(4) Qualification in PM on tracked vehicle, self-propelled vehicle, gama goat, and operate gama goat be required to better prepare the graduate for his first duty assignment.

b. Unit Training

(1) An Ad Hoc Committee that is representative of the air defense community be established to identify the SHORAD Threat. This threat should be limited to approximately 20 aircraft.

(2) VULCAN crewmen be trained in proper radar utilization techniques.

(3) An effective training program be established to fully qualify VULCAN crewmen in the tasks required by the CM. The use of TEC tapes be used in this program and by crewmen for refresher training as required.

(4) Maintenance equipment be made available to the batteries in accordance with existing TOEs. The acquisition of MWM-3 test sets should be expedited.

(5) A school should be established to train RCMAT pilots on a scheduled basis. Recognizing the difficulty in establishing a MOS for RCMAT pilots USAADS should consider this as an approach to solving the problem of pilots. Units (TASO, battalion, or division) should dedicate their RCMAT pilots to flying in support of ADA training exercises.

(6) All-weather tracking trains be acquired for each division to enhance tracking training for VULCAN gunners.

c. VULCAN Training System Evaluation

(1) The VTS be purchased for both the school and units for VULCAN training.

(2) A VTS training package be developed for the school and units.

936

VALE, David C., Assessment System Corporation, St. Paul, Minnesota.

DESIGN OF A TEST DEVELOPMENT SUBSYSTEM (Thu P.M.)

The test development component of a computerized testing system is responsible for maintaining a bank of test items and for assembling subsets of the bank's items into tests of specific forms. The bank maintenance portion of this subsystem must provide a means of entering items into the system, storing and retrieving the items, modifying the items, and keeping records on when the items were used and how well they performed. The test construction portion must provide means to combine these items into a test of a specified structure, to specify a scoring and score conversion algorithm, and to specify acceptable responses and methods for handling unacceptable ones. The system should provide communication of these specifications in such a way that a person skilled in test construction, but not computer programming, can manipulate them efficiently. In this paper the necessary and desirable characteristics of a test development subsystem are detailed and several alternative system design possibilities are discussed.

DESIGN OF A TEST DEVELOPMENT SUBSYSTEM

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

A computerized testing system must provide a means by which a test constructor can create the tests that are to be administered. The test development subsystem provides this means and includes three basic capabilities: 1) a means of getting items into the computerized testing system, 2) a means of modifying items already in the testing system, and 3) a means of specifying sets of items for administration as tests or test batteries. The goal in design of such a subsystem is to produce a system which is simple to use and which has sufficient power in each of these areas to assure that tests can be constructed efficiently from the test constructor's point of view. It is relatively easy to design a system that is simple but insufficiently powerful. It is nearly as easy to design a complex system that will do everything. The ideal system, however, one that is simple and will do everything, is much more difficult to design.

In the sections that follow, each of the three capabilities outlined above will be discussed. In some areas, the conflict between simplicity and power will be more striking than in others. In most areas, the optimal design choice will depend on such things as the types of tests and items to be administered as well as the capabilities of the individuals constructing the tests. This paper will not attempt to present the optimal design in any case but will, instead, point out alternatives and suggest which alternatives might be most appropriate in a number of situations.

Entering The Items

The first step in the process of developing a computerized test is usually the entry of test items into the testing system. During this process, the words or symbols comprising each of the items must be formatted and identified as a unique entity within the testing system. Each item record contains, at a minimum, two parts: the item text and item characteristic information. The item text is that part which will actually be presented to the examinee. The item characteristic section contains identifying information such as the item number, psychometric information such as item parameters, and logical information such as the number of lines and characters contained in the item and the types of responses that are acceptable. (See Vale and Brown, 1977, for a list of information stored by an operational item bank.) All of the item information must be entered virtually without error.

Item entry usually occurs in one of three situations: 1) many items from several tests are banked for later reorganization into several new computer-administered tests; 2) all the items for a single test are entered in preparation for computer administration of that test; or 3) a few new items are entered into the system as they are written or otherwise procured to allow trial administration and norming. The three situations differ in the amount of input required, in the speed with which entry must be completed, and in the types of individuals who will do the entry. In the first situation, a data-entry professional would probably be used and the time available for entry would probably be great. In the latter two instances, the new tests might be due on very short notice (i.e., yesterday) and test construction personnel might be used for data-entry operations. Different entry systems are indicated for these two sets of situations.

Computerized testing allows the possibility of several new item types, including items with spoken questions and responses and items composed of response-alterable video sequences. Although these types are intriguing, most current interest in computerized testing is limited to three more conventional types of items; textual items, graphic items, and pictorial items. Discussion here will be limited to these three item types.

Textual items. A textual item is an item comprised entirely of numbers, letters, and characters typically found on a standard typewriter or computer terminal. There are two basic ways in which textual items may be entered. One is in an unprocessed, line-by-line mode. Examples of the unprocessed mode are keypunch, keydisc, and OCR (Optical Character Recognition) entry. Keypunching is accomplished with a keypunch machine, keydisc is accomplished via a CRT terminal entering data directly to a computer disc, and OCR entry is accomplished by typing the data using a typewriter with an OCR element.

Unprocessed entry is most useful for the situation where many items are to be entered at once, as when an existing paper-and-pencil testing program is computerized. Unprocessed entry is useful, in that case, because entry can begin before the testing system is completed and this will accelerate the ultimate implementation of the system. In this situation, the mass of material to be entered warrants the hiring of a data-entry professional. The familiarity of such an individual with standard data-entry procedures makes such an unsupervised entry mode acceptable. The data-entry professional is aware of the need to adhere to specified input formats and is sensitive to accidental deviations from such formats. This, combined with standard verification techniques, can assure reasonably error-free input.

Unprocessed entry is not the ideal entry system for an operational testing system, however. The last two data entry situations listed above are probably more typical of an operational environment. Items are likely to be entered in small batches by test-construction personnel with basic knowledge of how to enter data into a computer but basic ignorance in error detection and verification procedures as well as inherent manual clumsiness. A processed entry system is thus a preferable means of assuring error-free items.

940

Perhaps the most elegant and simple entry system is the CRT screen-oriented entry system. Such a system, as illustrated in Figure 1, allows entry of an entire item onto a CRT screen. All necessary item-descriptive data are entered in prompted fields and are edited to assure that they are internally consistent. The keyed response, for example, is checked to assure that such a response is included in the item. Item parameters (such as the a, b, and c parameters of Latent Trait Theory; Birnbaum, 1968) can be checked to assure that they are within a reasonable range. The correct item format can be enforced by automatically fitting the item into a template. Since all entry is processed, the system can force the operator to recognize and correct errors. Although errors of spelling and punctuation are not likely to be detected by the system, format and some parameter errors will be corrected at entry time. This interactive entry procedure is more suitable for use by individuals who are not data-entry professionals by training.

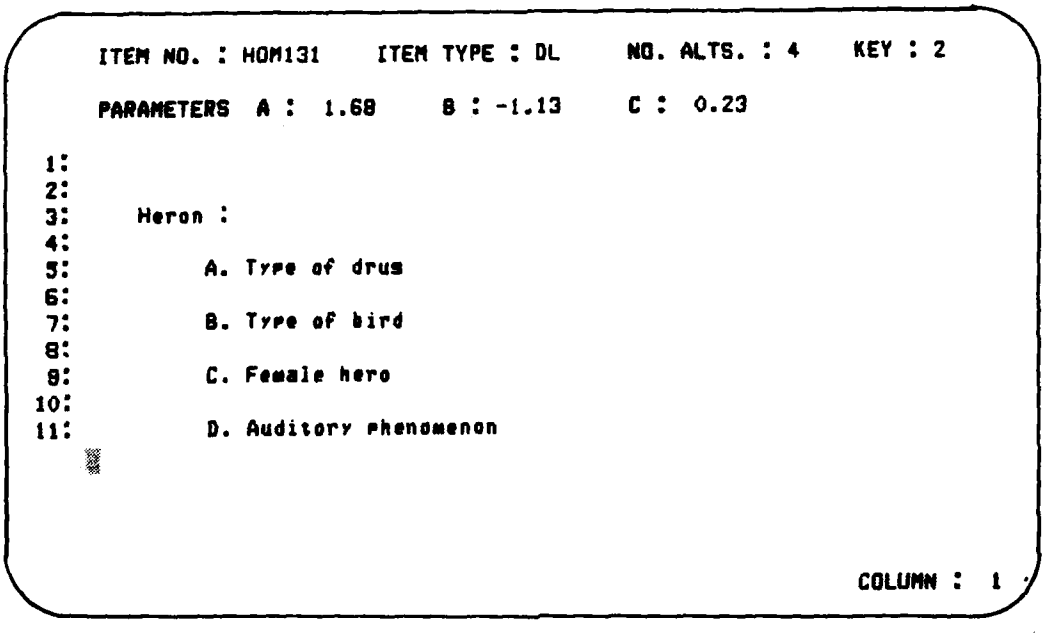


Figure 1: Screen-Oriented Entry System

Graphic Items. Graphic items consist of lines and curves and limited amounts of shading. Typical examples of graphic items are the pulley-and-gear items encountered in mechanical reasoning tests. Most items conventionally produced by mechanical drawing techniques are probably graphic items.

Graphic items are much more difficult to enter than are textual items since there exists no one-to-one correspondence between buttons on the keyboard and the symbols to be produced. There are three basic means of accomplishing graphic item entry: 1) brute-force specification of pen movement through coordinates entered in textual mode; 2) digitization (or analog-to-digital conversion) of mechanically drawn figures; and 3)

computer-aided figure construction. The typical graphics terminal has the internal capability to draw a line between two points. If the lines are made short enough, complex curves can be made by drawing point to point. The brute-force coordinate method of specification requires that the desired figure be laid out in coordinate form and all line endpoints be specified. This is a very tedious procedure for figures composed of anything but straight lines.

Digitization represents the opposite extreme in terms of ease of entry. It can be accomplished through either photographic or mechanical means. Photographic digitization is accomplished by converting a video image of a drawn figure to a digital representation. Mechanical digitization is accomplished by tracing the drawn figure over a mechanical converter or tablet (see Newman and Sproull, 1973 or Walker, Gurd, and Drawneek, 1975 for details of various forms of tablet digitizers). Both methods of digitization require a drawn version of the figure to be converted. The photographic process requires somewhat more equipment than the mechanical process. The mechanical process requires a steadier hand. Both are capable of the same end product.

The final mode of graphic item entry is the creation of the item directly at a graphics terminal with assistance from the computer. Conceptually, this appears an ideal method of item entry. Practically, this method requires more sophistication of the user than the previous two methods. The typical graphics terminal is capable only of taking coordinate specifications, as generated in the brute-force methods discussed above, and drawing lines between them. More complex actions require additional graphics software to assist the user. The utility of the computer assisted approach to item entry will depend on the abilities of the user as well as the power of the software.

Computer assisted graphics software typically consists of symbols, methods for defining new symbols, and procedures for manipulating the symbols. Examples of possible symbols are gears, pulleys, bullets, and maps of states. Some manipulations are positioning, scaling, rotating, and truncating (Walker, Gurd, & Drawneek, 1975). On-line creation of a graphic item is accomplished by selecting a symbol and then manipulating it. If an adequate supply of symbols is available, the computer assisted approach can be a very efficient way of entering graphic items. Specification of new symbols is typically accomplished by the brute-force coordinate specification or by a high level graphics language method. Once created, these new symbols can then be manipulated like any other symbol. The ultimate evaluation of this type of item entry will depend on the repetitiveness of a limited group of symbols. If a symbol is not repeated, this procedure degenerates to brute-force specification.

Pictorial items. Pictorial items may be thought of as very complex graphic items allowing varying shades of contrast. Theoretically, they can be entered in the same modes as can graphic items. Practically, however, photodigitization is the only entry technique that is not unacceptably tedious. The process used is, from the user's perspective, essentially the same used for graphic items.

Maintaining the Items

A computerized item pool rarely remains static over the life of the testing system. In addition to adding new items, old items are corrected, revised, or deleted to keep the item pool current and free of inferior items. Several different methods of item modification are possible.

Textual items. Two basic editing systems are possible for editing textual items. One is a line editor. The other is a screen editor. The line editor allows modification of an item on a line-by-line basis. A pointer is positioned at the line of interest and modifications are made to that line. The line editor requires minimal hardware capabilities and is typically used with low speed terminals where a minimal amount of output is desired. Line editors typically require a fair amount of operator training and have high vulnerability to mistakes since the material edited is usually seen only in part.

A screen editor looks very much like a screen-oriented entry system and may, in fact, be implemented using the same software. Instead of entering the item, however, the item is retrieved in its entirety and is displayed on the screen. Editing is accomplished by moving a cursor to the desired position and simply substituting a new character for the old one. All changes thus made are immediately visible to the user. Since the entire item, in its current configuration, is displayed on the screen, the screen editing system is considerably less vulnerable to gross errors in the editing process. An editing error that would destroy an item is quickly recognized and corrected before permanent damage is done.

Graphic items. Editing of graphic items is essentially a continuation of the item entry process. If the graphic item is defined by a coordinate set, the coordinate set can be treated as text and modified as a textual item. Using a computer assisted graphics terminal, editing of a graphic item amounts to retrieving the item to the graphics terminal screen and then proceding in the same manner as if the item was being entered. At the end of editing, the old item is simply replaced by the new one. For a digitized item, the simplest editing procedure is to modify and redigitize the original.

Pictorial items. Due to their complexity, little editing is typically done to pictorial items. The most satisfactory form of editing these items consists of retouching the original picture and then redigitizing it. No editor for these items is likely to be included in a computerized testing system.

Creating the Test

A pool or bank of items in a computerized testing system typically contains many more items than will ever be administered in a given test or test battery. A subset of items thus needs to be selected and organized. To specify such a test or battery, the instructions and tests must be sequenced, appropriate response editing modes must be chosen, test data to be retained must be specified, and the sequence of items within the individual tests must be determined. The first three acts

are discussed below as components to specifying the testing environment. The last consideration is discussed in a separate section on testing strategies.

Specification of the testing environment. In paper-and-pencil testing, it is the test proctor who creates the testing environment, telling the examinees which tests to take, how to take them, and when to stop. The environment is specified, more or less, by the instructions given to the proctor. In computerized testing, the computer assumes much of the proctor's role and is responsible for communicating what the examinee is to do and for watching over the examinee to make sure the required tasks are performed.

Specification of the testing environment is typically routine. It involves choices from a limited number of alternatives to a standard group of decisions. The decisions fall into three basic categories. Condition decisions determine what response configurations will constitute an error condition and how that condition should be handled. Retention decisions determine what data resulting from administration of the battery should be retained in the testing record. Sequencing decisions determine the order in which test and instruction sections will be administered. In complex test batteries, sequencing decisions may also decide which sections are to be skipped.

For small testing systems, specification of the testing environment can be accomplished using a simple question-and-answer specification procedure. The computer system asks a question regarding the environment and lists the alternatives. The test constructor then responds with an appropriate choice. This procedure works quite well if the user makes no mistakes and does not change his/her mind in the process of specification. Editing in this procedure can be cumbersome.

An alternative procedure that is likely to be preferable is a screen-oriented template-filling procedure illustrated in Figure 2. The shaded areas in Figure 2 represent computer prompts. The unshaded areas represent user entries. Such a procedure functions in much the same manner as a screen-oriented item-entry procedure. Appropriate entries are made in the spaces provided. All entries are processed for logical consistency (i.e., the edit limits must be reasonable and the branches specified must all be possible). Errors are corrected by simply refilling the incorrect field. This specification procedure is, for simple specifications, essentially equivalent to the question-and-answer procedure in which responses are made to questions instead of labelled blanks. The advantage of the screen-oriented procedure is the ease with which corrections can be made. Editing is an integral part of the screen-oriented system but must be a separate component of the question and answer system.

Specification of individual testing strategies. While specification of the testing environment may be routine, specification of individual tests typically is not. One major advantage of computerized testing is its ability to administer a tailored set of items to each examinee, tailored to the examinee's responses to the items as they are administered. Probably more than 100 different administration strategies exist.

```
BATTERY TITLE : ASVAB-14                              DATE : 82 OCT 03

NAME TEST A : INST-A14

     RESPONSE EDIT          ERROR DETECTION          SCORING MODES

     MAX. CHARS: 1          COACHED   :              PROP.CORR.: X
     ALPHA LIM.: NONE       COLLAB.   :              BAYESIAN  :
     NUMERIC   : 1..NALT    RANDOM    : X            LIKELIHOOD:
                            PANIC     : X            LATENCY   :
                            TIME(SECS): 120


     DATA RETENTION         NEXT MODULE

     SCORES    :            ALWAYS    : TEST B
     RESPONSES:             FLAG 1 SET:
     ID NUMBER: X           FLAG 2 SET:
     BIODATA   : X          FLAG 3 SET:
     TIME      : X
```

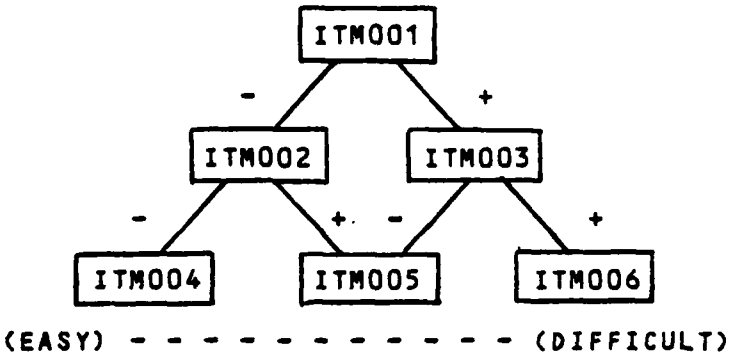
Figure 2: A Template-Filling Specification Procedure

Design of the strategy specification portion of the subsystem will depend on how many of these potential strategies the testing system must be capable of handling. A testing system for research use may need to be able to specify all of the strategies; a system for practical testing may need to differentiate only a few.

For a research system that must specify many strategies, two choices appear to exist for the specification system. The first is to specify strategies by explicit programming in a computer language such as FORTRAN for each new strategy. This route allows for specification of any testing strategy that can be logically described. Implementation of a new strategy requires the services of an individual skilled in the programming language, however, and may require a considerable investment of time. A large research testing system designed along these lines is operational at the University of Minnesota (DeWitt & Weiss, 1976; Lind & Messinger, 1980).

Alternatively, a special purpose programming language can be designed explicitly for specifying test structures. Such a language allows the flexibility of a regular programming language in the design of test structures but is limited to that application and is thus easier to use and to read. An example of one such language (Vale, in press) is shown in Figure 3. In that figure, a pyramidal adaptive testing strategy (Weiss, 1974) is diagrammed and specified. Each node in the diagram represents an item and items are arranged from left to right in order of ascending difficulty. After each item is administered, it is scored and testing proceeds with an easier item if the response is incorrect (-) and a more difficult item if the response is correct (+). The programming

language specifies the branch tree by providing branch items for incorrect and correct responses to each item. An item with no branches indicates an end to the test.




```
$
$ SPECIFICATION OF A PYRAMIDAL TEST
$
SET ITEMP = #ITM001
BRANCH
#ITM001 INCORRECT:#ITM002  CORRECT:#ITM003
#ITM002 INCORRECT:#ITM004  CORRECT:#ITM005
#ITM003 INCORRECT:#ITM005  CORRECT:#ITM006
#ITM004
#ITM005
#ITM006
$
$ END OF SPECIFICATION
$
END
```

Figure 3: A Test Specification Language

The advantage of such a specification language over a standard programming language is two-fold: First, it allows more efficient specification, from the test constructor's point of view, because much of the overhead programming (e.g., scoring) is done with a single statement and some strategies can be specified with much less coding. Second, the obvious relationship of the language to the test structure increases its readability for individuals not intimately familiar with the language and allows quick mastery by test constructors not trained in any of the standard programming languages.

For operational (i.e., non-research) systems needing only a few different strategies, a more automatic approach can be taken to the specification process. The required strategies can all be internally laid out in structural form and the process of test construction will amount to simply inserting appropriate items into the structure. Insertion

946

can be done in at least two different ways. In one, the structure can be presented graphically on the CRT screen and the items can be inserted in the blanks. This procedure works like the screen-oriented entry and edit procedures discussed earlier. To make the process even simpler, the system can be programmed to take a set of items and make optimal allocations within the structure. Specification then amounts to specifying the strategy to be used and the items to be included.

Although the distinction between operational and research systems has been made here, there is no reason that both designs cannot reside within a single system. Automatic specification and specification through a specification language are not mutually incompatible. A hybrid system may use the automatic specification process as a precursor to the language process. The automatic processor, given its input, produces a test specification in the language. The language produced can then be executed directly or can be optimized or otherwise modified. Although such a system may seem cumbersome in description, it may actually produce a system more satisfactory, from the user's point of view, than either or both of the methods implemented separately. The complexity is transparent to the user until s/he has need to take advantage of this hybrid system's capabilities.

Conclusions

Computer capabilities have changed radically over the past few years. The emphasis in system design has shifted from concern with design for efficiency from the computer's perspective to emphasis on the design of systems efficient from the user's perspective (Schneider, Weingart, & Pearlman, 1978). This paper has attempted to suggest some of the possibilities for design of a test specification system emphasizing efficiency from the user's perspective. That emphasis can be summarized in three guiding principles. First, the system should allow specification of the desired results with the minimum possible input. This is accomplished by making specification as automatic as possible without sacrificing required flexibility. Second, the entire specification subsystem should be simply masterable by test construction personnel. This objective is approached by making the specifications parallel the test, in form, rather than appear as a computer program. Finally, the system should be designed to give immediate feedback regarding actions taken by the test constructor. This is the reason for emphasis on processed entry whenever possible.

References

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

Dewitt, L. J. & Weiss, D. J. Hardware and software evolution of an adaptive ability measurement system. Behavior Research Methods and Instrumentation, 1976, 8, 104-107.

Lind, J. & Messinger, D. Personal Communication, October 1980.

Newman, W. M., & Sproull, R. F. Principles of interactive computer graphics. New York: McGraw-Hill, 1973.

Schneider, G. M., Weingart, S. W., & Pearlman, D. M. An introduction to programming and problem solving with PASCAL. New York: Wiley, 1978.

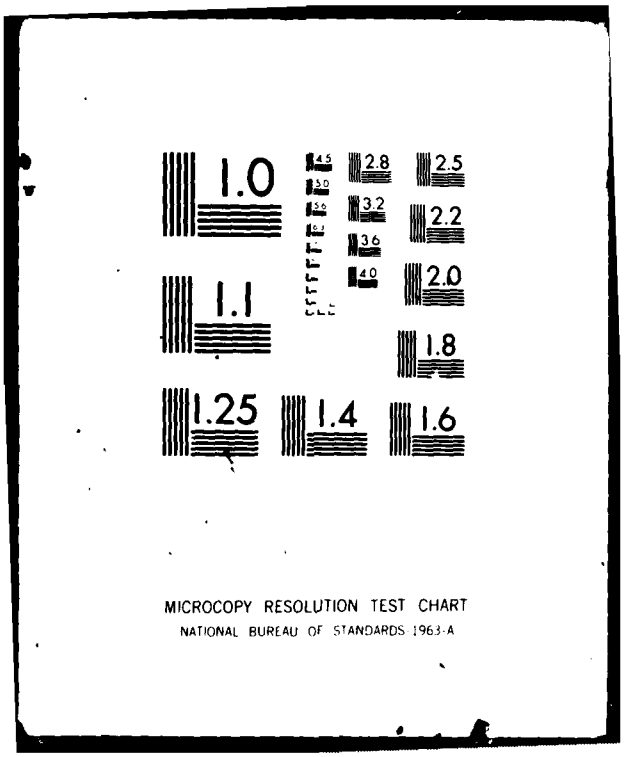
Vale, C. D. Design and implementation of a microcomputer-based adaptive testing system. Behavior Research Methods and Instrumentation, in press.

Vale, C. D., & Brown, J. Computerized item banking: A guide for other jurisdictions with emphasis on the Minnesota experience. St. Paul: Minnesota Department of Personnel, December, 1977.

Walker, B. S., Gurd, J. R., & Drawneek, E. A. Interactive computer graphics. New York: Crane Russak, 1975.

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

AD-A098 678 MILITARY TESTING ASSOCIATION F/G 5/10
PROCEEDINGS OF THE ANNUAL CONFERENCE OF THE MILITARY TESTING AS--ETC(U)
DEC 80
UNCLASSIFIED MTA-22-80-VOL-2 NL
5 6


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

VOORHEES, Phyllis P., U.S. Coast Guard Institute, Oklahoma City, Oklahoma.

ANOTHER USE OF THE MAPL PROCEDURE - AN ITEM EVALUATION TECHNIQUE FOR THE TEST CONSTRUCTOR (Wed P.M.)

One of the results of applying the Minimally Acceptable Performance Level (MAPL) procedure to criterion-referenced tests in the Coast Guard was the development of a technique to evaluate test item stems and responses.

The cut-score setting technique as described by Warm (1978) was used to determine what score represented a minimal level of competency on: (1) a training-needs diagnostic test, and (2) a proposed qualifying test for Deck Watch Officers.

Six subject matter experts scrutinized each test item and response for the purpose of assigning points to each response. During this process, unsuspected flaws in item stems and responses were discovered. These flaws, undetected, could have seriously impacted on test credibility and validity. The application of the MAPL procedure to these tests and the results are described in the paper. Application of the procedure to other types of military testing is also addressed.

ANOTHER USE OF THE MAPL PROCEDURE -- AN ITEM EVALUATION TECHNIQUE FOR THE TEST CONSTRUCTOR

Phyllis Peters Voorhees

Federal Aviation Administration Academy
Radar Training Facility
Oklahoma City, Oklahoma 73125

INTRODUCTION

A modified MAPL (Minimum Acceptable Performance Level) procedure as described by Warm (1978) was used to determine what score represented a minimal level of competency on a proposed qualifying test for Deck Watch Officers.

Six subject matter experts (SME's) were assembled and instructed to form a mental picture of a barely acceptable performer in their field of competency. Keeping this performer in mind, the SME's were instructed to read each of 200 test items and assign points to the wrong responses on the basis of whether a barely acceptable performer would know that a distractor was definitely wrong, or whether an individual could choose a particular wrong response and still be acceptable on the job.

On each item, the correct response was keyed and assigned two (2) points. Two (2) points were also to be assigned to any response the barely acceptable performer might not know was definitely wrong, and zero (0) points assigned to any distractor that even the barely acceptable performer should know was wrong. The SME's were instructed to assign one (1) point to each wrong response for which they could not give a definitive judgement.

For each item, the points assigned to all the choices were added and averaged. The average total points were divided into two, and the quotient termed the Alternative Similarity Index (ASI). Using this formula, an ASI of 0.25 would be the lowest ASI possible, an ASI of 1.00 the highest.

The SME's "MAPLed" two forms of the test -- Form 51 and Form 52. Each form contained 100 test items. After the MAPL procedure was completed, the 25 items on each form with the lowest ASI's were examined to determine, if possible, why these items had such low ASI's.

DISCUSSION

On both test forms, 72% of the items with the lowest ASI's contained item construction flaws, 36% - 40% performance predictor flaws. Some items were judged both poorly constructed and poor

451

predictors of performance.

A few months earlier, while "MAPLing" a training needs diagnostic test, these two kinds of flaws had been noted, but a formal review of the low ASI items was not conducted at that time. On the test for Deck Watch Officers, however, notes were taken of the SME's comments as they read the items and assigned points to the responses. Interviews were also conducted after the "MAPLing" was completed.

On Test Form 51 (see Illustration 1), 20% of the items with low ASI's were judged not related to the PERFORMANCE of an examinee, 20% judged highly questionable on that basis. Similar performance predictor flaws were also identified on Test Form 52 (Illustration 2).

As the SME's assigned two points to a response a barely acceptable performer might NOT know was definitely wrong, they were also evaluating the performance relevancy of the item. An item that called for the definition of a term or for a bit of nice to know information usually had a low ASI (see Example 1). The SME's assigned two points to each response on the basis of whether a Deck Watch Officer knew who authorized lines of demarcation or not, it would have little impact on job performance. In Example 2, the SME's suggested that the point of knowledge to be tested should have been when or under what conditions a lookout should be posted rather than the rule that called for the posting. Most of the responses to this item were assigned one or two points each. The responses in Example 3 were similarly assigned one or two points each, not because the point of knowledge was irrelevant to the job but because the responses were so close together in meaning that any one of the speeds would have represented caution. The SME's deemed it highly questionable whether "in the real world" the vessel would be slowed to, and maintained at, exactly 5 mph.

Examples 4 through 8 illustrate the most frequently identified item construction flaws on the items with low ASI's. The Coast Guard Institute's Exam Development Manual was used as the source for item construction standards. On Test Form 51, 16% of the low ASI items asked the question in the negative form, 48% had either inclusive stems or complex responses that were either inclusive or contained secondary qualifications that made them inclusive. It was interesting to note that negative questions without enough information in the stem to lead the examinee to either a single correct answer or a small set of correct answers appear to bother the SME's as much as trainees. More research needs to be done in this area, as in the area of the complexity of responses. Even the SME's tended to think the test constructors were trying to "trick" them on inclusive responses or responses with secondary qualifications.

952

CONCLUSION

The modified MAPL procedure as applied to the qualifying test for Deck Watch Officers not only provided a cut-score setting technique but also produced a test instrument that was better constructed and more related to job performance. The MAPL procedure appears easily adaptable to the constraints of time, people, and money. Although much research needs to be done on the efficiency of the technique and its "adaptations", the MAPL technique appears to be an extremely effective means for the use of training units and training facilities to help produce realistic, valid test instruments where a cut-score or pass/fail score is implemented.

(95.3)

Illustration 1

TEST ITEMS WITH LOW ASI'S

Test Form 51

ITEM STRUCTURE FLAWS 72%

- Inclusive Stem 4%
- Ambiguous Stem 4%
- Negative Stem 16%
- Complex Stem 4%
- Complex Response
  - a. Secondary Qualification 16%
  - b. Inclusive 28%

PERFORMANCE PREDICTOR FLAWS 40%

- Non-Performance Related 20%
- Highly Questionable 20%

ITEMS WITH NEITHER TYPE OF FLAW 8%

Illustration 2

TEST ITEMS WITH LOW ASI'S

Test Form 52

ITEM STRUCTURE FLAWS 72%

- Negative Stem 8%
- Complex Response
  - a. Secondary Qualification 24%
  - b. Inclusive 32%
- Unparallel Response 8%

PERFORMANCE PREDICTOR FLAWS 32%

- Non-Performance Related 16%
- Highly Questionable 16%

ITEMS WITH NEITHER TYPE OF FLAW 12%

(75-4)

Example 1

NON - PERFORMANCE RELATED

Lines of demarcation are authorized by the _____.

A. United States Congress
*B. Commandant of the Coast Guard
C. Army Corps of Engineers
D. President of the United States

Example 2

NON - PEFRORMANCE RELATED

A vessel underway is required to keep a proper lookout in accordance with the _____.

A. Steering and Sailing Rules
B. Pilot Rules for Inland Waters
C. General Prudential Rule
* D. Rule of Good Seamanship

Example 3

When a vessel passes within 200 feet of a stationary dredge during dredging operations, the vessel's speed is limited to _____ miles per hour.

A. 2
B. 3
* C. 5
D. 7

QUESTIONABLE RESPONSES

955

Example 4

NEGATIVE STEM

In which situation does the rule of Special Circumstance NOT apply?

A. Meeting several vessels at one time
B. Meeting a tug bound downstream in a heavy current
*C. Meeting a vessel end-on or nearly end-on
D. Encountering a vessel engaged in laying cable

Example 5

NEGATIVE STEM

Which signal is NOT an international signal of distress?

A. November Charlie
B. SOS in Morse Code
*C. Rockets or shells throwing green stars
D. Explosives fired at one-minute intervals

QUESTIONABLE ANSWER

Example 6

INCLUSIVE STEM

Three short blasts on a vessel's whistle indicates that the vessel is backing at _____ speed.

A. one-third
B. two-thirds
C. full
*D. any

INCLUSIVE RESPONSES

956

Example 7

What light(s) must be displayed by a 120-meter ship at achor?

A. An all around red light at the masthead and a white stern light

*B. An all around white light forward and an all around white light aft lower than the forward light

C. An all around white light forward and an all around white light aft higher than the forward light

D. An all around white light forward only

COMPLEX RESPONSES
(INCLUSIVE)

Example 8

When a vessel sounds one long blast followed by three short blasts, this indicates that the vessel is

A. backing from a pier

*B. backing from a pier in sight of another vessel

C. backing from a pier at FULL speed astern

D. backing from a pier at FULL speed astern in sight of another vessel

COMPLEX RESPONSES
(SECONDARY QUALIFICATIONS AND INCLUSIVE)

WALDKOETTER, Dr. Raymond O., US Army Research Institute for the Behavioral and Social Sciences, Fort Sill Field Unit, Fort Sill, Oklahoma.

THE FIREFINDER RADAR TRAINER: A TRAINING DEVELOPMENT ANALYSIS APPROACH (Thu A.M.)

Using a trainer with high fidelity and computerized operation almost directly assures such functions as training transfer and effectiveness. To observe the sequence of hardware design with human factors and personnel requirements accounted for, should lead to a direct conversion analysis for any testing and evaluation aims trying to measure trainer and hardware interface effects. If the sequential development of a trainer is purposefully documented, then knowing which critical task behaviors to evaluate becomes relatively obvious. However, the analysis of training development needs is constrained by not having the analysis approach and model which employs only several operators or mechanics and parallels a "test pilot" analogy. With accelerated development of computerized training systems and design of individual or small crew performance criteria, alternative analysis approaches for training development are required to compensate for any abbreviation in the hardware and weapon system development process and projections for sequenced training development objectives. An approach is presented as one alternative analysis technique to augment training development where certain preliminary development milestones were bypassed for the Firefinder Radar Trainer (A17E11). The trainer reference and orientation seems to best support a comprehensive review of each facet of training tailored to prepare the operator for the Artillery locating Firefinder Radar (AN/TPQ-37).

THE FIREFINDER RADAR TRAINER: A TRAINING DEVELOPMENT ANALYSIS APPROACH[1]

Raymond O. Waldkoetter
US Army Research Institute for the Behavioral and Social Sciences
Fort Sill Field Unit, P.O. Box 33066, Fort Sill, Oklahoma 73503

and

Thomas Curran
Radar Division, Counterfire Department
US Army Field Artillery School, Fort Sill, Oklahoma 73503

INTRODUCTION

As may happen the development of a program-of-instruction (POI) for a new type of equipment, such as the Artillery-locating Firefinder Radar (AN/TPQ-37), can run into certain gaps in personnel and technical documentation. Projection of certain personnel skill requirements and related human factors could not get the detailed specification desirable to follow on from equipment system development to guide the extensive objectives for the POI design. Especially when the equipment has a companion training device or simulator like Firefinder Trainer (A17E11), there is an inherent difficulty in making sure the trainer and equipment are maximally interfaced to decide confidently what instruction is really needed and how the trainer skills will most efficiently transfer to the actual radar system. Where initial development procedures were accelerated or combined both for the equipment and trainer, the gaps in training guidance were temporarily accepted knowing that the engineering aspects were operationally sound and technically valid. A situation then is produced in which the responsible supervisors and instructors must integrate their training design by relying on the equipment and trainer manuals for operators while experimenting with student subjects or test players before the official training program is begun. As the instructors progressed with their training development it became evident that training skills and tasks should have more emphasis and clarity to adjust the instruction sequence and practical exercises toward the acquisition of the requisite operator tasks. A questionnaire design was constructed as an alternative analysis technique to obtain a review format which would furnish omitted information while interpreting the pertinent training development issues and task behaviors for enhanced operator training. This analysis approach was conceived in terms of augmenting those minimal training development documents which should have optimally furnished the automated task acquisition procedures from initial diagnostic steps and later analysis options during the equipment and trainer design phases.

---

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

961

It must be noted that this paper will not attempt to critique the A17E11 Trainer nor the Q-37 Radar equipment. The primary attention is devoted to the trainer, however, as a simulated training device to better describe its unique affect on the full range of Firefinder radar training and the application results from the particular questionnaire analysis. In spite of certain omitted training information for development of task performance standards according to qualified personnel requirements, such a trainer usually assures critical task functions will attain relative training transfer and effectiveness standards. This is at least a potential consequence since the A17E11 Trainer was highly engineered with computerized, operational, and visual similarity to the radar equipment. As one viable alternative analysis approach the review format design for the questionnaire is presented as a training development tactic to more thoroughly help define instructional content and system proficiency levels when such information is not delivered from the equipment or weapon system development process and for guidance of sequenced training task objectives.

METHOD

Human factors and personnel requirements were evidently accounted for to some extent in the equipment system design and trainer man-machine interface decisions without making a continuous and explicit set of documents. An analysis for training transfer and operational evaluation (Kolstrom et al., 1979) measured selected test issues against criteria identified by the Field Artillery School, but the student, trainer and equipment system interfaces were not measured with the precision that might seem truly conclusive for the tested issues. Basically, adequate data were observed for generalizations regarding training effectiveness and transfer and instructor-student interface.

Yet that vivid network of obviously projected behavioral task objectives was not evident for the training design stage. The functional task analyses did not proceed to trace across the student, trainer, and equipment system interfaces in sufficient documentation to infer what test procedures would most readily assess the training issues. Measurement precision in this context was diminished by not being able to decide the critical training task implications for trainer performance and subsequent instructional program construction and validation. This limitation seemed to arise in part because many of the simulated training device (A17E11) and Q-37 system output measures did not relate to specific criteria for the rather general test issues. Sequential development of a fully documented trainer should then clearly identify which specific critical training task behaviors to train and operationally evaluate.

As the detailed information was not available for the instructional course design to more conveniently prepare the training sequence for Firefinder operators, because of the accelerated development schedule, the questionnaire review format was designed to augment the decision process for modifying training task behaviors under instruction. The interview questionnaire methodology was developed for a training analysis review within the instructional development process and with certain format characteristics to explore training course content dimensions and why task behavior may or may not be performed within expected boundaries

of skill and time. This questionnaire format structure differs from those Army Research Institute instruments adopted for an early training effectiveness analysis (Finley & Tremble, 1977), a task validation procedure (Tremble & Finley, 1978) and human factors type of evaluation (Griffith, 1979), in that the interactive experiences of test-player students and instructors are inventoried to identify the student, training device and equipment interface perceptions which may yet receive sufficient modification to assure enhanced course content and operator proficiency.

A first draft questionnaire was prepared after reviewing a preliminary program-of-instruction (POI) composed by the senior non-commissioned instructor, immediate manual and contractor handbook documents, and observing six test subject students follow that trial instructional sequence on the A17E11 training device with complementary classroom activities. The draft questionnaire was administered to test students, discussed with the instructors still familiarizing with the training device and Q-37 Radar, and revised based on a critique of radar division officers, NCOs, and all Firefinder instructors. Questionnaire items were edited and formatted to acquire data on training background, specific functional task behavior, training device operations, equipment/device interface, instructor attributes, training attitudes, training media and modes, classroom activities, and course evaluative procedures. None of the content areas is exclusive but defining nine possible content sources could increase the utility and interpretive value of the anticipated data. To avoid the chance of creating a too obvious relationship among instrument items and content areas (i.e., context effect), items were not grouped by category or in any fixed sequence. Some authorities might argue in favor of more accurate measurement with factor-oriented areas, but the intent was to obtain relatively independent judgment on each item regardless of any prior awareness of the factual or attitude items. Using the somewhat random item arrangement had helped acquire apparently independent respondent answers on a survey instrument for selected missile extension training materials (Waldkoetter & Milligan, 1980).

A final working version of the Firefinder radar training inventory and interview form was jointly agreed on with the Field Artillery School to be administered to test-player subjects preparing via their training to conduct pre-scheduled developmental and operational Q-37 tests in a field environment. Player personnel would sample experienced enlisted soldiers and NCOs with several warrant officers to get that experienced assessment of training device and equipment instruction and operation at the end of a full but condensed POI rehearsal. It is agreed that the technical aware and trained soldier in this situation would yield an objective and informed set of responses which would be sufficiently independent from instructors and uninitiated students to accurately qualify instructional development policy, procedures, content and simulated training transfer practices. With two test-player classes scheduled it was also decided that two separate administrations would be given several weeks apart, as well as combining the instrument data, to determine whether each class has similar or different experiences and if overall findings might indicate any definitive or divergent trends. Between 50 and 55 test players were assigned for participative training sessions and would render, it is believed, an acceptable level of expert judgment about the Firefinder training strengths or deficiencies.

In addition to this procedure the TRADOC System Manager (Wickliffe, 1980) recommended that the questionnaire be given later to one of the classes of about 25 personnel to see if intensive operation of the Q-37 Radar might cause expression of some other analytical perspective from which the POI could receive positive training changes and note critical equipment operational incidents. Comments are requested on over 35 percent of the questionnaire items which will encourage an in-depth interpretive analysis of each specific item eliciting comment and those which have clustered relationships with any one or more of the items receiving comment. In one sense the comments can function as a qualifying criterion benchmark, because if the items are generally purposeful and precise, minimal commentary could occur, and if so, just where instruction is not usually perceived in adequate detail.

Participation by instructors in assisting with the content selection for the questionnaire instrument and their planned review of completed questionnaires served as other direct evaluative uses of such an instrument to diagnostically examine their content approach to the POI and next to think in terms of prescriptive instructional techniques. With perception of instructional skills and tasks viewed in greater accord with student learning and retention performance objectives, their instruction has gained added flexibility and choices for modifying student behavior and task proficiency. During the formative stage of POI training design and practice, instructor completion of the review questionnaire was seen as an internal audit by the supervisory instructor who was interested in gaining a high level of cohesion among instructors and establishing a common basis for explaining student, training device, and Q-37 functional relationships. Another use of this training development analysis approach could be observed in having instructors analyze their own simulated training experiences by inventory instrument items which permitted instructors to review the criteria for user/operator transactions with automated A17E11 and Q-37 systems.

Questionnaire items vary in number of alternatives depending on the item response requested and the possible range of judgments which can logically be executed in reaching one specific opinion or solution. The majority of items have multiple choice judgment and scaling formats with about 20 percent being "yes" or "no" queries, and 20 percent requesting factually verifiable data and which can offer a content reliability check for the subject test players. Eighty-one items were presented with one designed for an alternate entry code to assist the subsequent data processing of 80 card items. While several statistical tests are to be applied in the final report phase to account for the differential significance of all items and the applicability of the findings, a principal technique will be to convene a review analysis panel (RAP) to examine the data and findings through a series of iterative discussions to compare and decide on the actions suggested by the results. Implications for instructor development, student accessions, POI changes, extension training materials, and system modifications were to be discussed and evaluated as a summary function of the training survey.

RESULTS

Data obtained from the questionnaire are testable by merely observing any patterns related to the types of items and intended use for POI development and the degree to which the review analysis format may have assisted in gathering valuable information. A more specific presentation of results is found when the student, instructor, training device, and Q-37 equipment interfaces are analyzed for item responses and correlations. Format effects have some probable interaction with those discernable student, instructor training device, and equipment variables. The command and instructor/review analysis panel (RAP) will use a comparative process by which these expected relationships can be identified and recommendations put in implementing training priority and terms of operation. Those results having feedback for system development and testing procedures are interpretable in light of directed user/operator transactions. These transactions are thus related to given items and as such transactions are projected to given standardized behaviors or products from design through system operation.

Questionnaire structure and item characteristics depend somewhat on the order and expression of content material and the style of item display. Whether these characteristics affect the results obtained is never wholly answerable. Training issues in course development were thoroughly exhibited by the variety of types of data requested and item formats. Not only significant proportions of respondents are measurable but there are numerous techniques by which item answers and comments are open to cross-referencing interrogatories to exhaust item meaning and context order. Definition of A17E11 training device strengths and constraints was demonstrated by obviously significant trends across many items, which were screened by instructors and senior test-player observers who recognized nearly instantly where different or corrective instruction would effect training changes to get a faster, efficient or accurate student task behavior.

Instructor responsiveness to questionnaire items took on a judgmental set and answers with comments tended to give a critique of the instructional process and problems the student-players were meeting. Some items were not answered because instructors thought they were student oriented only, but they would explain why the items were helpful if answered by the students. Most of the instructors had not taught in a simulated training device facility nor with such a system as the A17E11 trainer using a scenario data package translating a sequence of training requirements into an interactive system between the student and computer. Scenarios in the trainer system carry the emulation of prime system (Q-37) functions as a full sequence of scenario exercises using techniques of computer-assisted and computer-managed instruction (CAI/CMI), furnishing an objective background for instructor use of the questionnaire or for "troubleshooting" on-going training. The evaluative purpose of the questionnaire items is not unlike many thought procedures an instructor would follow, guiding the trainer system of six student stations from the instructor console and monitor, when making tentative evaluations of student activity. But the extremely heavy demands on a working instructor usually discourage complete expression of similar inventory questions probing cause and effect for student task behaviors.

Test-player students though differing from the normal student in-take could draw on their prior Army experiences as students, technicians and supervisors to assess whether items would yield information to modify the POI, trainer operations, or equipment system training as requested. Student answers are seen as the most important source for determining the suitability of course content to result in adequate learning and retention and to insure operational proficiency at a unit site. Statistical significance is not a single criterion in deciding whether instructional changes or curriculum revision might seem advisable. A review panel can look at and discuss associations between items and if one result may stem from the same or a separate cause affecting student item responses. At a minimum a significant difference between student replies or comments will bring a fresh evaluation of why a given relationship may exist within the training process or equipment system. Where little consistent MOS identity was found for the student test players and they appear randomly sampled across Artillery, missile and computer duty positions, any basic group difference will occur by paygrade or rank if at all because of equivalent personal data for education, PMOS training and time-in-unit assignments. In searching for practical measures of student progress the questionnaire has some guidance use in finding which items are related to student learning problems and attitudinal performance on the Firefinder task sequences. Even though the training largely is dependent on instructor skill in spite of an automated trainer, knowledge of student perceptions will make the training more responsive to achieving task objectives.

Classifying answers by trainer or equipment system relevance and supporting comments can describe another dimension of analytic data. These data can indicate characteristics specifically for each system and the transfer elements which are interface criteria to estimate the efficiency of simulated training. A difficulty may arise in trying to partial out a trainer practice effect since skills and tasks oriented initially in a trainer practice sequence will affect a later assessment of radar equipment training and guide that operational learning. By making any distinctions which may differentiate the trainer and equipment, while recognizing high trainer fidelity, will also permit the design of flexible media and instruction modes to support preferred and proficient training practices. Perhaps the most particular value from this analysis is to help maximize trainer experience and that transfer to equipment without compounding student inhibition by emphasizing minor trainer constraints.

Results for one of two test-player groups has demonstrated, along with specific item functions, at least three directly observed patterns of responses. Instructor, trainer/equipment, and training content and delivery categories seem rather visible as the data summary is examined. Instructor related items seem to confirm a desire to keep an instructor immediately on-hand for aid and training control for fewer students, though the systems are extensively automated using CAI/CMI techniques. Trainer and equipment items indicate similarity in task performance difficulty within and across systems, a desire to get as much "hands-on" time as possible on each system for skill proficiency, and indicate too, some moderate difficulty to operate the systems. Items which appear more training content and delivery focused suggest that instructor skill varies to a degree, POI content should remain about the same but minor system differences need attention, the automated trainer evaluation is very helpful to guide the student simulation, and student mental ability must fall in the upper-average range to expect reliable unit performance.

(666)

DISCUSSION

Because of the training design limitations imposed by a rapid system development schedule for the Field Artillery's Q-37 Firefinder Radar and its A17E11 training device, analysis of automated training development needs appeared constrained. A training device was devised to reduce costs that would accrue if soldiers were trained only on an operational system, where numbers trained would be limited and length of time to train hard to control. Such a training device can concentrate attention and effort to accomplish a better integrated instructional process with attainment of individual and crew task objectives. However, with the rapid schedule, early training design documents could not fully accompany the radar and device thereby constraining development of test and evaluation acceptance programs. These programs could furnish training design guidelines if planned with greater detail for human factors and personnel support from the actual radar and device design products. An inventory form for Firefinder training requirements was constructed to alleviate part of the information constraint, and was oriented on the training device to maximize the utility of the commitment to simulated training and transfer to the Q-37 Radar.

The questionnaire approach to analyze training development needs or consequences is not unique, yet a generalized review format was newly formulated to recover performance objectives rather implicitly expected in the systems design. There is an innovative procedure, additionally, in gathering and synthesizing information for course design which was not previously referenced nor based on immediately observed training conditions. Moreover, the methodology application has pointed to finding a further clarification and coherent integration of system design procedures which can project specified training guidelines and test requirements so that an economical and accurate strategy will guide the parallel activities of Artillery training development and test system evaluation. If the total coordination of system design, test and training task objectives is conceptualized and implemented, training device and equipment systems should comply with the proposed operational standards and still collect any suggested modification data as test and training decisions are formulated well before system installation.

A convergence of Firefinder course design and test requirements occurred at the concept evaluation and user test phases (Lovell et al., 1980) indicating two apparent conditional training constraints. First, there was a limited concept available of what should constitute testable training on the A17E11 device and Q-37 Radar. And secondly, the user test was compelled to base evaluation of training and suitability on part of the first conditional constraint findings, while next extracting measures from training still subject to other revisions for content and proficient task behaviors. Under these conditions there were training issues and measures that would not reach an optimal level of precise testing, because the student, instructor, training device and equipment system interfaces were not identified in sufficient terms to evaluate just which behaviors, operations or system features defined the best test of system capability. In the face of pending training development and test schedules other information gathering alternatives were not proposed. Neither could an analysis approach nor model concurrently evolve which would more economically "test" a small number of operators or mechanics

in a manner similar to a structured "test pilot" evaluation (Kratochwill, 1978). Some alternative analysis approach was inherently recommended to bring a degree of synthesis between course design objectives and test sanctions, to accommodate training and clarify test results for any course, device, or equipment changes.

Questionnaire results indicated that the instrument was effectively constructed to describe student concerns and equipment system relationships. This observation is confirmed by the Firefinder training device item responses given by students and instructors and a critique of interested reviewers. Though a "one-time" instrument, the questionnaire review format for training analysis may suggest some type of standardized approach by which critical training issue and equipment capability measures can move toward increased utility and precision. Surely, other alternative analysis approaches for training development will generate with the expected development of more such advanced computerized training systems.

If design of performance criteria are begun with the system creation, later evaluation of test requirements and training can proceed directly from those documented decision options and design guidelines which formulate and assimilate human factors and personnel requirements into system engineering and functional operations. It is not too optimistic, perhaps, to forecast a test and training development system from which each performance requirement is so specified with standards and man-machine interface controls, that a few selected test-players will reliably exercise the system operational capabilities and automated training requirements through a completely instrumented scenario. Group test evaluation procedures may soon seem slightly anachronistic to measure normative task behaviors if a simulated training device like the A17E11 displays a high degree of fidelity, satisfies rigorous design parameters, and would have full design evaluation guidelines for test and training procedures.

An interim analysis instrument as advocated in the paper does yield significant training material for course content and review of simulated performance criteria. When training information documents may have omitted certain simulated and prime system training and instructional guidelines during accelerated development, that situation should not deter an effort to construct an auxiliary training inventory and interview form to obtain the best available data. Having a flexible questionnaire format to interpret user/operator transactions with an automated system (Berger & Hawkins, 1979), will remain a viable approach to reinforce and support training development under constrained conditions. Using this approach, simulated training device and equipment operations were favorably complemented in the Firefinder Radar System. By applying these questionnaire results formatted on instructor, student, and training device/equipment interface perceptions, a progressive enhancement of operation training was planned.

REFERENCES

Berger, B.M., and Hawkins, H.H. Occupational analysis: An automated approach. In T. Abramson, C.K. Tittle, and L. Cohen (Eds.), Handbook of vocational education evaluation. Beverly Hills, CA: Sage Publications, Inc., 1979.

Finley, D.L., and Tremble, Jr., T.R. An analytic training effectiveness analysis for a CTEA update (Research Memorandum 77-19). Fort Benning, GA: US Army Research Institute for the Behavioral and Social Sciences, November 1977.

Griffith, D. TACFIRE OT 056 human factors evaluation (Research Problem Review 79-5). Fort Hood, TX: US Army Research Institute for the Behavioral and Social Sciences, March 1979.

Kolstrom, F.R., Trovato, A.E., Kagawa, F.E., Sisney, M., and Bonacci, N.L. Concept evaluation program: Firefinder radar operator trainer, A17E11, Test (Final Report ACN 24488). Fort Sill, OK: US Army Field Artillery Board, October, 1979.

Kratochwill, T.R. (Ed.). Single subject research: Strategies for evaluating change. New York: Academic Press, 1978.

Lovell, J.A., Klinger, D.R., Trovato, A.E., McLeod, E.S., Raney, R.W., and Meisenzahl, K.A. On-site user test of Firefinder radar operator/maintenance trainers, device A17E11 and device A17E12 (Final Report ACN 24488/80 OTN 435). Fort Sill, OK: US Army Field Artillery Board, July 1980.

Tremble, Jr., T.R. and Finley, D.L. Task validation for the AN/TPQ-36 radar system (Research Problem Review 78-17). Fort Benning, GA: US Army Research Institute for the Behavioral and Social Sciences, September 1978.

Waldkoetter, R.O., and Milligan, J.R. Extension training materials: Differential perceptions among USAREUR Lance missile personnel (Research Report, Draft). Fort Sill, OK: US Army Research Institute for the Behavioral and Social Sciences, April 1980.

Wickliffe, P.T. Personal communication. Fort Sill, OK: US Army Field Artillery Center, August 1980.

WALDKOETTER, Dr. Raymond O., US Army Research Institute for the Behavioral and Social Sciences, Fort Sill Field Unit, Fort Sill, Oklahoma.

MTA PUBLISHING REVIEW GROUP (PRG) WORKING SESSIONS (Mon P.M.)

The Military Testing Association has existed now for over 21 years. Programs resulting in annual proceedings dealt initially with the personnel and proficiency test instrument interests of the Navy, Army, and Air Force. Within a few years the Marine Corps, Canadian Forces, Goast Guard, and Allied Forces began participating with the annual programs being attended also by government and business organizations. Programs gradually increased content to include all areas where relationships with personnel measurement and evaluation might arise. Topic coverage in proceedings presently spans behavioral measurement, occupational analysis, manpower analysis, simulation models, training programs, selection methodology, survey and feedback systems. An opportunity exists to originate and author a scholarly scientific book which will diligently review the principal content areas and critically integrate materials spanning MTA from its origin through the projected 25th MTA Conference and Silver Anniversary, Those thematic developments across the 25 years will become more professionally and technically identified by this task to indicate their effects on intra-and interservice military activities with additional integration of the associated personnel research of various governmental, educational, business, industrial and private organizations.

THE FIRST MILITARY OPERATIONAL APPLICATION OF ITEM RESPONSE THEORY.

T.A. Warm, LTJG G.J. Edge, and LTJG C.R.Pastene
U. S. Coast Guard Institute
Oklahoma City, Oklahoma

Abstract: In this first military operational application of Item Response Theory, the 9 subtests of the 2 alternate forms of the Coast Guard Warrant Officer Selection Battery were calibrated using item response theory. A linking test composed of items optimally selected from the 2 calibrated forms was constructed, adminstered, and calibrated. All items from the 2 original forms were then put onto the linking test scale, and their Score Information Curves compared. The results demonstrated that in general the subtests of the 2 original forms were not even "weakly" parallel as designed by classical test theory criteria.

Finally, an empirical approach was used to determine the critical point on the ability scales that maximized the value of the test for its intended purpose, and the tests were revised to provide the greatest possible score and test information at that point.

This paper was originally entitled, "The first operational application of item response theory." We then heard that the Educational Testing Service had used item response theory on its English-as-a-second-language test. So we had to change the title to "The first military operational application..."

The Coast Guard Warrant Officer Selection Battery (WOSB) consists of six tests, which may be subdivided into nine subtests. See Table I.

	Test/Subtest	# items	
I.	General Ability		100
	A. Verbal Analogies	40	
	B. Arithmetic Reasoning	30	
	C. Mechanical Comprehension	30	
II.	Coast Guard Knowledge		50
III.	English		50
	A. Grammar	30	
	B. Reading Comprehension	20	
IV.	Math		30
V.	Science		40
VI.	History and Social Studies		30
			300

Table I. Components of the Warrant Officer Selection Battery (WOSB).

The philosophy of the test is to select individuals with the equivalent of 2 years of college. The reasoning behind the philosophy is that Warrant Officers, who come from the enlisted ranks, serve as liaison between the officers, who are generally college graduates, and enlisted personnel, who are generally high school graduates.

The WOSB reflects that philosophy. The General Ability section is analogous to a college entrance examination. The last four sections cover content which is common to most college core curricula. The Coast Guard Knowledge section (section II) might be considered analogous to core curricula in a military academy.

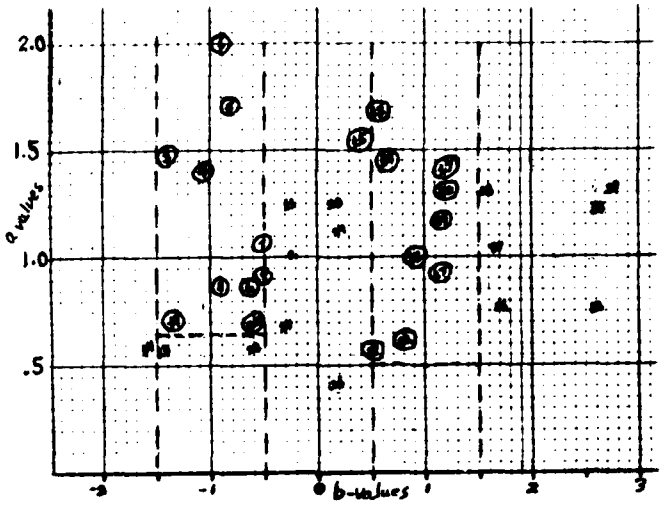


FIGURE 1. Example from Verbal Analogies subtest, Series 8, of selection of items for the "linking" test. Selected items (item numbers circled) were chosen with $.5 \leq |b| \leq 1.5$, and the a-value as high as possible.

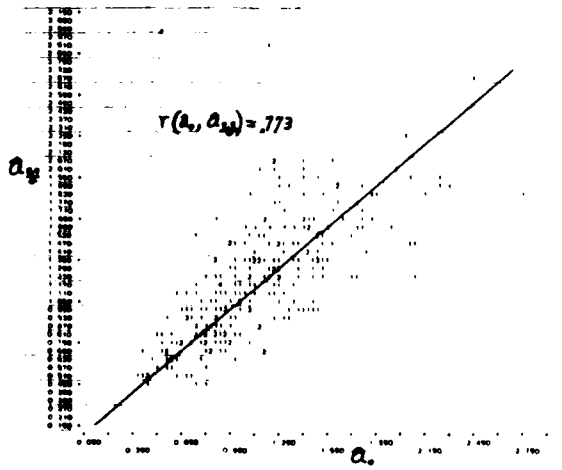


Figure 2. Scatterdiagram of a-value calibration on the linking test (a_o), and on either Series 8 or Series 9, after conversion to the linking test scale ($a_{8,9}$). A perfect correlation is represented by the diagonal line.

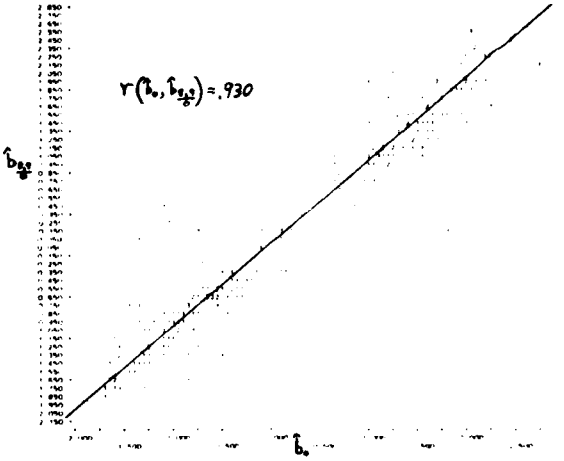


Figure 3. Scatterdiagram of b-value calibration on the linking test (b_o), and on either Series 8 or Series 9, after conversion to the linking test scale ($b_{8,9}$). A perfect correlation is represented by the diagonal line.

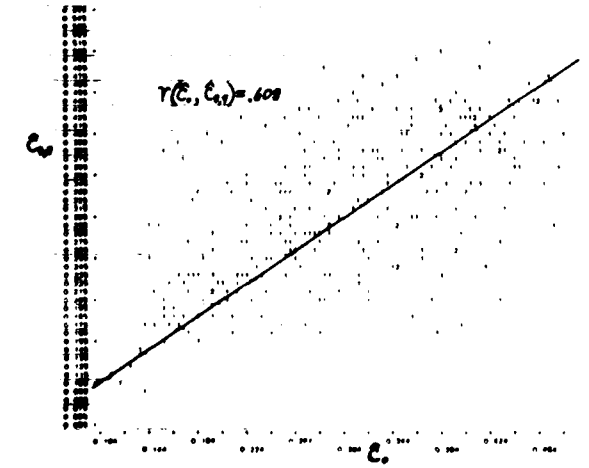


Figure 4. Scatterdiagram of c-value calibration on the linking test ($\hat{c}_o$), and on either Series 8 or Series 9. The correlation is spuriously high due to correlated error introduced by calibration program. A perfect correlation is represented by the diagonal line.

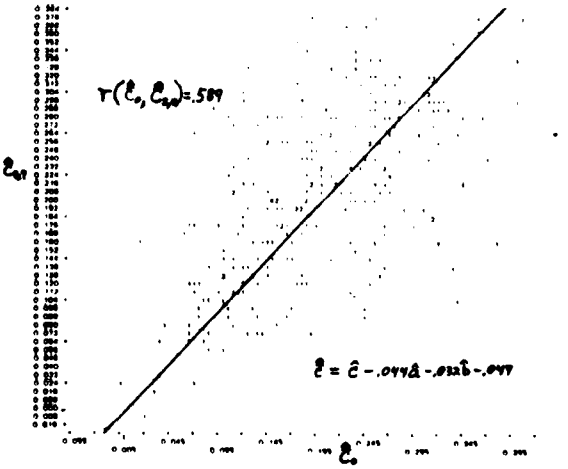


Figure 5. Scatterdiagram of c-values after correction ($\hat{c}$), making them uncorrelated with $\hat{a}$ and $\hat{b}$. Negative $\hat{c}$'s should be adjusted up to zero. The correction formula is shown in the figure.

(974)

The WOSB is administered to about 1300 candidates in November of each year. Until this year there have been two nominally parallel forms of the battery, which were alternated from year to year. The two forms were constructed to be parallel by classical test theory criteria.

This paper is a report of a four year study that began soon after the Coast Guard Institute obtained the technical capability to apply Item Response Theory.

This study was designed in seven phases.

Phase I: Calibrate items of the two "parallel" batteries.

Phase II: Construct, administer, and calibrate a "linking" battery.

Phase III: Equate the scales of the item parameters.

Phase IV: Test the assumptions of parameter invariance and unidimensionality.

Phase V: Evaluate parallelism of the batteries with the Test and Score Information functions.

Phase VI: Construct a battery that maximizes effectiveness to accomplish its purpose.

Phase VII: Evaluate effectiveness of the battery.

We have completed the first six phases of the study. The last phase is awaiting administration of the battery next month.

Each phase was conducted "by the book" whenever practical circumstances permitted. In fact, very few compromises with reality were necessary. Thus, this study is nearly as theoretically pure as can be done.

In Phase I we calibrated the items after the 1977 and 1978 administrations of the two forms of the WOSB. We used the item calibration computer program, OGIVIA-3, which we obtained from Dr. Vern Urry of the Office of Personnel Management. Ree (1978) found that OGIVIA is the most robust of several item calibration programs.

In Phase II we constructed a third form to be used as a "linking" test. The linking test was used to link the ability scales of the original two forms, so that all item calibrations would be on the same scale. To construct the linking test we selected half of the items of each subtest from each of the two forms, according to the item parameters.

First, we plotted the item a-values against the b-values separately for each subtest of each form, representing the points by the item numbers. Figure 1 shows how this was done. Second, we selected half the items from each subtest, according to the following criteria:

(1) half the b-values above zero and half below;

(2) absolute values of the b-values between 0.5 and 1.5; $.5 \leq |b| \leq 1.5$;

(3) a-values as high as possible;

(4) if items were tied on the above criteria, the one with the lower c-value was chosen.

These criteria allow the scale equating process to be as accurate as possible. b-values above 1.5 and below -1.5 were not used, because extreme b-values contain considerable error as do their associated a-values. Only in the Reading Comprehension subtest did we have to violate

these criteria somewhat. The violation was caused by the fact that Reading Comprehension items were blocked into groups that were based on the same reading passage. This grouping gave us very little freedom of item selection in that subtest.

From the selected items we constructed the linking test and administered it in November 1979. We then calibrated the items from that administration. We now had all items calibrated at least once, and one half the items of each test calibrated a second time on the linking test.

Since the item a-values and b-values are invariant except for linear transformations, it was a simple matter to discover the linear transformation needed. (See Warm, 1979, chapter 16.) This is done by setting the means and standard deviations of the separate calibrations of b-values equal. The a-values were similarly transformed. For simplicity we chose the linking test scale as the base scale and transformed all item parameters to that scale. Item c-values are always on the same scale, so no transformation is necessary. We now had all 600 items calibrated on the same scales.

Our next step was to test the item invariance assumption. Figures 2, 3, and 4 are scatter diagrams of the two calibrations of the 300 items on the linking test after equating. Only 275 items are plotted, 25 items having been lost for various reasons, such as non-convergence. In each figure the horizontal axis is the item parameter estimate on the linking test, which we designated Series 0. The vertical axis is the item parameter estimate on either of the two original calibrations, designated Series 8 and Series 9. The subscripts of the variables indicate the calibration series. Table 2 gives the means, standard deviations and intercorrelations of these variables.

Figure 2 shows the bivariate distribution of the two calibrations of the a-values, and represents a correlation of $r(\hat{a},\hat{a}) = +.733$. This correlation is between two estimates. The correlation between the estimate and its true value is the square root of that correlation, or $r(a,\hat{a}) = .856$, which is a very respectable correlation, and typical of those found in monte carlo studies where the true values are known.

Figure 3 is the same as Figure 2, but for the b-values instead. It represents a correlation of $r(\hat{b},\hat{b}) = +.930$, or $r(b,\hat{b}) = +.964$. Usually, monte carlo studies observe $r(b,\hat{b}) \geq .98$, but who in the behavior sciences can complain about a correlation of .964 or even .93. To us it is terrific. As you know eeking those last few correlation points out of real data is extremely difficult, so we are satisfied.

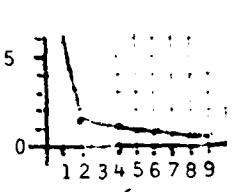
Figure 4 is for the c-values, and represents a correlation of $r(\hat{c},\hat{c}) = +.608$, or $r(c,\hat{c}) = +.780$. No respectable researcher would believe a result that high and we do not. We have two reasons to be skeptical. First, monte carlo studies consistently show correlations between c and $\hat{c}$ to be very low, for example, from $r = .20$ to $r = .35$. Second, other studies have shown that estimates of c-values are highly correlated with the estimates of the a-values and b-values. This means that OGIVIA introduces correlated errors into the estimates of the c-value. So it is the correlated errors that have given us the unbelievable high correlation. Furthermore, the equation

$$r(c,\hat{c}) = \sqrt{r(\hat{c},\hat{c})}$$

assumes that the errors are random and uncorrelated, and in this case we know that they are not.

This reasoning led us to a discovery that we feel may be rather significant. Using the intercorrelations of all the parameters, we were able to compute the multiple regression equation for an estimate of the c-value that is uncorrelated with the a and b parameters.

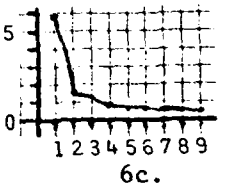
$$\hat{\hat{c}} = \hat{c} - .044\hat{a} - .032\hat{b} - .047 \qquad \text{Equation 1}$$



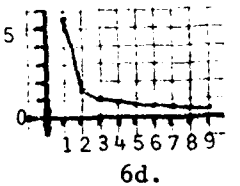
6a.



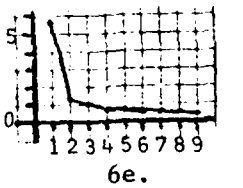
6b.



6c.



6d.

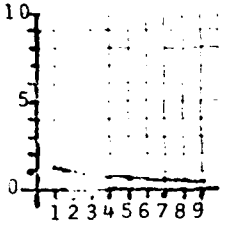


6e.

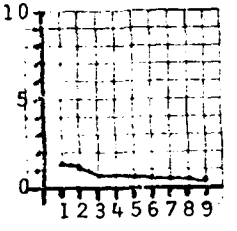


6f.

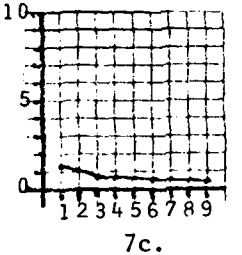
Figures 6a. to 6f. Eigenvalues of first nine factors of six real, 20-item unidimensional tests. The test had 20 items. Each scree test shows one common factor.



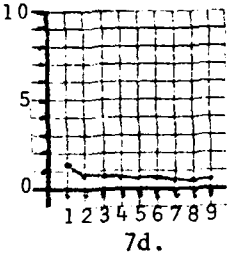
7a.



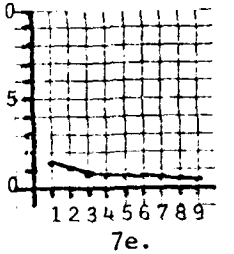
7b.



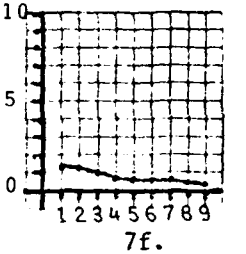
7c.



7d.



7e.



7f.

Figures 7a. to 7f. Eigenvalues of first nine factors of six, 20-item sets of random data. Each scree test shows no common factors.

WARM-5

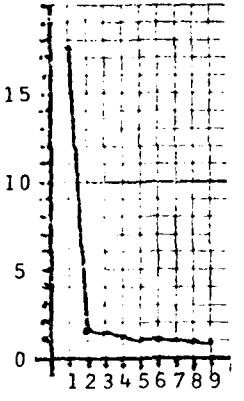


Figure 8. Scree test of pure unidimensional, monte-carlo generated, 60-item test.

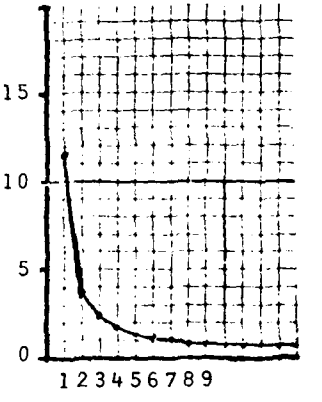


Figure 9. Eigenvalues of 80 multi-dimensional real items. Scree test shows four real factors.

where $\hat{e}$ is the estimate of c that is uncorrelated with $\hat{a}$ and $\hat{b}$, and $\hat{a}$, $\hat{b}$, and $\hat{c}$ are the original estimates of a, b, and c.

Figure 5 shows the scatter diagram of the corrected c-value ($\hat{e}$) on the two calibrations. Interestingly, it represents a correlation $r(\hat{e},\hat{c}) = +.589$, or $r(c,\hat{e}) = +.767$, not much change, but now $\hat{e}$ is not correlated with the $\hat{a}$ or $\hat{b}$ values. It is also interesting that the corrected and uncorrected estimates of c correlate $r(\hat{e},\hat{c}) = .89$. It is difficult to imagine what correlated error might still be in the estimates of c. Thus, we think that $r(c,\hat{e}) = +.767$ is the real relationship within sampling fluctuations. We have not cross-validated the regression weights in Equation 1. However, we have reason to believe that they are rather stable. If we are correct, then this estimate of c($\hat{e}$) will be the first time that any decent estimate has been found. And, as was shown last year (Warm, 1979b), improvements in the estimation of c have the greatest potential of improving our estimates of ability.

The high correlations we have found between item parameter estimates (particularly the a and b parameters) support both the invariance principle and the assumption of unidimensionality.

We also tested unidimensionality with the "scree" test, which is also called the Eigenvalue Test (Warm, 1979a). The scree test is a way to discover the number of common factors in the data. A single factor is a sufficient condition for unidimensionality.

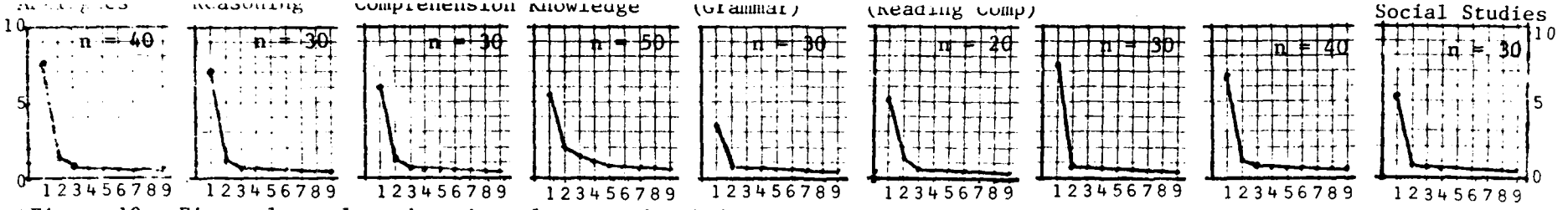
The scree test is done by performing a common factor analysis (CFA) on the matrix of interitem tetrachoric correlations with the largest row absolute value in the diagonal (communality). Eigenvalues are then plotted against the factor ordinal. Unidimensionality is indicated if only the first factor eigenvalue is much larger than the others. We will not get into an explanation of factor analysis, because a few examples are all that is necessary to learn how to interpret the eigenvalues for the scree test. To illustrate consider Figures 6a to 6f, which show the scree test for data for six real 20 item tests that were shown to be essentially unidimensional by three different tests of unidimensionality (McBride and Weiss, 1974). In each figure the first factor eigenvalue is much larger than the rest, and all others are small and about the same. Therefore, they each have only one real factor.

In Figures 7a to 7f are the scree tests for six sets of random data (McBride and Weiss, 1974). By definition each of these should have no real factor, and in fact we see that that is the case. Figure 8 shows the scree test for Urry's test data for OGIVIA. This data is monte carlo generated to have exactly one pure factor. Indeed, the scree test shows just one very strong factor. One more example. Figure 9 shows the scree test for 80 real items from the WOSB. We took 20 verbal analogy items, 20 arithmetic reasoning, 20 mechanical comprehension, and 20 CG Knowledge. We chose these four sets of items because they are as different in content as we could imagine. The first three are traditional so-called "pure" abilities, and CG knowledge is a hodge-podge of knowledge items which should have little to do with the other three. We wanted to see if the scree test would identify four "real" factors. Indeed, it did as we can see in Figure 9.

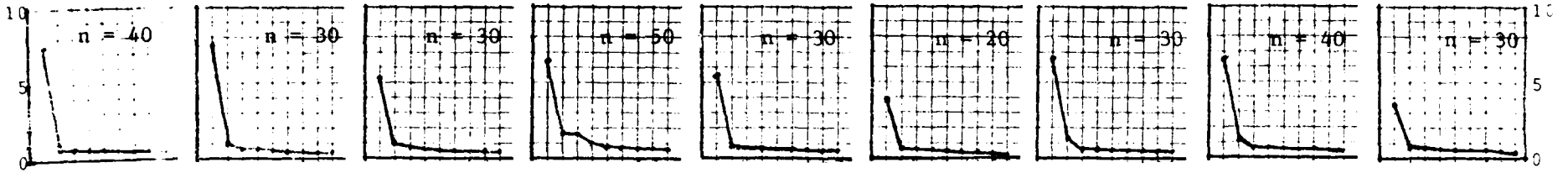
There are two traditional rules of thumb for determining the number of "real" factors. One rule is that any factor with a percent of total variance of greater than 10% is a real factor. That means that any factor is real if the eigenvalue is greater than the number of items divided by ten. Guttman's rule is that any factor is real if it has an eigenvalue greater than one. Neither rule is consistent with what we know about the number of factors in these examples. But then neither rule was developed to be used with data that has as much error as does multiple-choice test data, so that the factor analysis process can capitalize on chance.

Therefore, we have developed our own rule of thumb. We claim a factor is real only if it's eigenvalue is greater than 1.75. This rule works very well with the examples as you can verify

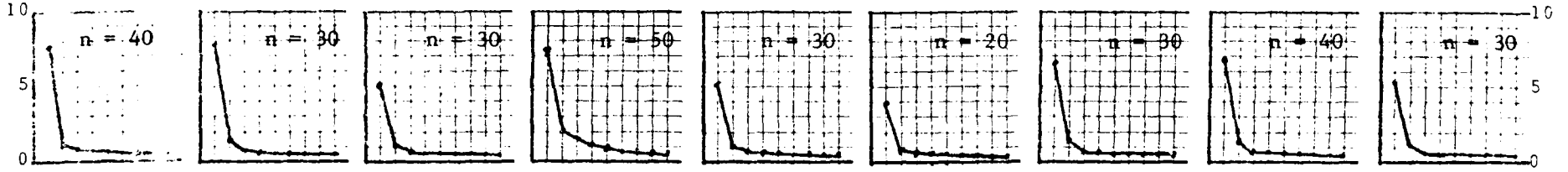
(97P)



↑Figure 10. Eigenvalues plotted against factor ordinal for each subtest of the WOSB, Series 8.



↑Figure 11. Eigenvalues plotted against factor ordinal for each subtest of the WOSB, Series 9.



↑Figure 12. Eigenvalues plotted against factor ordinal for each subtest of the WOSB, Series Ø.

↓Table 2. Means, standard deviations, and correlations of estimated item parameters from two separate calibrations of 275 items from the USCG Warrant Officer Selection Battery.

	$\hat{a}_{0/0}$	$\hat{b}_{0/0}$	$\hat{c}_{0}$	$\hat{a}_{8/9\ 0}$	$\hat{b}_{8/9\ 0}$	$\hat{c}_{8/9}$	$\hat{\hat{c}}_{0}$	$\hat{\hat{c}}_{8/9}$
MEANS	1.1673	0.2962	0.3031	1.2117	0.2970	0.3179	0.1952	0.2081
STANDARD DEVIATIONS	$\hat{a}_{0}$	$\hat{b}_{0}$	$\hat{c}_{0}$	$\hat{a}_{8/9\ 0}$	$\hat{b}_{8/9\ 0}$	$\hat{c}_{8/9}$	$\hat{\hat{c}}_{0}$	$\hat{\hat{c}}_{8/9}$
	0.4060	1.1777	0.0927	0.4241	1.1703	0.0918	0.0819	0.0814
CORRELATION MATRIX	$\hat{a}_{0}$	$\hat{b}_{0}$	$\hat{c}_{0}$	$\hat{a}_{8/9}$	$\hat{b}_{8/9}$	$\hat{c}_{8/9}$	$\hat{\hat{c}}_{0}$	$\hat{\hat{c}}_{8/9}$
	1.0000	0.0191	0.2201	0.7330	0.0436	0.1306	0.0222	-0.0408
	0.0191	1.0000	0.4189	-0.0511	0.9300	0.3336	0.0099	-0.0401
	0.2201	0.4189	1.0000	0.1282	0.3472	0.6078	0.8915	0.4963
	0.7330	-0.0511	0.1282	1.0000	-0.0552	0.1836	0.0088	0.0032
	0.0436	0.9300	0.3472	-0.0552	1.0000	0.4140	-0.0444	0.0193
	0.1306	0.3336	0.6078	0.1836	0.4140	1.0000	0.5061	0.8952
	0.0222	0.0099	0.8915	0.0088	-0.0444	0.5061	1.0000	0.5892
	0.0408	-0.0401	0.4963	0.0032	0.0193	0.8952	0.5892	1.0000

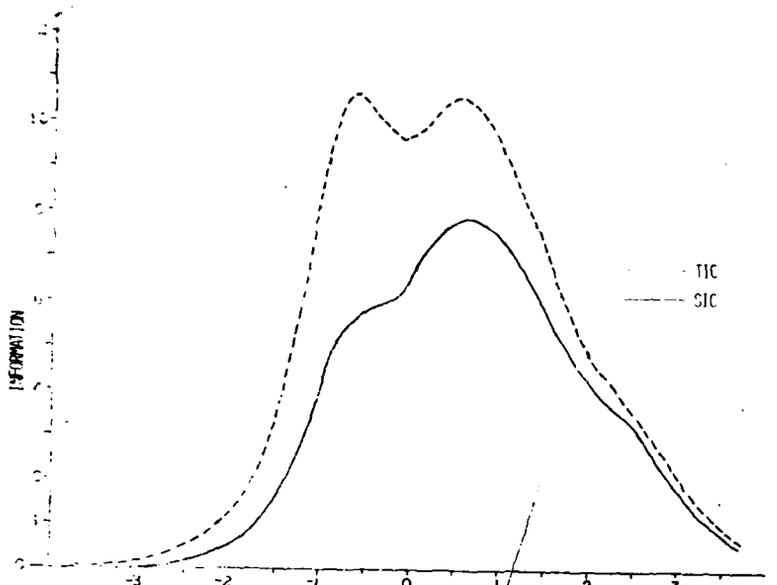


FIGURE 13. TEST INFORMATION AND SCORE INFORMATION CURVES OF SERIES 8 VERBAL ANALOGIES SUBTEST

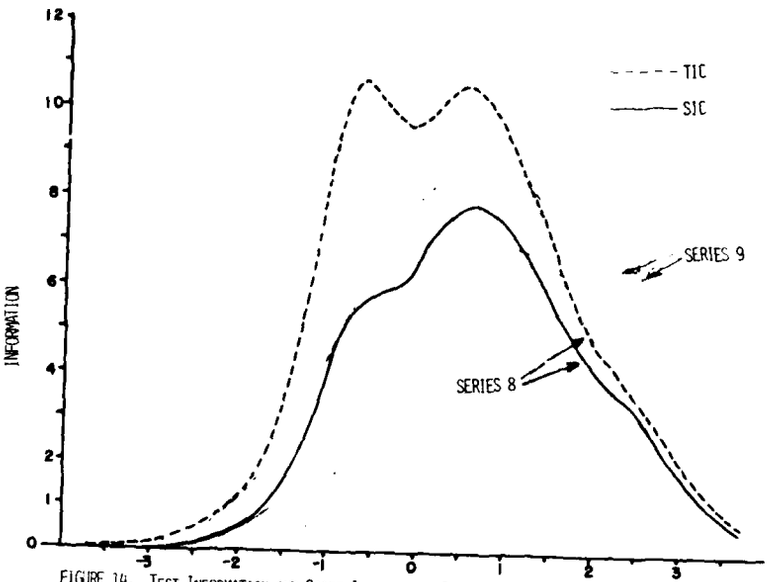


FIGURE 14. TEST INFORMATION AND SCORE INFORMATION CURVES OF SERIES 8, AND SERIES 9 VERBAL ANALOGIES SUBTESTS.

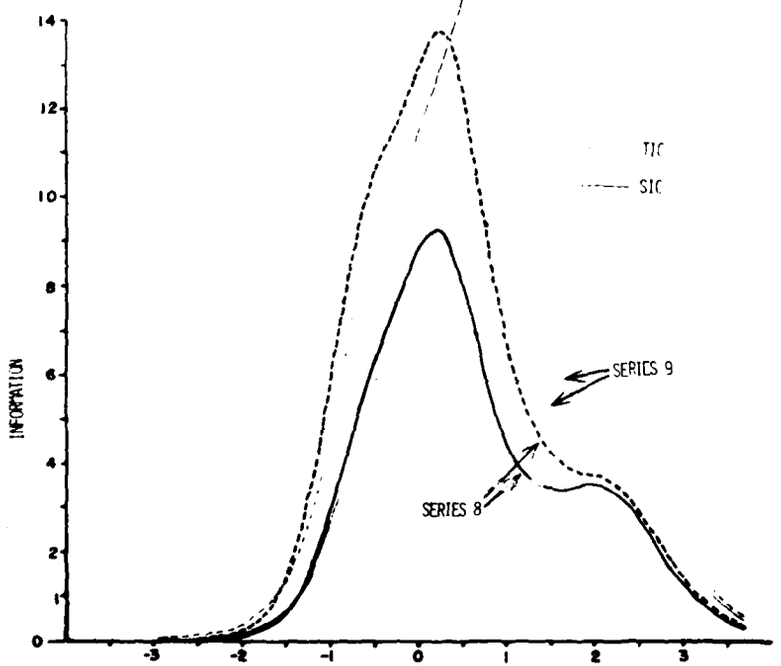


FIGURE 15. TEST INFORMATION AND SCORE INFORMATION CURVES OF SERIES 8, AND

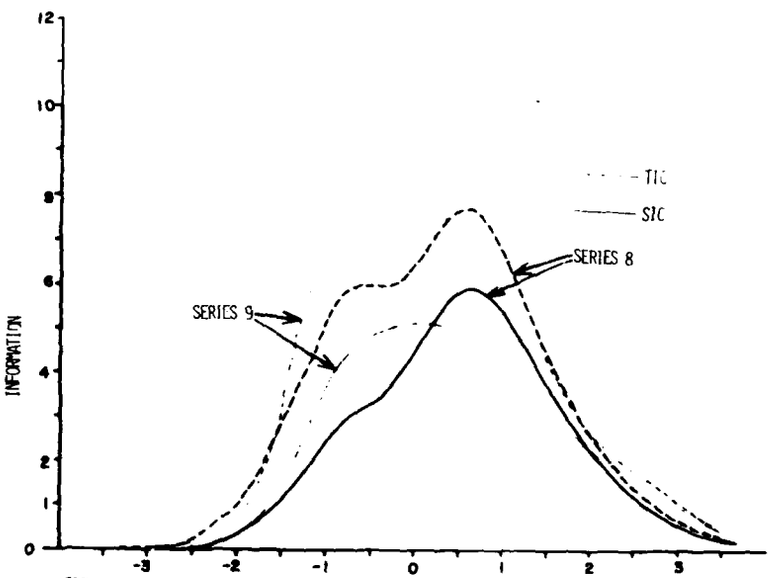


FIGURE 16. TEST INFORMATION AND SCORE INFORMATION CURVES OF SERIES 8, AND SERIES 9 MECHANICAL COMPREHENSION SUBTESTS.

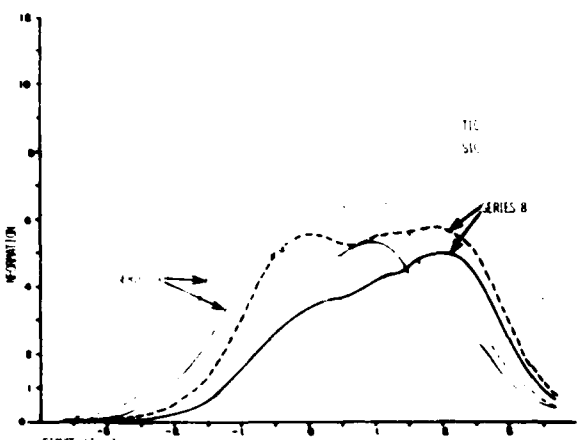


FIGURE 17. TEST INFORMATION AND SCORE INFORMATION CURVES OF SERIES 8, AND SERIES 9 (FIRST GRADE NUMERICAL SUBTEST)

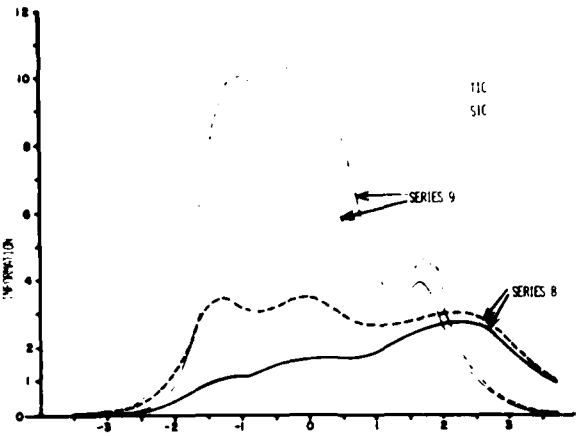


FIGURE 18. TEST INFORMATION AND SCORE INFORMATION CURVES OF SERIES 8, AND SERIES 9 (ENGLISH GRAMMAR SUBTESTS)

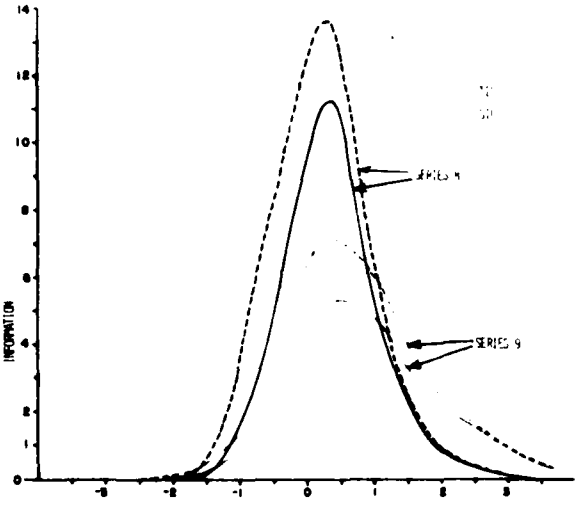


FIGURE 19. TEST INFORMATION AND SCORE INFORMATION CURVES OF SERIES 8, AND SERIES 9 ENGLISH READING COMPREHENSION SUBTESTS

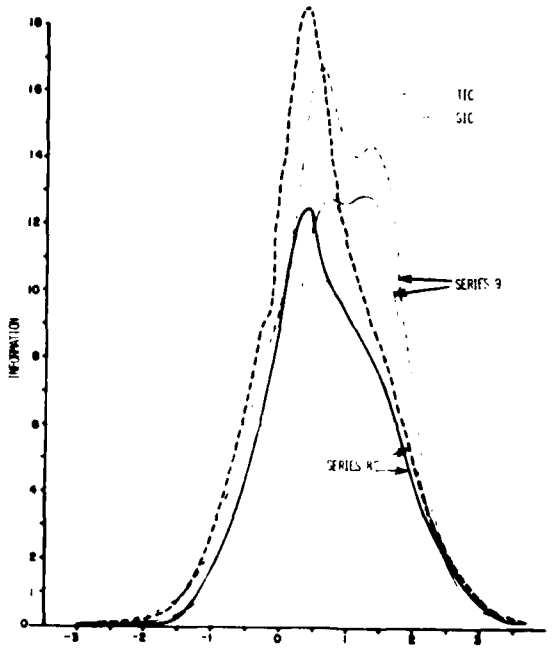
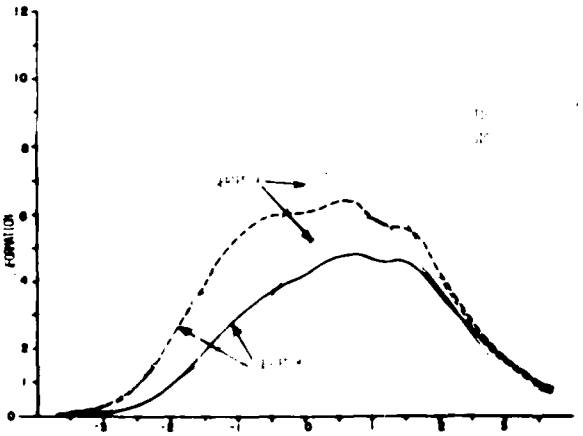
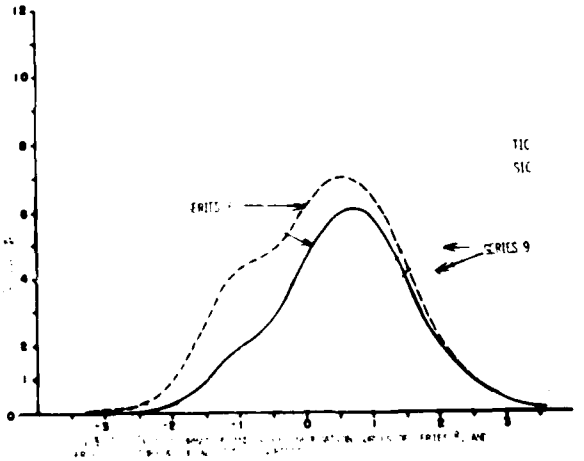


FIGURE 20. TEST INFORMATION AND SCORE INFORMATION CURVES OF SERIES 8, AND SERIES 9 MATHEMATICS SUBTESTS





for yourselves. With that small bit of experience we can do the scree test on the three administrations of the nine subtests of the WOSB.

Figures 10 and 11 are for the two original forms. Figure 12 is for the linking test. In each graph we find one strong factor. Only in the Coast Guard Knowledge subtest is there a second factor with an eigenvalue greater than 1.75, and that other factor is weak.

With the parameter invariance and scree test evidence to support us, we concluded that all of the subtests are in fact unidimensional or unidimensional enough to apply Item Response Theory.

In Phase V we investigated the parallelism of the original forms. Samejima (1977) has distinguished between two types of parallelism--i.e. strongly parallel and weakly parallel.

Strongly parallel means that there is a one-to-one correspondence between items and item parameters on the tests. This means that for each item on one test with particular parameters there is an item on the other test with those same parameters.

Weakly parallel means that the test information functions of the parallel tests are nearly identical, although there may be no parallelism at the item parameter level.

In conventional tests such as we have here there is no advantage of strong parallelism over weak parallelism, and, since strong parallelism is an unnecessarily stringent standard, we have chosen weak parallelism as our criterion. Moreover, since the operational use of these tests is always with number-right scoring, we are particularly concerned with the test's Score Information Curve rather than Test Information Curve. However, we looked at both.

Figure 13 shows the Test Information (TIC) and Score Information Curves (SIC) for the Series 8 Verbal Analogies test. The higher, dashed curve is the TIC. It represents the maximum information that can be gotten from the test. The solid line is the SIC. The SIC is always lower than the TIC, because of the inefficiency of the number-right score as a measure of ability. At high abilities it is usually pretty close to the TIC, but at low abilities the SIC falls away from the TIC dramatically, because there is a lot of guessing going on there.

If the Series 9 test is parallel, then its SIC will be about the same shape as the Series 8 SIC. Figure 14 graphs the information curves for both Series 8 and Series 9 Verbal Analogy subtests. It is obvious that the two forms are not even weakly parallel with respect to either the TIC or SIC. Figures 15 to 22 show the information curves for the other subtests. Clearly, the classical test theory criteria failed to make any of the subtests parallel.

In Phase VI we constructed a battery from the calibrated items.

Since about 7% of the candidate population are selected and promoted each year, we decided to build the new battery so that it would best help to select the top 7% of the candidates. That suggested to us that we should build the test with high test information at the point on the ability scale that cuts off 7% of the candidates. Such a test would cause all those above the point to score high on the test, and those below the point to score very low. That was the general idea. But, if only 7% scored high on <u>each</u> subtest, then probably no one would score high on <u>all nine</u> subtests. So we began looking for a way to determine the point to build with high information, so that about 7% would score high on all nine subtests.

All of the subtests are correlated to some extent, and we know those correlations. There is probably some theoretical way, using those correlations in a multi-variate (9 variable) distribution to select the point. If so, it is extremely complicated, and we could not find out how, nor could we find anyone who knew how. So we resorted to an empirical method. We obtained the standardized test scores of candidates from a previous administration. We then picked a standardized score, and counted how many candidates got above that score on all the tests. If more than 7% did so, we picked a higher score, and counted again. If less than 7% scored above the point on all tests, we picked a lower score and counted again. We continued this iteration process until we found the score above which exactly 7% of the candidates scored on all tests. That score corresponded to .4 of a standard deviation above the mean. It is equivalent to a Navy Standard Score of 54, and on the ability scale, $\Theta = .4$.

We then computed the item information at $\Theta = .4$ for all 600 items in our calibrated bank. We selected for the new battery the 300 items which had the highest information at $\Theta = .4$ for each subtest. Those items became the new battery.

The new battery will have extremely high test and score information at the most critical point, which will maximize the test's effectiveness in accomplishing its purpose.

Perhaps next year we can report to you what we found.

REFERENCES

McBride, J.R., and Weiss, D.J. A Word Knowledge Item Pool For Adaptive Ability Measurement. Research Report 74-2, Psychometric Methods Program, Dept. of Psychology, University of Minnesota, Minneapolis, MN.

Ree, M. Estimating item characteristic curves. AFHRL-TR-78-68. Personnel Research Division, Brooks AFB, TX 1978.

Samejima, F. A use of the Information Function in Tailored Testing. Applied Psychological Measurement, vol 1, #2, Spring 1977, pp.233-247.

Warm, T.A. A Primer of Item Response Theory. Technical Report 940279, 1979. U.S. Coast Guard Institute. National Technical Information Service, AD-A063072.

Warm, T.A. Analysis of and Comments on Dr. Ree's Paper: The Effects of Errors in Estimating of Item Characteristic Curve Parameters. Proceedings of the 1979 Conference of the Military Testing Association, Navy Personnel Research and Development Center, San Diego, CA, 1979.

WARNOCK, A. Timothy., WATERS, Brian K., Air War College Associate Programs, Air University, Maxwell AFB, Montgomery, Alabama.

OBJECTIVES-BASED TESTING IN THE AIR WAR COLLEGE ASSOCIATE PROGRAMS
(Tue A.M.)

This paper examines the use of objectives-based, norm referenced, multiple choice examinations in the non-resident senior professional military education program of the Air War College. First, the authors describe the purpose, organization, methodology, evaluation, and curriculum of the Correspondence and Non-resident Seminar Programs; i.e., the Air War College Associate Programs. Also, they explain why an objectives-based, multiple choice examination is used to measure student achievement. Next, they describe the procedures for test construction, administration, and analysis, and examine problems of test reliability and score standardization. Then, they discuss the results of the examination program over a period of two years. The authors compare the examination responses of Seminar students to those of Correspondence students and offer possible explanations of differences in performance.

OBJECTIVES-BASED TESTING IN THE AIR WAR COLLEGE ASSOCIATE PROGRAMS[1]

A. Timothy Warnock and Brian K. Waters

Air War College[2]
Maxwell AFB, AL 36112

INTRODUCTION

The Mission.

The Air War College is the senior professional military education institution of the United States Air Force. Its mission is to prepare senior officers for key command and staff assignments in which they may be responsible for developing, managing, and employing air power as a component of national security. The School of Associate Programs, as an integral part of the Air War College, provides a professional military education to senior officers unable to attend the School of Resident Programs.

Organization.

The School of Associate Programs is organized into three departments, i.e., Seminar Studies, Correspondence Studies, and Curriculum Development. Both the Department of Seminar Studies and Department of Correspondence Studies enroll non-resident students, utilize the same curriculum materials, have virtually the same eligibility requirements for enrollment, and have similar requirements for course completion. Students enrolled in Seminar Studies meet in groups weekly for forty weeks. The non-resident seminar methodology gives students a chance to discuss the materials with members of a group. Each member has a chance to venture his/her opinions and reactions, state his/her ideas, and compare his/her experiences with other members of the seminar. On the other hand, Correspondence students generally study over a period of about one year individually and independently to complete the course.

Students.

Students eligible to enroll in the Associate Programs must be senior majors, lieutenant colonels, colonels, or civilians with GS-13 or higher rank. As shown in Table I, the majority of Associate Programs' students are active duty Air Force officers. Air National Guard and Air Force Reserve officers make up a substantial proportion of the enrollment, as do active and reserve officers from other armed services. Civilians make up the remainder of the enrollment. Air National Guard and Air Force Reserve officers make up a much greater proportion of the Correspondence enrollment than of Seminar enrollment.

---

[1]Paper presented at the Military Testing Association Annual Meeting, Toronto, Canada, 28 October 1980.

[2]Views or opinions expressed or implied are the authors and are not to be construed as carrying official sanction of the Air University (ATC) or the Department of the Air Force.

Curriculum.

The Associate Programs' curriculum, prepared by the Department of Curriculum Development, consists of two volumes of twenty chapters each. Each chapter contains a bibliography, an introduction describing the purpose of the chapter and the nature of its materials, a set of behavioral objectives, a set of suggested questions for study and discussion, and from six to fourteen readings selected from books, periodicals, monographs, and government publications. Table II is an outline of the overall curriculum objectives.

Students enrolling in the Associate Programs undertake an in-depth study of essential, critical issues in international and military affairs. They are expected to develop knowledge, skills, and attitudes significant to the profession of arms, particularly aerospace power. In general, problems that must be dealt with in higher echelons of command and staff are so complex and diversified and frequently include so many intangible and imponderable factors as to preclude solution by formula or precedent. Consequently, the School of Associate Programs neither provides approved solutions to problems nor suggests what the solutions should be. Instead, students are expected to develop possible solutions through individual study, research, analytical reasoning, and, in the case of seminar students, thorough discussion and debate.

Members of the Department of Curriculum Development are professional writers with graduate education and extensive writing experience in history, military and international affairs, and political science. Each writer is responsible for certain chapters and assembles the readings and bibliography; writes behavioral objectives, introductory materials, and study questions; and constructs examination items. The curriculum is revised on an annual basis, with each revision being referred to as an edition. Presently, the curriculum writers are preparing the 15th Edition, which will become available to students in July 1981.

Evaluation.

The School of Associate Programs requires students to write a research paper on each volume of study and to complete four multiple choice examinations satisfactorily. These are the chief instruments of evaluation of student achievement. The School relies on the research reports to evaluate higher levels of cognitive learning and communications abilities. However, the primary interest of this paper is the use of objectives-based, norm-referenced, multiple choice examinations for evaluation purposes.

In the Associate Programs, the chief purpose of examinations is to evaluate student comprehension of the curriculum. In addition, examinations serve as incentives to spur students to study diligently, to explore complex topics in depth, and to discern views and arguments presented in the required readings of the texts. Also, examinations, with adequate and timely feedback, aid students in learning the material and in demonstrating their knowledge. Aggregate data on student performance on examinations provide a basis for decisions and planning in curriculum and in administration.

(PP)

The present system of objectives-based, multiple choice, closed-book examinations was instituted in January 1978 in place of multiple choice, open-book examinations. Although there was some discussion of essay testing at the time, a decision was made to use multiple choice examinations for several reasons. The writers were already experienced in writing multiple choice questions. In addition, test review, administration, and analysis procedures were already in place. Thus, the multiple choice examinations could be controlled, graded, and otherwise handled more efficiently. The curriculum writers received 20 hours of additional training in writing behavioral objectives and related test items.

ESSAY EXAM PILOT PROGRAM

However, objections to closed book, multiple choice examinations continued on philosophical and practical bases. One proposal was to switch to essay examinations. Arguments for essay examinations included the following:

a. Students generally perceive essay tests as more appropriate to graduate level professional military education than multiple choice examinations.
b. Students would gain freedom of expression in answering items.
c. Students could better demonstrate higher levels of cognitive learning.
d. Students could demonstrate written communications abilities.
e. Students would feel better being evaluated with essay exams.

As a result, the School of Associate Programs conducted an essay exam pilot program to assess the feasibility and desirability of introducing this type evaluation. Fifty-two Seminar students were administered an essay examination during the period 28 April-1 May 1980. The examination was given in controlled circumstances in a single two-hour session. It was strictly closed book; students were not permitted to use any texts, references, or notes. The exam consisted of four essay questions, with students having to answer the first item and any one of the three remaining items.

Workload to Prepare and Grade Essay Exams.

A crucial question of the pilot program was the assessment of the feasibility of implementing an essay examination without overloading the faculty, particularly in grading and feedback workload. The workload may be split conceptually into two segments: that dealing with exam preparation which is most sensitive to the number of questions prepared; and that dealing with grading, which is most sensitive to the number of examinees, the number of responses per examinee, and the number of faculty graders.

Exam preparation time included the preparation of study questions, construction of four test items, and the writing of a grading guide for each item. The workload for these procedures totaled 97 manhours. The resource investment per exam question for preparation was 24.3 manhours per question.

(989)

The grading of the fifty-two examinations required the expenditure of 251 manhours; thus each examination required 4.93 manhours to grade. If the pilot program exam grading workload were linearly extrapolated to the number of students who typically take a multiple choice examination, the total would be 2032 manhours, 288 mandays, or 57.6 manweeks. The Department of Seminar Studies has only six officers available as exam graders. If they graded a typical two-response essay examination for a typical complement of seminar students, they would have to work full-time for ten weeks to complete grading one set of exams. There are presently four multiple choice examinations given each year. If essay examinations were substituted, almost all the time of the faculty monitors would need to be devoted to grading answers. The Department of Correspondence Studies would face similar problems.

Student Response to Essay Exams.

Each response was graded independently by two faculty members, using a standardized assessment sheet. Responses were assigned adjectival grades of outstanding, excellent, satisfactory, or unsatisfactory. Any differences between the two graders on a given response were reconciled by consultation between the graders. If no agreement could be reached, a third grader independently graded the paper. The team of graders then agreed on a grade for the entire examination. Results by question and for the examination are given in Table III.

Student responses were not of as high a quality as faculty graders had expected. Generally, graders perceived that students did not exercise their freedom of expression on the essay questions nor clearly demonstrate higher levels of cognitive learning. Also, because students had been told that they would not be graded on grammar, they made numerous grammatical errors which detracted from the overall quality of answers on the essay examination.

Other disadvantages of essay examinations were perceived:

a. Essay examinations can only sample very small, selected portions of the curriculum. Multiple choice examinations permit sampling of virtually all the curriculum content.
b. Grading of essay examinations would be far more costly than is the case for multiple choice examinations. This cost arises from the manpower and time required for grading essay questions. With multiple choice examinations, the Extension Course Institute, another organization of the Air University, automatically grades the tests by computer at much lesser cost.
c. Danger of test compromise would be greater with essay than with multiple choice examinations.

Rejection of Essay Exams.

The School of Associate Programs rejected the use of essay examinations primarily because of the tremendous workload and cost involved in grading the examinations. Other factors contributing to this decision were the ease of administration of multiple choice examinations, the successful use of research reports to measure certain types of student achievements not reached easily by multiple choice questions, and the lack of

significantly better performance of students on essay examinations in comparison with multiple choice tests.

OBJECTIVES-BASED, NORM-REFERENCED, MULTIPLE CHOICE EXAMS

Writing Objectives.

Each writer constructs the behavioral objectives and related test items for his/her chapters. Objectives are stated in terms of Bloom's taxonomy of learning (Benjamin S. Bloom, Ed., Taxonomy of Educational Objectives: Handbook I--Cognitive Domain, New York, David McKay Co., Inc., 1956). While some objectives are knowledge level, most are comprehension or application level. Test items are related directly to samples of behavior specifying what students should be able to do and giving a good indication of what the test items will be like. Each chapter has a general objective, two to five specific objectives, and ten to fifteen subobjectives or samples of behavior.

The objectives are representative samples of the learning outcomes and content areas to be measured. They provide general direction toward learning levels that cannot be completely achieved. Learning outcomes depend on the cumulative effect of diverse learning experiences that can be organized and integrated in many different ways. Thus, the examinations are objectives-based. However, the test questions measure concepts, trends, comparisons, and complex interrelationships not easily understood without repeated review of the text materials.

Constructing Items.

The writers face several problems in writing items at the higher levels of cognitive learning. First, the examinations must measure student understanding of complex, interdependent relationships from as many different angles and with as much representativeness as possible. In addition, the curriculum materials on which the examinations are based are diverse, complex, and written by many different authors with contradictions in viewpoints frequently occurring. Thus, writers face the problem of testing students on an opinion of a topic where more than one argument appears in the text or where multiple perceptions are likely to be found among students. Also, materials often deal with specifics, but questions written to test specific detail are generally at a knowledge level of learning. Consequently, writers must generalize to write questions at the comprehension or higher levels. Finally, the Air War College philosophy is that there are no school solutions, however, the use of closed-end questions appears to give school sanctions to one answer for each significant issue.

Curriculum writers construct seven items per chapter, each item being directly related to a sample of behavior. A trained educator reviews and edits objectives and items. After revision, objectives and items go to a test review committee made up of all the writers, the test reviewer, and one member each from the Departments of Seminar and Correspondence Studies. Finally, two examinations per volume are prepared for the printer. The first examination covers the first ten chapters of the volume and the second exam covers the remaining ten chapters.

Taking the Test.

Whenever a student completes ten chapters of a volume, he takes an examination under the direction of a local test control officer (TCO). The test control officer sends the completed answer sheet to the Extension Course Institute for grading. ECI sends the student a feedback form identifying each missed item by reference to a related sample of behavior and giving the overall exam grade of satisfactory or unsatisfactory. If the student scored unsatisfactory, ECI also sends the TCO a retake examination for the student to accomplish.

Students may receive an unsatisfactory on both the initial and the retake examination and still continue in the course. However, they must make satisfactory scores on the second half of the volume examination to pull up any unsatisfactory scores on the first half to complete the volume successfully. Students failing both examinations or failing to score high enough on the average of both examinations for a satisfactory score are automatically dropped from the enrollment list.

Scoring the Exam.

Examinations may be interpreted with reference to a group, i.e., norm-referenced, or with reference to a criterion, i.e., criterion-referenced. Given the nature of the Associate Programs' curriculum, the mission of the Air War College, and the difficulty of establishing a valid criterion of performance, the examinations are norm-referenced. Nevertheless, until August 1980, student performance on examinations had been judged by a set criterion. Students had to answer 60 percent of the items correctly to receive a satisfactory score on an examination.

When closed-book, multiple choice examinations were first instituted, the Associate Programs set the cutting score at 70 percent. The reasoning behind this policy was that the Air War College is a professional military graduate education and students should be held to a minimal criterion on examinations. However, a pilot testing program showed that the examinations were entirely too difficult for a 70 percent cutting score. Subsequently, the cutting score was lowered first to 50 percent and then later raised to 60 percent with the goal of 70 percent never being reached.

Basically, the problem is that examinations cannot be written consistently at a given level of difficulty and reliability when less than 20 percent of the text materials and 12 percent of the objectives and questions are common from one edition to another. Too much extraneous variance can creep into different forms of the same examinations, as well as in different tests on different materials, to permit fair scoring simply on the percentage of correct answers. For example, 68 percent of the Seminar respondents passed the second part of the Volume I examination (see Table (IV). The Seminar students' average score on that examination, including retakes, was 63 percent and the Correspondence students' average score was 69 percent. Due to the high difficulty level of the examination, a decision was made to multiply student scores by 1.2, so that students would not be unfairly penalized. The result was an inflated average score on the 13th Edition examination compared to what the score would have been without the multiplier (see Table (IV).

Recognizing the problem with a criterion-referenced cutting score, the Dean of the Associate Programs has directed that a system of standardized scoring be developed and applied to the examinations in the 14th Edition. Presently, the School is working out a system which will be fair, while providing immediate feedback to students. Many of the problems of instituting a closed-book multiple choice examination derived from the imposition of the criterion-referenced cutting score for what is basically a norm-referenced examination.

Handling Reliability Problems.

Test analyses of student responses in the 12th Edition revealed rather low internal reliabilities. Reliabilities ranged from .57 on the first Volume I examination to .70 on the third examination. Using individual item analyses, several items on each examination were rewritten in an effort to increase test reliability. The only examination in which reliability did not increase significantly as a result of rewrites of individual items was the fourth one. The changing reliabilities may be seen in Table V.

In the 12th Edition, 42 to 50 percent of the enrolled Seminar students took the examinations (see Tables IV and VI). Both Correspondence and Seminar students had to request that examinations be sent to local TCOs. Whenever a request was received, ECI would mail the examination. In an effort to encourage a greater proportion of Seminar students to take examinations, all 13th Edition examinations for every Seminar student were mailed simultaneously. Seminar students no longer had to request examinations but knew that the TCO was holding the tests for their convenience. The change in procedure resulted in eight to eleven percent more Seminar students taking the 13th Edition examinations than did so in the 12th Edition (see Tables IV, VI, and VII).

However, examinations sent out to TCOs cannot be replaced in any feasible, cost-effective manner. Previously, rewritten examinations could be printed and sent to students easily since the Associate Programs maintained a test bank in which items could be edited letter by letter, word by word, line by line, or item by item. Rewritten examinations after initial test analyses would be available only to a limited number of Correspondence students, and the number of students (less than 200) who would benefit from a rewritten examination was too low to justify rewriting and reprinting the tests. Thus, the only changes in examinations which could be made in the 13th Edition was the elimination of certain items from scoring. This particular effect of sending all examinations to Seminar students in bulk instead of on request was not considered at all in the decision-making process.

Explaining Differences in Seminar and Correspondence Performance.

The School of Associate Programs instituted the objectives-based, multiple choice examination program in January 1978. As of 1 June 1980, results on all the 12th Edition examinations and on Volume I, 13th Edition examinations were available. A comparison of Seminar and Correspondence responses to examinations showed a number of interesting differences. These differences are summarized in Tables VI and VII.

One obvious difference is that a greater percentage of Correspondence students passed examinations with satisfactory scores than did Seminar students. In addition, on every examination, Correspondence students consistently made higher average scores than did Seminar students. Another evident trend is that the percent of Seminar students taking examinations is consistently larger than that of Correspondence students. On the first three examinations, where responses are practically complete, six to ten percent more Seminar students took the tests than did Correspondence students. Correspondence students, who do not have time invested in attending seminars, are less likely to take the examinations than are Seminar students.

These differences between Seminar and Correspondence students apparently rest in the ways in which the two populations approach the tests. One reasonable proposition is that only highly motivated Correspondence students, who are likely to do well, take the examinations. On the other hand, many Seminar students with less likelihood of doing well probably take the examination because of the time they have already invested in the meetings. Peer pressure may be another factor encouraging Seminar students to complete examinations. Another reasonable proposition is that Seminar students may tend to rely on seminar discussions as a primary means of studying for examinations. On the other hand, Correspondence students study alone and must rely entirely on what they read about the subjects. Seminar students may often not read the material as carefully or prepare as adequately as Correspondence students and thus may not do as well on the examinations.

CONCLUSION

Several approaches may be taken to writing test items, but an objectives-based system emphasizes behaviors that can be measured on examinations. The quality of test items depends on the goals and objectives on which they are based. The goals of the educational process provide the foundation for deriving specific objectives which define expected student behavior. Behavioral objectives state what students should be able to do and, when shared with students, as in the Associate Programs, give them an indication of the related test items. Of course, objectives for learning concepts, trends, and other complex knowledge of the Associate Programs' curriculum are much less precise and therefore more difficult to measure than are objectives for learning job tasks and performance, such as in a training program.

To measure student achievement, the School of Associate Programs instituted multiple choice, objectives-based examinations. However, there was a belief that essay examinations would be a better form of evaluation. Therefore, an essay pilot examination was conducted to see how feasible the use of essay tests would be. The outcome showed that, in comparison to multiple choice examinations, essay tests would be entirely too costly. Because of the unacceptable time and manpower requirements, essay examinations were rejected as a feasible form of evaluation of student achievement.

Over a period of two and one half years, the examination program revealed differing responses from Seminar and Correspondence students. For example, although Correspondence students tend to make better average

scores, a greater proportion of Seminar enrollees take the examinations and go on to graduate. Evidently, these are two different populations and probably should not be treated the same in every instance. Each program has its strong points, although a question remains as to whether the differences rest in characteristics of the populations or the differing methodologies for non-resident studies.

Perhaps the greatest problem of the examination program was the setting of a criterion score based on the percentage of correct answers. A basically academic course of study does not lend itself to criterion-referenced testing, as the School of Associate Programs discovered. Objectives stated in behavioral terms are useful to students studying for the examination, but students cannot completely master all objectives, as required by the instructional systems development model. Instruction at the Air War College Associate Programs emphasizes professional education rather than training. Educational objectives are goals or targets rather than minimum levels that all must achieve, and students are not expected to attain fully the objectives of the curriculum. This philosophy recognizes individual differences, has greater utility for planning, and allows us to challenge all our students, including the very best. Given this philosophy, norm-referenced examinations are more suitable than are criterion-referenced tests for the Associate Programs.

(96)

TABLE I

ENROLLMENTS AS OF 30 JUNE 80

	Correspondence Studies			Seminar Studies			Associate Programs	
	#	% of Corr.	% of Total	#	% of Sem.	% of Total	#	% of Total
AF Active	1035	58%		771	76%		1806	64%
AF Reserve Component	470	26%		49	5%		519	19%
Other Services	261	14%		61	6%		322	11%
Civilian	33	2%		132	13%		162	6%
Total	1799	100%	64%	1013	100%	36%	2812	100%

WARN-10

TABLE II

AIR WAR COLLEGE ASSOCIATE PROGRAMS' CURRICULUM OBJECTIVES

Volume I. Military Environment and Decision Making

To comprehend national, international, and military factors which affect U.S. security and decision making.

Unit 1. "The National and International Environments"

To comprehend significant factors and events in the national and international environments which affect national security and well-being.

Unit 2. "Threat Assessments"

To comprehend potential Communist external military and non-military threats to U.S. security and well-being.

Unit 3. "Formulation of National Security Policy"

To comprehend the formulation of national security policy and the role and influence of major participants.

Unit 4. "Management of Defense Resources"

To comprehend factors contributing to leadership, command, and management of human resources in Department of Defense; analytical techniques for decision making; and principal methods used in defense resource management.

Volume II. Military Strategy and Aerospace Power

To comprehend selective dimensions of USAF doctrine and strategy; the capabilities, doctrines, and strategies of strategic and general purpose forces; and strategic appraisals of and force posture in unified command areas.

Unit 1. "Military Doctrine, Strategy, and Capabilities"

To comprehend significant influences on the doctrine, strategy, and capabilities of strategic and general purpose forces, particularly air forces, for strategic nuclear and theater warfare.

Unit 2. "Support Functions"

To comprehend the capabilities and employment of those forces which support the defense posture of the United States.

Unit 3. "Theater Appraisals"

To comprehend strategic factors and military posture of U.S., Allied, and significant Communist forces in unified command areas.

TABLE III

ADJECTIVAL GRADES ASSIGNED FOR PILOT ESSAY EXAMINATION

	OUTSTANDING	EXCELLENT	SATISFACTORY	UNSATISFACTORY
Question 1:	5	13	32	2
Question 2:		2	5	
Question 3:	2	12	18	3
Question 4:		1	9	
Overall Exam	1	14	33	4

(998)

TABLE IV

SEMINAR AND CORRESPONDENCE RESPONSES

	12th Edition								13th Edition			
	Volume I				Volume II				Volume I			
	Exam 1 Chs. 1-10		Exam 2 Chs. 11-20		Exam 3 Chs. 1-10		Exam 4 Chs. 11-20		Exam 1 Chs. 1-10		Exam 2 Chs. 11-20*	
	Sem	Corr	Sem	Corr	Sem	Corr	Sem	Corr	Sem	Corr	Sem	Corr
# enrolled	617	782	622	782	642	642	643	642	848	829	851	829
# scored	297	310	263	282	320	259	272	182	473	170	453	68
% tested	48	40	42	36	50	40	42	29	56	20	53	8
% passed	90	94	68	82	79	91	92	98	88	89	97**	91**
Ave Score	70	75	63	69	67	71	74	78	69	72	82	86

* All scores for these examinations are adjusted by a factor of 1.2.

**Only 88% of the seminar students and 84% of the correspondence students would have passed this test if the factor had not been added into the scores.

(GOOD)

TABLE V

TEST STATISTICS

12th Edition

	Exam 1		Exam 2		Exam 3		Exam 4	
	Original	Rewritten	Original	Rewritten	Original	Rewritten	Original	Rewritten
N=	153	392	438	198	378	177	141	105
R=	.57	.68	.68	.72	.70	.74	.63	.62
ADI=	.19	.23	.28	.29	.24	.23	.21	.19
AEI=	72.52	73.82	63.88	65.81	72.42	70.70	78.16	76.99

WARN-14

13th Edition

	Exam 1 Original	Exam 2 Original
N=	504	453
R=	.68	.66
ADI=	.25	.24
AEI=	71.38	72.06

N/number of respondents

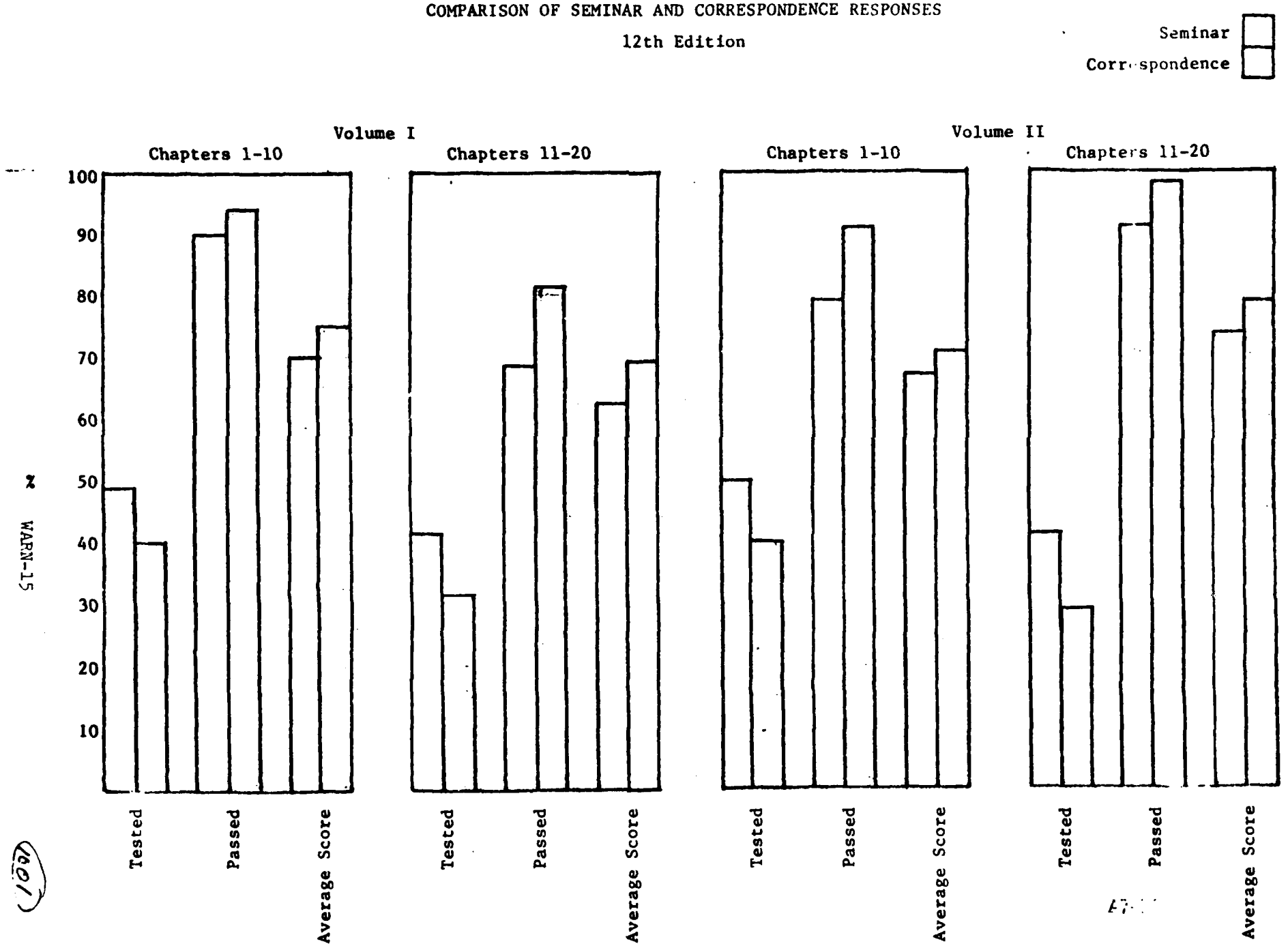
R/reliability

AEI/average ease index

ADI/average discrimination index

COMPARISON OF SEMINAR AND CORRESPONDENCE RESPONSES

12th Edition



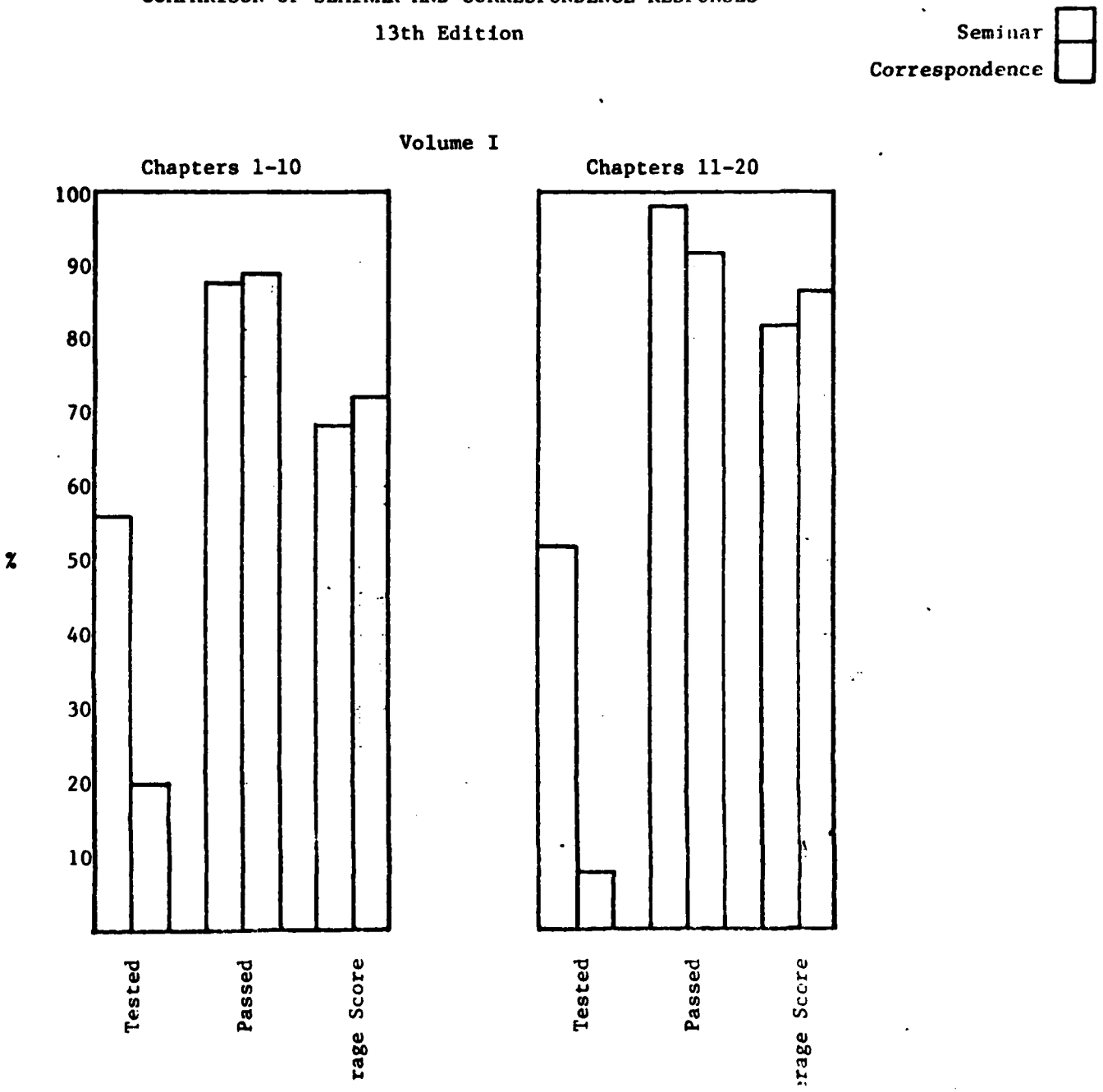
WARN-15

(101)

TABLE VII

COMPARISON OF SEMINAR AND CORRESPONDENCE RESPONSES

13th Edition



WEEKS, Joseph L., Air Force Human Resources Laboratory Manpower and Personnel Division, Brooks Air Force Base, Texas.

VALIDITY COMPARISION OF VERBAL AND NONVERBAL MEASURES OF VOCATIONAL INTERESTS (Thu A.M.)

Previous research has resulted in both a verbal measure of vocational interests, the Vocational Interest-Career Examination (VOICE), and a nonverbal measure of vocational interests, the Pictorial Interest Inventory (PII). Each instrument is designed forluse in the assignment of Air Force enlistees to job specialties. The need for the development of separate measures issued from observations which indicated wide differences in the reading skills of enlistees. The study reported in this presentation concerns the relative validity of these two measures as indicators of job satisfaction. Multiple linear regression analysis was used to assess the validity of these measures for different occupational categories within different reading ability groups. Special steps were taken to control for the confounding effects of extraneous factors. Although the study is still in progress, preliminary data analysis has indicated that both the verbal and nonverbal measures of vocational interests are valid indicators of job satisfaction.

COMPARISON OF THE VALIDITY OF THE PICTORIAL INTEREST INVENTORY AND THE VOCATIONAL INTEREST-CAREER EXAMINATION

By

Joseph L. Weeks
Manpower and Personnel Division
Air Force Human Resources Laboratory

I. Introduction

One of the major organizational goals of the United States Air Force is to obtain maximum returns from personnel investments. The high costs of personnel procurement and training in combination with growing budgetary constraints have increased the importance of selecting the "right person" for the "right job." The "right person," in this context, is the qualified applicant; the "right job" is the job which can be adequately performed by the individual and which results in job satisfaction. The underlying rationale is that the assignment of qualified recruits to satisfying jobs will result in lower attrition. Empirical support for the relationship between job satisfaction and attrition is abundantly available in research literature (Mobley, Griffeth, Hand, & Meglino, 1979; Porter & Steers, 1973; Price, 1977).

For many years the Air Force has relied on aptitude information as a major component in the assignment of personnel to jobs. Actual job assignments are based on a combination of enlistee aptitudes, job aptitude requirements, and the needs of the Air Force. Enlistees enter occupational specialties by two basic routes; assignment to a guaranteed job prior to enlistment or occupational assignment during basic military training. The satisfaction of job aptitude requirements is a prerequisite regardless of the route followed. Assignments to guaranteed jobs occur at the recruiting stage. Recruits select an occupational specialty from those available either on the basis of prior job experience or written job descriptions. Job assignments occurring during basic training are accomplished with the help of career guidance counselors. Enlistees are provided with a list of job specialties in the career/aptitude area in which they enlisted (i.e., mechanical, electronics, general, or administrative), and they indicate a preference among those jobs that are available. In the case of both guaranteed job assignments and basic training job assignments, enlistees frequently establish occupational preferences in the absence of comprehensive, job description information. Furthermore, when such information is available, the associated technical wording may be incomprehensible to enlistees with limited work experience or inadequate reading skills. Such a situation may result in occupational assignments that eventually lead to job dissatisfaction.

1005

For a number of years, private industry has used measures of vocational interest as a supplement to aptitude information in making job assignmment decisions. These measures reduce the necessity of the individual knowing the specific tasks associated with various jobs, register the individual's interests, and estimate the relative magnitude of those interests. Reference to such information for job assignment decisions increases the probability of placing individuals in satisfying jobs.

A general occupational interest inventory, the Vocational Interest-Career Examination (VOICE), has been developed for the Air Force and was designed for use during pre-assignment career counseling. The items comprising the inventory were developed by the Educational Testing Service under contract (Echternacht, Reilly, & McCaffrey, 1973). Through subsequent research, inventory scales were refined and validity and normative information were produced (Alley, Berberich, & Wilbourn, 1977; Alley, Wilbourn, & Berberich, 1976; Berger & Berger, 1977). With this inventory, recruits indicate their preferences for a variety of different activities. Written descriptions of job activities, as contained in the VOICE, require a certain amount of reading comprehension and verbal ability on the part of the individual completing the inventory. This has caused concern as to the value of the instrument for determining the vocational interests of enlistees who possess limited verbal or reading abilities. In response to this concern, research was initiated to design and construct a nonverbal interest inventory.

Wilbourn and Alley (1978) developed the Pictorial Interest Inventory (PII) for assessing the vocational interests of enlistees with limited reading skills. The inventory consists of 180 35mm slides which depict tasks associated with job specialties in each of the four aptitude areas (i.e., mechanical, electronics, general, and administrative). It was concluded in the developmental study that photographs of job tasks represent a viable alternative to verbal descriptions in eliciting the vocational interests of job applicants. Furthermore, on the basis of available information, it was concluded that the PII was an effective means of identifying the vocational preferences of individuals with low verbal or reading abilities. However, an empirical study providing evidence of the validity of the PII was not conducted at that time due to the lack of job satisfaction criterion. Such a study is necessary prior to the use of the inventory in an operational setting. In addition, it should be established whether or not the PII is more valid than the VOICE for the assessment of the vocational interests of personnel having limited reading skills. The objectives of the present study were to provide evidence of the validity of the PII and to compare the validity of the PII and VOICE for enlistees who differ in reading ability.

II. Method

The interest inventories were administered in the 1975-76 time frame. During that period, 11,863 Air Force enlistees were tested with the PII and VOICE during basic military training. The interest inventories were administered in a counterbalanced design. The 180 slides of the PII were projected on a screen one at a time, allowing 15 seconds for the examinees to indicate their preference for the activities represented. Examinees coded item responses to both the PII and the VOICE in a standardized format (i.e., Like, Indifferent, or Dislike). Total testing time was approximately 1 1/2 hours: 45 minutes for the PII and 45 minutes for the VOICE.

Demographic information and aptitude scores from the Armed Services Vocational Aptitude Battery (ASVAB) were obtained for each subject from personnel records maintained by the Technical Services Division of the Air Force Human Resources Laboratory. The ASVAB is an aptitude test battery used by the Department of Defense (DoD) for the selection and classification of military enlistees (Weeks, Mullins, & Vitola, 1975). It requires approximately 2 hours of testing time and yields four different aptitude scores (i.e., mechanical, electronics, general, and administrative) generally referred to as aptitude indexes (AI's).

The criterion employed in this study was job satisfaction. A survey was mailed to a subset (i.e., N = 8,944) of the total group administered the PII and VOICE. At the time the survey was mailed, enlistees had performed their jobs for a period of 1 to 2 years. The return rate was approximately 57%. Job satisfaction was determined from responses to the survey item:

How satisfied are you with your present Air Force job specialty?

Item responses were weighted 800 = Very Satisfied, 600 = Moderately Satisfied, 400 = Moderately Dissatisfied, and 200 = Very Dissatisfied. The theoretical midpoint of the scale was 500.

The job specialties to which enlistees were assigned were also determined by survey. Job specialties were initially categorized into 20 occupational groups based on the DoD Occupational Conversion Table (U.S. Department of Defense, 1977), and descriptive statistics representing job satisfaction were determined separately for each of the 20 DoD occupational groups. In subsequent analyses, the 20 DoD occupational groups were further collapsed into six homogeneous occupational categories based on a factor analytic study performed by Watson, Alley, and Southern (1979).

After eliminating all those subjects with either incomplete or invalid data, a total of 4,482 cases were available for analysis. To insure that this sample was representative of Air Force enlisted populations, the sample average on the general AI of the ASVAB (i.e., a measure of general mental ability) was compared to that for the total 1975 and 1976 Air Force enlisted populations. The differences among

these means ranged from one to two score points. Therefore, the sample was considered to be representative of typical Air Force populations in terms of general intelligence.

The primary objective of this study was to provide evidence of the validity of the PII. To attain this objective, the first step was to identify subjects in the total sample who had worked on their jobs for a period no greater than 1 year. A sample consisting of 1,856 cases was derived and served as the basis for obtaining estimates of the relationship between the PII and job satisfaction. This sampling procedure was followed for two reasons. First, it was necessary to insure comparability between validity estimates for the PII and VOICE. The VOICE was validated against job satisfaction reported after 1 year on the job. Second, this procedure should result in more accurate estimates of the validity of the PII since the most dissatisfied enlistees may have separated from the Air Force by the end of the second year. Multiple linear regression analyses were performed on this sample to determine the contribution of the PII in the prediction of job satisfaction. The effects of occupational category, sex, and aptitudes were also taken into account in the analyses.

The second objective of the study was to determine the relative validity of the PII and the VOICE for individuals who possess limited reading skills. The total sample (N = 4,482) was employed for these analyses. The use of the total sample would mean that job satisfaction data collected from 1 to 2 years on the job would serve as criteria. The probable effect of this procedure will be that the absolute magnitude of the relationships between job satisfaction and vocational interests as measured by the inventories will be underestimated. However, it is the relative rather than the absolute validity of the inventories which is of interest; therefore, the total sample was considered suitable for this purpose. Subjects in this sample were assigned to different reading ability groups on the basis of their scores on the general AI of the ASVAB. Previous research (Mathews, Valentine, & Sellman, 1978) had established a high relationship (i.e., r = .79) between scores on the general AI and reading grade levels as determined by commercial reading tests. Two reading ability groups were established in this manner. Individuals scoring above the 89th percentile on the general AI formed the high reading ability group (N = 1,186) and those scoring below the 60th percentile formed the low reading ability group (N = 1,517). Multiple linear regression analyses were accomplished within each reading group separately to determine the relative validity of the PII and the VOICE in the prediction of job satisfaction. As in the previous regression analyses, the effects of occupation, sex, and aptitudes were taken into account.

III. Results

In describing the results of this study, consideration will first be given to the descriptive statistics associated with each interest

(100%)

inventory and the job satisfaction criterion. The validity evidence derived for each of the interest inventories will be discussed last.

The 180 slides of the PII represent 11 independent interest scales which are common to males and females. One additional scale, shop skills, is derived for males only. Table 1 presents average scale scores for males and females, scale score ranges, and a nominal description of each scale. The differences between the scale averages for males and females were tested for significance using Bonferroni t values. The Bonferroni t technique was employed in order to maintain a constant error rate while performing multiple comparisons (Miller, 1966). Except for Shop Skills and Air Traffic Control, all mean differences are statistically significant. Males scored higher in the area of Electronics, Aircraft Weapons Maintenance, Pararescue, Law Enforcement, Heavy Duty Equipment Operator, Cable/Power Line Maintenance, and Automobile/Aircraft Mechanics. Females scored higher in the areas of Office/Administration, Medical/Dental, and Food Services.

The form of the VOICE employed in this study consisted of 400 items from which 18 independent scale scores were derived (Alley, 1978). These scales measure general interests in a variety of different content areas. Table 2 presents average scale scores for males and females, scale score ranges, and a nominal description for each VOICE scale. As with the PII, the differences between VOICE scale average for males and females were tested for significance using Bonferroni t values. Except for the Science and Drafting scales, all mean differences are statistically significant. Males scored higher in the interest areas of Electronics, Heavy Construction, Outdoors, Mechanics, Law Enforcement, and Marksman. Females scored higher in the areas of Office Administration, Medical Service, Aesthetics, Food Service, Audiographics, Mathematics, Agriculture, Teacher/Counselor, Craftsman, and Automated Data Processing.

Table 3 presents job satisfaction score means and standard deviations for each DoD occupational group for males and female separately. For males, mean job satisfaction scores for all occupational groups except 640 - Armaments and Munitions; 820 - Material Receipt, Storage, and Issue, and 83-A - Security Police are above the theoretical midpoint of the scale (i.e., 500). For females, mean job satisfaction scores for all occupational groups except 6X0 - General Mechanic and 820 - Material Receipt, Storage, and Issue are above the theoretical midpoint. Comparing the job satisfaction of males and females within each occupational group, we find that males were significantly less satisfied than females in 3X0 - Miscellaneous Medical and Dental Specialties and 510 - Administration. Females were significantly less satisfied than males in 6X0 - General Mechanic, 602 - Aircraft Accessories Mechanic, and 720 - Utilities Maintenance. As a group, females were significantly more satisfied with their jobs than were males. However, the mean satisfaction score for the total group (i.e., mean = 578, N = 4,482) exceeds the scale midpoint and suggests that Air Force enlistees are generally more satisfied than dissatisfied with their jobs.

Table 4 presents the results of multiple linear regression analyses which were accomplished to determine the magnitude of the predictive relationship between pre-service interests as measured by the PII and satisfaction reported after 1 year on the job. Briefly, this analytical technique involves the computation of an R^2 for a set of predictor variables (full model), and another R^2 for some subset of these predictor variables (restricted model). The difference between the two R^2's is tested for significance. If no significant difference is found between the two R^2's, the interpretation is that those variables in the full model that are not in the restricted model add nothing in predicting the criterion and can be discarded from the predictor set without affecting validity (Ward & Jennings, 1973). All regression analyses were accomplished separately within each of six homogeneous occupational categories (OC). Occupational categories were established on the basis of a factor analytic study of the 20 DoD occupational groups (Watson, Alley, & Southern, 1979) previously employed in the distributional analyses. The original 20 DoD groups were not used in these regression analyses because of small sample sizes for some occupational groups.

The first comparison in Table 4 tests the R^2's resulting from a full model which consists of separate predictor vectors for each sex by occupational category and a restricted model which employs only the unit vector. The comparison indicates that predictor vectors representing sex and occupational group membership make a significant contribution to the prediction of job satisfaction. This is not surprising in view of the sex and occupational group differences displayed in the distribution of job satisfaction scores. This comparison established a baseline model for evaluating the separate contributions of aptitudes and interests in the prediction of job satisfaction. The second comparison in Table 4 tests the significance of the predictive contribution of aptitudes to the baseline model. Although aptitudes as measured by the ASVAB add significant predictive variance to the baseline model, its practical contribution is so small as to be negligible. The third comparison in Table 4 tests the predictive contribution of the PII to the baseline model. Clearly, the PII makes both a significant and practical contribution to prediction when variables representing sex and occupational group membership are already in the predictor system.

Table 5 presents the results of multiple linear regression analyses accomplished separately in low and high reading ability groups to determine the relative validity of the PII and VOICE. Job satisfaction reported after 1 to 2 years on the job served as the criterion. In the analyses for both reading ability groups, the baseline model used for testing the separate effects of the PII and VOICE consisted of separate predictor vectors for the ASVAB and each sex by occupational category. The ASVAB was included in the baseline model because previous analyses indicated that it made a significant contribution to prediction even though a small one.

1010

In the low reading ability group, the first comparison indicates that the baseline model makes a significant contribution to prediction. The second and third comparisons indicate that pre-service interests as measured by either the PII or the VOICE add signifiant and unique variance to the baseline model in the prediction of job satisfaction. In the high reading ability group, the pattern of results are basically the same. Pre-service interests as measured by either the PII or the VOICE add significant and unique predictive variance to the baseline model. These comparisons clearly indicate that job satisfaction is related to pre-service interests as measured by either the PII or the VOICE; however, the major purpose of these analyses was to compare the relative validity of the PII and VOICE for subjects having limited reading skills. In the low reading ability group, the R^2 for the full model including the PII (R^2 = .1440) is practically equivalent to the R^2 for the full model including the VOICE (R^2 = .1525). Evidently, the VOICE is just as useful as the PII for registering the pre-service interests of subjects with low reading skills when job satisfaction is the criterion for validity. Furthermore, a comparison of the R^2's for the VOICE and PII models in the high reading ability group indicates that the VOICE is slightly more valid than the PII for subjects with high reading skills.

IV. Summary and Conclusions

In summary, the questions which served as the basis for this study were whether or not the PII was a valid instrument for registering the pre-service vocational interests of Air Force enlistees and whether the PII was more valid than the VOICE for enlistees having limited reading skills. Job satisfaction reported after 1 to 2 years on the job served as the criterion. Distributional analyses of both PII and VOICE scale scores indicated significant differences in pre-service interests depending on sex group membership. Analyses of job satisfaction scores indicated significant differences in satisfaction depending on both sex and occupation group membership. Regression analyses clearly indicated that the PII was a valid measure of the vocational interests of Air Force enlistees when satisfaction with the job assignment after 1 year was used as the criterion. Additional regression analyses indicated that both the PII and VOICE were valid measures of vocational interests regardless of the reading abilities of Air Force enlistees completing the inventories. Furthermore, analyses accomplished in the low reading ability group indicated that the separate validities of the PII and the VOICE were practically equivalent. In the high reading ability group, the data indicated that the VOICE adds slightly more to prediction than does the PII.

On the basis of these results, it is concluded that the VOICE is the preferred measure of pre-service vocational interests for Air Force enlistees. In addition to its demonstrated validity, it is easier to administer than the PII and it provides a greater range of vocational interest information.

REFERENCES

Alley, W.E. Vocational Interest-Career Examination: Use and application in counseling and job placement. AFHRL-TR-78-62, AD-A063 657. Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, October 1978.

Alley, W.E., Berberich, G.L., & Wilbourn, J.M. Development of factor-referenced subscales for the Vocational Interest-Career Examination. AFHRL-TR-76-88, AD-A046 064. Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, June 1977.

Alley, W.E., Wilbourn, J.M., & Berberich, G.L. Relationships between performance on the Vocational Interest-Career Examination and reported job satisfaction. AFHRL-TR-76-89, AD-A040 754. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, December 1976.

Berger, F.R., & Berger, R.M. Vocational Interest-Career Examination: Norming and standardization on a nationwide high school sample. AFHRL-TR-77-69, AD-A047 762. Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, September 1977.

Echternacht, G.J., Reilly, R.R., & McCaffrey, P.J. Development and validity of a vocational and occupational interest inventory. AFHRL-TR-73-38, AD-774 573. Lackland AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, December 1973.

Mathews, J.J., Valentine, L.D., Jr., & Sellman, W.S. Prediction of reading grade levels of service applicants from Armed Services Vocational Aptitude Battery (ASVAB). AFHRL-TR-78-82, AD-A063 656. Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, December 1978.

Miller, R. Simultaneous statistical inference. San Francisco, McGraw-Hill, 1966.

Mobley, W.H., Griffeth, R.W., Hand, H.H., & Maglino, B.M. Review and conceptual analysis of the employee turnover process. Psychological Bulletin, 1979, 86(3), 493-522.

Porter, L.W., & Steers, R.M. Organizational, work, and personal factors in employee turnover and absenteeism. Psychological Bulletin, 1973, 80, 151-176.

Price, J.L. The study of turnover. Ames: Iowa State University Press, 1977.

United States Department of Defense, Office of the Assistant Secretary of Defense (M&RA). Occupational Conversion Table. DoD 1312.1.DA PAM 611-11. Washington, DC: U.S. Government Printing Office, 1977.

(1012)

Ward, J.H., & Jennings, E. Introduction to linear models. Englewood Cliffs, NJ: Prentice-Hall, 1973.

Watson, T.W., Alley, W.E., & Southern, M.E. Initial development of operational composites for the Vocational Interest-Career Examination. Proceedings of the 21st Annual Conference of the Military Testing Association. San Diego, CA: October 1979.

Weeks, J.L., Mullins, C.J., & Vitola, B.M. Airman classification batteries from 1948 to 1975: A review and evaluation. AFHRL-TR-75-78, AD-A026 470. Lackland AFB, TX: Personnel Research Division, Air Force Human Resouces Laboratory, December 1975.

Wilbourn, J.M., & Alley, W.E. Pictorial Interest Inventory development. AFHRL-TR-78-36, AD-A060 089. Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, August 1978.

TABLES

Table 1. PII Mean Scale Scores[a] by Sex

PII Scales	Range	Midpoint	Female Mean	Male Mean
1 Electronics	20-60	40	37	40*
2 Office Administration	20-60	40	42	33*
3 Medical/Dental	20-60	40	38	32*
4 Shop Skills[b]	20-60	40	-	36
5 Aircraft Weapons Maintenance	20-60	40	32	39*
6 Pararescue	17-51	34	31	34*
7 Law Enforcement	10-30	20	18	19*
8 Food Service	7-21	14	11	9*
9 Heavy Duty Equipment Operator	8-24	16	12	15*
10 Cable/Power Line Maintenance	8-24	16	12	14*
11 Air Traffic Control	11-33	22	25	25
12 Automobile/Aircraft Mechanics	11-33	22	16	21*

[a]Means are rounded to nearest whole number. Total sample consists of 4,482 cases (male N = 3,426; female N = 1,056).
[b]Shop Skills is scored for males only.
*Mean difference significant at the .05 level.

Table 2. VOICE Mean Scale Scores[a] by Sex

VOICE Scales	Range	Midpoint	Female Mean	Male Mean
1 Office Administration	20-60	40	37	31*
2 Electronics	20-60	40	30	39*
3 Heavy Construction	20-60	40	27	33*
4 Science	20-60	40	36	37
5 Outdoors	15-45	30	36	37*
6 Medical Service	20-60	40	39	32*
7 Aesthetics	15-45	30	31	25*
8 Mechanics	15-45	30	23	30*
9 Food Service	15-45	30	25	20*
10 Law Enforcement	15-45	30	26	28*
11 Audiographics	10-30	20	21	20*
12 Mathematics	12-36	24	22	21*
13 Agriculture	15-45	30	30	28*
14 Teacher/Counselor	10-40	25	21	19*
15 Marksman	7-21	14	10	15*
16 Craftsman	7-21	14	11	9*
17 Drafting	7-21	14	13	13
18 Automated Data Processing	7-21	14	14	13*

[a]Means are rounded to nearest whole number. Total sample consists of 4,482 cases (male N = 3,426; female N = 1,056).

*Mean difference significant at the .05 level.

Table 3. Job Satisfaction Score Means[a] and Standard Deviations by Occupational Group and Sex

Occupational Group	Score Range[b]	Males			Females		
		N	Mean	SD	N	Mean	SD
Electronic Equip Rpr (1X0)	200-800	270	607	177	42	605	192
Radio/Radar Equip Rpr (100)	200-800	275	613	164	71	580	154
Comm and Intel Specs (2X0)	200-800	19	589	210	9	645	126
Radar & Air Traf Cont (220)	200-800	107	630	217	29	665	200
Medical/Dental (3X0)	200-800	74	627	187	53	701	144*
Medical Care (300)	200-800	101	651	161	76	658	155
Technical and Allied (4X0)	200-800	64	635	179	41	664	156
Admin & Clerks (5X0)	200-800	411	585	172	215	603	175
Administration (510)	200-800	310	526	185	188	577	182*
General Mech (6X0)	200-800	129	637	163	12	484	223*
Gen Acft Mech (600)	200-800	249	574	180	35	583	181
Acft Engine Mech (601)	200-800	59	597	200	27	519	198
Acft Accessories Mech (602)	200-800	195	567	178	54	508	199*
Armaments/Munitions (640)	200-800	121	484	190			
Utilities Maintenance (720)	200-800	206	621	179	26	500	224*
Fire Fighter (780)	200-800	127	661	150			
Misc Services/Supplies (8X0)	200-800	145	542	198	47	528	182
Material Receipt/Storage (820)	200-800	86	484	171	69	478	193
Security Police (83A)	200-800	413	487	199			
Law Enforcement (83B)	200-800	65	627	178	63	657	170
	Total	3,426	575	189	1,056	589	188*

Note. Groups 640, 780, and 83A consist of occupations restricted to male entrants.

[a]Means and standard deviations are rounded to the nearest whole number.

[b]200 = Very Dissatisfied, 400 = Moderately Dissatisfied, 600 = Moderately Satisfied, 800 = Very Satisfied.

*Mean difference is significant at the .05 level.

Table 4. Results of Regression Analyses Demonstrating the Validity of the PII

Comparison	Models[a]		R^2				
	Full	Restricted	Full	Restricted	DF1	DF2	F
1	OC x Sex	Unit	.0481	.0000	11	1844	8.46**
2	OC x (Sex + ASVAB)	OC x Sex	.0671	.0481	24	1820	1.54*
3	OC x (Sex + PII)	OC x Sex	.1274	.0481	72	1772	2.24**

[a]Predictor models consist of the following predictor vectors:

Sex - Two binary predictor vectors representing sex category membership (male, female).

OC - Six binary predictor vectors representing membership in one of the following homogeneous categories: Mechanical (Dod occupational groups 2X0, 600, 601, 602, 6X0, and 780), administrative (DoD occupational groups 510, 5X0, and 820), security and support specialties (DoD occupational groups 220, 640, 83A, 83B, and 8X0), medical (DoD occupational groups 300 and 3X0), utilities maintenance and technical specialties (DoD occupational groups 4X0 and 720), and electronics (DoD occupational categories 100 and 1X0). Refer to Table 3 for a nominal description of each DoD occupational group.

ASVAB - Four continuous predictor vectors representing scores on the mechanical, administrative, general, and electronic aptitude indices of the Armed Services Vocational Aptitude Battery.

PII - Twelve continuous predictor vectors representing scale scores on the Pictorial Interest Inventory. Refer to Table 1 for a nominal description of each scale.

VOICE - Eighteen continuous predictor vectors representing scale scores on the Vocational Interest-Career Examination. Refer to Table 2 for a nominal description of each scale.

*R^2 difference significant at the .05 level.

**R^2 difference significant at the .01 level.

Table 5. Results of Regression Analyses Demonstrating the Relative Validity of the PII and VOICE for Subjects[a] at Different Reading Ability Levels

Comparison	Models		R^2		DF1	DF2	F
	Full	Restricted	Full	Restricted			
		Low Reading Ability Group (General AI 60) N = 1,517					
1	OC x (Sex + ASVAB)	Unit	.0607	.0000	35	1481	2.73**
2	OC x (Sex + ASVAB + PII)	OC x (Sex + ASVAB)	.1440	.0607	72	1409	1.91**
3	OC x (Sex + ASVAB + VOICE)	OC x (Sex + ASVAB)	.1525	.0607	108	1373	1.38**
		High Reading Ability Group (General AI 85) N = 1,186					
1	OC x (Sex + ASVAB)	Unit	.0261	.0000	35	1150	1.62**
2	OC x (Sex + ASVAB + PII)	OC x (Sex + ASVAB)	.1513	.0261	72	1078	1.84**
3	OC x (Sex + ASVAB + VOICE)	OC x (Sex + ASVAB)	.1894	.0261	108	1042	1.69**

Note. Refer to Table 4 for a description of the variables included in each predictor model.

**R^2 difference significant at the .01 level.

WEISSMULLER, Johnny J., and MOORE, B.E., The Center for Cybernetic Studies, University of Texas, at Austin.

CODAP: NEW APPLICATIONS AND THEIR IMPLICATIONS FOR HIGH-LEVEL DESIGN
(Wed A.M.

In the environment of a major university, the Comprehensive Occupational Data Analysis Programs (CODAP) system reaches many new and varied users. The range of suggested applications greatly exceeds the scope of applications which have been traditionally utilized within the military. For this reason, The Center for Cybernetic Studies is engaged in a long-range project to reformulate and adapt the CODAP system to easily and efficiently address changing needs. Examples of current needs will be presented and the design implications of those needs will be discussed. These examples include: assessing selection/promotion panels in light of anti-discrimination laws, manpower modelling and the special problems encountered in surveying managerial/professional level personnel. The concept of High-Level Design will be defined and its properties as an effective strategy will be explored.

CODAP: Applications and Their Implications For Higher Level Design

Johnny J. Weissmuller
Brian E. Moore
Michael C. Thew

The Center for Cybernetic Studies
The University of Texas at Austin
Austin, Texas 78712

INTRODUCTION

There is available to universities and governmental agencies a fundamental technology which can support both operational and research occupational analysis programs. The core of this technology is called CODAP, an acronym for the Comprehensive Occupational Data Analysis Programs. The CODAP "system" is a set of analysis tools and procedures which use, as raw material, information provided by members of the occupational field being studied. This system may be used to improve classification structures, assess job related skills, verify the relevance of training courses and a host of other applications in which an accurate knowledge of job content at the task level is desirable.

At the last Military Testing Association Conference, an overview of the entire CODAP system was presented by the United States Air Force Human Resources Laboratory, the originators of the CODAP approach. As many of the attending organizations were from state and local agencies, questions were raised as to the applicability of this technology to the personnel problems within the civilian sector because the primary development and use of CODAP has been in the military setting with particular emphasis on training. This paper describes how the CODAP system may be used in specific applications outside of the traditional military setting and in areas of pressing concern to the civilian sector. Moreover, the purpose of this paper is to describe the ways in which these applications differ from the traditional military applications and explain why the Center for Cybernetic Studies at the University of Texas at Austin has undertaken the project to redesign CODAP for use in the civilian sector.

TRADITIONAL MILITARY APPLICATIONS

Traditional military applications involve the development and validation of training programs along with some review of the classification structure. The military's concern for training is demonstrated by the fact that the Air Force's main operational organization using CODAP (USAF Occupational Measurement Center) is functionally part of the Air Training Command (ATC) and the Army's equivalent organization is part of their Training and Doctrine Command (TRADOC). The routine use of CODAP for classification has apparently been limited to the level necessary for making training decisions. Hence,

review of the hierarchical clustering results was limited to identifying coarse job types. Recommendations regarding the classification structure were made only if major training problems could be solved by such an action or if such recommendations were specifically requested by the career field management. The operational programs of the other military services and the Coast Guard generally follow the same emphasis, although the Marine Corps has probably made the greatest use of the classification capability.

Several important observations can be made. First, these applications used composite job descriptions for large groups of incumbents. In other words, the emphasis was on management at the career field level, not the position level. This implied a lesser need for precision in the definition of particular jobs. Secondly, with an overriding aim to simply adjust training for future entry-level personnel, the implication was that there was no particular need to be able to match the current incumbent's responses to his or her personnel record. In other words, no attempt was made to retain this information on an individual basis as a means of documenting experience in a highly useable form. Thirdly, there seemed to be little attention given to surveying officers -- that is, the management level personnel who do not receive extensive and expensive technical training. And fourthly, there was no evidence to indicate an active effort to exploit the possible uses in the areas of job redesign and productivity assessment.

CURRENT CIVILIAN CONCERNS

Civilian organizations are begining to realize their acute need for these and more advanced techniques. Two of the primary forces motivating this search are mandated court actions and a basic recognition of employees as both the human resources of the organization and as individuals with unique needs. Not much has to be said regarding the impact of court decisions except that they demand immediate solutions to complex problems. In a wider perspective, however, the courts seem to be asking for a responsible effort to establish job requirements and for the use of selection and placement procedures tied as nearly as possible to those requirements.

It is a statement about the maturing nature of our system that more recognition is being given to the dual aspect of employees as the human resources of the organization and as individuals with differing needs. Long range concerns are begining to hold their own against short sighted tactics. This viewpoint is forwarded in a defintion of manpower planning by Eric Vetter:

> The process by which management determines how the organization should move from its current manpower position to its desired manpower position. Through planning, management strives to have the right number of people and the right kinds of people, at the right places at the right time, doing things which result in both the organization and the individual receiving maximum long-run benefit. (Vetter, 1967)

Using this definition, manpower planning seems to be the most generic term for all the concerns facing the civilian sector. The highest objective

is to accomplish the mission, and now that mission includes the well being and personal growth of the employees. One of the first civilian agencies to use CODAP in such an innovative manner was Ontario Hydro. In Ontario Hydro's survey of managerial personnel, the purpose was to develop experience profiles and assess how each manager could profit from the available in-house training courses. The next logical step would seem to be to evaluate how well potential canidates match vacant position requirements and interview high matches for critical positions and medimum matches for career broadening positions. In the medimum match cases, those areas showing the least overlap could be identified in advance and qualification training be provided, even before taking charge of a new position.

In a related point, the University of Texas at Austin became familiar with the CODAP system as the result of development work on the Static Multi-Attribute Assignment Model in a contract with the Navy (Charnes, 1973). One vital point in this goal programming model was the existence of both accurate job descriptions and job ideal & minima constraints. At that point in time, the CODAP system could address the first need, but not the second. The Task Factor Applications Package added to CODAP in 1975-76 provided the tools necessary for that type of analysis, albeit too late for that project. An overview of that package is being presented at a different session of this conference.

IMPLICATIONS FOR HIGHER LEVEL DESIGN

Thus far, new applications have been discussed. But what are the implications of these applications for design purposes? There are three conflicting trends and each of these trends implies a different design strategy. The first trend is that of stagnation. This position holds that the CODAP system was in development for 20 years and must therefore be complete by now. This viewpoint, however, is extremely untenable since the CODAP system is designed to be flexible and adaptable to changing needs. To state that personnel needs have not changed in the past 20 years is to ignore the point of the preceding section.

This position was further damaged by an excellent presentation given at the International Personnel Management Association Assessment Council (IPMAAC) in Boston this last July. Mr Ted Darany, IPMAAC President, talked about the revolution in the computer industry and how it will undoubtedly affect the personnel function. He said it was now incumbent on personnel managers to posit ideal personnel systems because the technology is rapidly making possible whatever people request. He stressed that these requests should be well thought out because the initial systems may quickly proliferate, and initial market leaders may dominate the field for years to come, despite the later emergence of significantly superior systems. His message was clear -- the microcomputer revolution can be a boon or a bane to the personnel field, and our planning and forethought will be the deciding factor.

The second trend is represented by the CODAP redesign project being done by Texas A&M under the direction of the Navy. This trend is to enhance and modernize the software designed to support the traditional military

applications. As pointed out earlier, restricting attention to simply training considerations makes CODAP take on the appearance of any other statistical package. The Texas A&M design, in fact, consciously emulates such packages. This direction was made even more apparent by the publication of a Navy Technical Report entitled "Methods to Evaluate Scales and Sample Size for Stable Task Inventory Information" (Pass, 1980). This is a viable trend, given the mutually agreed upon limitations on the subject area and the equipment profile of the projected users (large, main-frame computer systems). The potential failing in this approach, however, will come if the users decide they wish to expand their programs in the directions apparently being pursued by the civilian sector.

The third trend is the approach represented by the Center for Cybernetic Studies at the University of Texas. This design effort is aimed at addressing the pressing and near future needs of the civilian sector. This approach calls for greater discrimination in individual job descriptions, not less (99 point scale?). This approach calls for data files that are compatible and cross referencable with personnel data files for both individual level records and for comparison to composite level descriptions. Census sampling (100 percent of all incumbents) is the expected order of the day, and total maintenance of this experience history is anticipated. This approach calls for a small core of CODAP programs which do things essentially unique to CODAP, but with easy interface to any number of available, off the shelf statistical packages. More importantly, interface to other stand alone packages (e.g. Static Multi-Attribute Assignment Model) is crucial to a system with a reasonable life expectancy.

What things are in the core of CODAP? Basically, the following:

1. The ability to generate, store, manipulate and report individual or composite job descriptions. (JOBSPC, COMGEN,FACPRT)

2. The ability to empirically define potential groups (OVRLAP, GROUP, DIAGRM).

3. The ability to identify the consensus of managerial/expert opinion or policy at the task level (REXALL).

4. The ability to derive task level factors from incubment background data (AVALUE, AVGPCT).

5. The ability to generate and report an interaction between the vectors from items 3 & 4 above with the data vectors in item 1 for individuals or the groups identified in item 2. (VARGEN, VARSUM, PLTVAL)

What needs to be added to CODAP is:

1. An easy way to incorporate module level

(102)

policy with task level policy in a realistic hierarchy.

2. An easy way to handle "Text Factors" for use in detailed task analysis projects.

3. An easy profile similarity assessment, reporting and perhaps clustering mechanism.

4. An easy way of restructuring the data (tasks versus background questions) to fit the analysts needs.

5. An easy way to record and report job requirements by job description.

6. An automated way to do job type selection and highlight differentiating tasks (William Phalen of AFHRL is on the verge of a major breakthrough in this area).

The basic emphasis of the University of Texas design is for modularity with rapid growth potential and easy interface to other systems to minimize duplication of effort and maximize cooperation. Much of the design concern is directed at the fundamental difference that civilian organizations desire position level management in addition to, not in place of, management at the career field level. This desire is very appropriate in response to both the court mandated review of the selection and promotion systems and the new emphasis on employee development and satisfaction. Only with a system that is integrated from the lowest levels can one reasonably hope to accomplish the widely diverse aims that are arising in the civilian sector.

References

Charnes, A., Cooper, W. W., & Niehaus, R. J., "Dynamic Multi-Attribute Models for Mixed Manpower Systems" Austin, TX, Center for Cybernetic Studies, 1973.

Moore, B. E., "Models for Organizational Design and Staffing," Proceedings of NATO Conference on Manpower Planning and Organization Design, (with E. Bres, F. Leader, and D. Sholtz) New York, Plenum Publishing Company, 1978.

Pass, J., & Robertson, D. W., "Methods to Evaluate Scales and Sample Sizes for Stable Task Inventory Information" NPRDC 80-28, NPRDC, San Diego CA, May 1980.

Vetter E. "Manpower Planning for High Talent Personnel" Ann Arbor Michigan, University of Michigan, Bureau of Industrial Relations, 1967.

WELSH, John R., Air Force Manpower & Personnel Centre, Randolph AFB, Texas.

CRITERIA DEFINITION AND MEASUREMENT IN THE AIR FORCE PROMOTION SYSTEM
(Tue A.M.)

This is an informational report designed to document the role of Specialty Knowledge Tests (SKTs) and Promotion Fitness Examination (PFE) in the Air Force's Weighted Airman Promotion System (WAPS). Specifically examined is the process which ties the test content domain to specialty knowledge requirements. Also examined is the role of supporting documents and occupational surveys used to establish the relevant specialty knowledge domain and difficulties associated with the process. Current issues in the test development process are discussed, including the formatting of the occupational survey results, the role of the survey in development of other fundamental documents used to establish the specialty knowledge domain, literacy requirements, and job vs specialty concepts.

CRITERIA DEFINITION AND MEASUREMENT IN THE USAF PROMOTION SYSTEM

MAJOR JOHN R. WELSH
Air Force Manpower and Personnel Center
Randolph AFB, Texas

INTRODUCTION

The United States Air Force uses a weighted factor system to rank order airmen eligible for promotion. This system was developed from policy capturing research on enlisted promotion boards by the Air Force Human Resources Laboratory (AFHRL) and documented in several technical reports (Koplyay, 1969a; 1969b). The system has been in effect since 1969, and provides several unique features for enlisted airmen seeking promotion. The Weighted Airman Promotion System (WAPS) system is objective in that quantitative values are given for each factor, and it's highly visible compared to the relative secrecy of closed promotion boards. All people competing for promotion know where they stand in relation to others competing for the same promotions and what factors aided or prevented their promotion.

Test Development Branch of the United States Air Force Occupational Measurement Center (USAFOMC) is responsible for the development and revision of the Air Force Specialty Knowledge Tests (SKTs) and Promotion Fitness Examinations (PFEs) used in the Weighted Airman Promotion System. Other tests used by the Air Force are also developed by the Test Development Branch but will not be discussed in this paper. A sister branch of the USAFOMC is charged with the development and reporting of occupational survey data. The purpose of this paper will be to report the current test development process and how the test content is tied to specialty knowledge requirements with training and job defining documents, as well as occupational survey data. Problems associated with the process will be discussed as well as recurring issues associated with the difficult process of tying test content to the job knowledge domain.

The SKTs and PFEs are currently used in the Air Force promotion system as two out of six weighted factors used to rank order promotion eligible airmen under WAPS. The other factors include: Time in Grade (TIG), Time in Service (TIS), Awards and Decorations (DECs), and the Airman Performance Report (APR). As the system is currently designed, a certain number of points are assigned to individuals for each of the six factors, based on a published list of points awarded for different values of each factor. The total WAPS score of a given person then determines his/her standing relative to all other eligible airmen in a given Air Force specialty competing for promotion to a given pay grade.

Implied from the above description, the total WAPS composite score and all its components are the proper scrutiny of any analysis dealing with potential adverse impact. The SKT and PFE scores then, are a part of the composite WAPS score which forms part of the Weighted Airmen Promotion System. These tests are not predictors of potential performance in the traditional sense, but are, in fact, designed to be achievement tests reflecting attained knowledge. From this reasoning the appropriate validity of SKTs and PFEs, in their present use, is determined using content validation studies, as opposed to criterion related or construct validation research.

It will be an underlying assumption of this paper that SKTs and PFEs are in fact criteria for determining promotion eligibility of airmen, and as such, the content validity of these tests must be established in the procedures used to develop them. For the sake of convenience, the general information relayed in this paper about how content validity is established will be discussed in terms of the SKT, but also can be assumed to apply to the PFE.

This paper will primarily attempt to explicate the process which ties SKT content domain to the job knowledge domain. As necessary, the parts of the test development process which serve to satisfy EEO requirements relating to SKT content validation will be described.

Content Validation of SKTs. While the concept of content validity has been discussed and covered thoroughly in many test and measurement texts, some basic features of the concept need to be set forward here in order to form a basis for subsequent discussion. Content validity requires a representative sample of performance in a domain of situations (APA Standards, E12, 1966). In sum, a test can be said to be content valid if it is based on a representative sample of all the possible knowledge requirements that could have been included on the test. To operationalize this concept is difficult at best. As Guion, 1965 observes, content validity is determined by more by expert judgement than by empirical correlation (P. 124). By injecting human judgement into the process in order to determine what constitutes a representative sampling of the possible job, one removes to some extent the ability to verify by empirical means the veracity of the process.

Conventional wisdom as well as EEO standards (Uniform Guidelines section 14e) require that expert judgement be directly supported by official documents and job analyses which serve to define the entire scope of a given job. The documents and job analysis used in defining the Air Force specialty content will be explored in some detail below. The Air Force requires, by operational necessity, that individuals within a specialty perform

or have knowledge of the broad range of functions served in that specialty. A specialty then is the unit of discussion, not the individual's current job within that specialty. This distinction is important, for it would be virtually impossible to develop thousands of individual tests for each job being performed in a specialty. It's difficult enough dealing with over 230 specialties.

Described below are a number of relationships which serve to anchor the test (and the test outline) to various controlling documents and the USAF occupational surveys. Any test outline is important for insuring content validity because it determines the number and kinds of questions appearing on the final test. A properly constructed outline will help insure a representative sampling of the domain of the specialty knowledge.

TEST DEVELOPMENT

All SKTs and the PFEs are revised annually. Annual test revision accomplishes the dual purpose of reducing likelihood of test compromise and allowing the highest degree of currency of test content possible. The test revisions/development is accomplished using teams of two to four Subject Matter Specialists (SMSs) and a test psychologist to generate new SKTs. It is in the process of test development that the content validity of the SKTs is established and will be the focus of the following discussion.

Domain Defining Documents. There are two official documents which control the development of the test outline: the Air Force Regulation 39-1 (career field introduction, career field progression chart and specialty description); and the Specialty Training Standard (STS).

The first of these documents, the AFR 39-1, serves as the official specialty description. It contains broad descriptions of the duties and responsibilities of airmen and NCOs at each of the various skill levels in a given specialty. It also provides minimum aptitude and training requirements. As a general guide, these documents are invaluable (example at Appendix A). The test development teams of subject matter specialists are required to have one member of the team read aloud each sentence of the AFR 39-1 to other members of the team at the start of each test development project (SKT Handbook, para 4.3). This procedure is also used for each paragraph of the Specialty Training Standard. As each document is orally reviewed, each team member must agree with the meaning of each statement on both documents before proceeding.

This exercise has two major purposes. The review helps set the SMSs' frame of mind for an Air Force "total" view of the specialty, and allows differing interpretations of the specialty descriptions to be resolved and/or surfaced before development of the test outline begins. At this point in the process, SMSs make

formal recommendations for changes in the two control documents to reflect the most recent developments in the specialty, and to resolve existing difficulties in the classification process. It is important to note that subsequent test outline development is not based on recommended changes; but significant conflicts over meaning and content must be resolved before the test construction project proceeds.

Test Outline Development. As mentioned above, the most critical part of any SKT/PFE development project is the development of the test outline. The outline, in its finished form, contains the weighting to be applied to each area of the test outline, and consequently, the number of items appearing on the final test. The techniques of outline development are described elsewhere (SKT Handbook, Chapter 4). The important aspects of outline development for the purposes of this paper are the decision on what content to include and what emphasis to place on that content. Additionally, topics and task areas selected are arranged in the sequence in which they will appear in the final test, and there is a general progression of easier to more difficult test content. The determination of the weights (emphasis) for the various outline areas is governed by a number of considerations. The determination of the types of tasks and their emphasis drives the content validity of a test. Before discussing how the emphasis or weighting is done, test content at this point must be directly tied to the Specialty Training Standard (STS). In fact, each specialty specific STS paragraph must be reflected on the test outline. That is, each major test outline area must have reflected on the outline itself the appropriate STS paragraph number(s) relating to an outline area. Some STS paragraph numbers can be zero weighted for reasons mentioned below, but all must be included under some major area of the test outline. This is an important first step in defining the test content and anchoring that content to the job knowledge domain.

Zero weighting of certain outline areas occurs when the STS paragraph covers certain duties or equipment that are no longer appropriate or, the STS covers duties that are not performed frequently. The zero-weighting allows subsequent test development teams and any other concerned parties to know that the content was considered, and not included intentionally. The determination as to which tasks are performed infrequently is made with the aid of the USAF Occupational Survey - the job analysis used by the Air Force. The role of the survey will be discussed in more detail later.

It is crucial for later arguments to understand the function of the STS. The STS is a driving document for a number of Air Force programs including technical training course development (in terms of course content) and On-the-Job Training (OJT). The knowledge requirements are specified by alphabetic designators which outline not only the knowledge areas required at specific

(1032)

skill levels but provide a hierarchy of values that serve, in a rough fashion, to define the extent or depth of knowledge required of the subject or task by an individual. The example at Figure 1 (adapted from the SKT Handbook, para 6-4, 1979) shows the STS proficiency code key defining how task and knoweldge areas are established on the STS.

Subject Matter Specialists use these codes to help define the type of knowledge required at a given skill level. At a later stage of the test development process, during the item writing phase, the test psychologist and SMSs use these codes as a general guide to generate items appropriate for a given level of content validity as depicted in the hierarchy. During the outline development phase, these codes help SMSs apportion test content and emphasis to the tests to be developed for the various pay grades. The outline eventually describes content and emphasis for a three-skill level proficiency test called the Apprentice Knowledge Test (AKT), and three SKTs - one for pomotion to E-5, and two others used for promotion to E-6 and E-7. An example of an STS and a test outline is provided at Appendices B and C respectively.

An important point to note at this stage of the discussion is that the areas of the STS paragraphs are worded in very general terms. The STS provides useful guidance up to this point in defining the content of the test, but the general wording of most STS paragraphs often does not provide the more exact or refined task statements that would be required in order to develop good test items tied to specific specialty knowledges required to perform specific tasks. As can be seen from the example of the specialty description at Appendix A, the 39-1 specialty description provides even broader statements - useful for its intended purpose but of limited help in developing a test outline with specification of clear knowledge areas.

The USAF Occupational Survey Report (OSR) is used to resolve problems of test content determination and emphasis. The Occupational Survey information is used in a number of ways to provide a more refined and realistic view of the Air Force specialty as it is actually being used in the field. One of the most important services provided by survey data is to expand the base of experience. The test development team, as mentioned above, is composed of about four SMSs. While this may seem to be a small number relative to the total number of airmen in a given Air Force Specialty (AFS), they generally represent between 50 and 100 years cumulative experience in the area. Even with this experience base, one needs to measure their decisions on test content against a bona fide job analysis.

1033

STS PROFICIENCY CODE KEY

	Scale Value	DEFINITION: The Individual
TASK PERFORMANCE LEVELS	1	Can do simple parts of the task. Needs to be told or shown how to do most of the task. (EXTREMELY LIMITED)
	2	Can do most parts of the task. Needs help only on hardest parts. May not meet local demands for speed or accuracy. (PARTIALLY PROFICIENT)
	3	Can do all parts of the task. Needs only a spot check of completed work. Meets minimum local demands for speed and accuracy. (COMPETENT)
	4	Can do the complete task quickly and accurately. Can tell or show others how to do the task. (HIGHLY PROFICIENT)
* TASK KNOWLEDGE LEVELS	a	Can name parts, tools, and simple facts about the task. (NOMENCLATURE)
	b	Can determine step by step procedures for doing the task. (PROCEDURES)
	c	Can explain why and when the task must be done and why each step is needed. (OPERATING PRINCIPLES)
	d	Can predict, identify, and resolve problems about the task. (COMPLETE THEORY)
** SUBJECT KNOWLEDGE LEVELS	A	Can identify basic facts and terms about the subject. (FACTS)
	B	Can explain relationship of basic facts and state general principles about the subject. (PRINCIPLES)
	C	Can analyze facts and principles and draw conclusions about the subject. (ANALYSIS)
	D	Can evaluate conditions and make proper decisions about the subject. (EVALUATION)

HIERARCHY OF CONTENT VALIDITY

STS CODES	MATERIAL	ITEM
	Simple recall	Single factor
1,a,A	Simple fact, term, equipment part, or component Nomenclature	WOOTF/WOTF What Where
	Steps in performing a task	Single and double factor
	Relationship between two	Sequence
2 or 3	or more basic facts	WOOTF/WOTF
b	Identification of operating	When
B	principles	Where How
	Analysis and application principles	Purpose Prediction
	Theory	Multifactor
4	Evaluation	Troubleshooting
c or d	Problem solving	What set of data
C or D	Most complex tasks	Why Where How

Figure 1.

(1034)

The Role of the Occupational Survey. The Air Force uses the Occupational Survey for a variety of important purposes - especially in defining course content in formal training, training requirements in general, and OJT. The survey results are also used to develop the test outline. The usefulness and the necessity of using Occupational Survey data to develop SKTs has been described in several previous reports (Bills, 1978; Phalen, 1978; Vaughan 1976, 1977). The previous papers dealt primarily with specific procedures used, or proposed for use, in developing test outlines directly from survey data. This present effort deals with more conceptual issues and practices which allow test psychologists and SMSs to generally relate test content to the specialty domain.

As a matter of practice, the test psychologist will withhold the survey data until the SMSs have generated a tentative outline. The test psychologist will compare the first-brush outline with the survey report in order to spot areas of potential conflict and to verify the judgement of the SMSs. Additionally, the test psychologist gets the benefit of a fresh look by having the SMSs develop the new outline "cold."

After the SMSs develop the first rough outline, the Occupational Survey is used in a number of ways to refine decisions on test content and weighting. The outline emphasis is checked by the test psychologist. The psychologist also checks the Survey data to determine if all appropriate content is included on the test outline. Serious deviations in content from STS paragraphs or questions concerning content and emphasis are generally resolved using a special survey package developed just for test construction. This package has tailored Occupational Survey information specifically designed to aid in verifying test outline headings and weighting.

The special testing package has the task listings ordered in terms of the STS paragraph areas in addition to the more familiar OSR data like percent members performing by pay grade and the training emphasis index for each task. Task information is also ordered in a second part of the special SKT package in terms of the time spent performing the task by pay grade for various groups of airmen.

The OSR data is particularly useful in resolving questions about the amount of time actually spent on specific tasks by different groups of people in the career field. Percent-members-performing is also used by the test psychologist and SMSs in breaking down the broad STS paragraph areas in more specific tasks, and in deciding relative emphasis of content.

The use of OSR data in helping to verify outline content and emphasis is still somewhat limited by tne fact that even OSR task statements are sometimes too broad for use in outline development. Additionally, SMSs must take these statements and use their experience and judgement to determine what specific knowledge is required for an individual to successfully perform the task. Even with these limits, however, the survey expands the base of SMS experience to include the task response information on a substantial portion of all airmen performing in the specialty. This is an important feature in securing the content validity of an SKT. All tasks appropriate for testing are given appropriate consideration for inclusion on the test, and the broad task/knowledge area emphasis is verified against an objective, and in many cases, exhaustive inventory of all tasks performed and percent-members actually performing the tasks. In the future, the special SKT package of OSR data will include not only training emphasis information, but will include an index of testing importance. This index represents a combination of percent-members-performing and training emphasis in terms of time required for training. This new index will provide a handy means to guide test developers in assuring the job performance domain is reflected in the test content. In fact it will serve as an index of importance of the task for testing.

Interrelationship Among Controlling Documents. An extremely important fact in the establishment of the job performance domain for test content validation is the relationship between the various controlling documents (STS and AFR 39-1), the Occupational Survey data, formal technical training course content, and the Air Force On-the-Job training program.

These relationships are important because all official documents and survey results are related in various ways, and it is this interrelationship which serves to conceptually define the range of job performance within a specialty.

For example, the STS is used to help training curriculum course developers establish what will be trained in formal training courses and what will be trained in On-the-Job training. Occupational Survey data is being used to refine judgements about the curriculum for formal training courses as well as for redefinition of STS areas.

In turn, the STS is also used in the Air Force OJT program. In order for an airman to become eligible for consideration for promotion under WAPS, he must attain a specified skill level (3-level, 5-level, 7-level). The airmen train to the appropriate skill level in OJT and by using Career Development Courses (CDC). The CDC is a correspondence course developed from STS and other material as well. Additionally, the CDC serves as a principle reference for test item development. In this way, test content is indirectly related to the job performance content that is trained

on the job and in formal training. Attainment of the appropriate specialty knowledge is measured both by award of the skill level upon successful completion of OJT, and also when the individual competes for promotion by taking the SKT. This is the same job knowledge that appears in the appropriate training standard for OJT; it is also the same knowledge taught in his/her CDC, and the formal courses.

Occupational survey data is most extensively used, at the present time, to make recommendations on training requirements. This use, in fact, helps keep the STS and formal course training standards in line with what work is really being done "out there." At the present state of the art in developing training standards, many other considerations must play in the final structuring of the STS - considerations that somtimes tend to make the STS less than optimal for test development purposes. By using STSs, AFR 39-1, CDCs, and specially prepared occupational survey data in combination with expert judgement of SMSs, one can reasonably assure content valid tests, because all of these documents help serve to define what is to be trained, and what is to be required of the airmen at each skill level and pay grade in the specialty.

The SMSs on the test development team also have an opportunity to bring their experience to bear in recommending changes to all documents they use in defining test outline content. They formally critique the compatability of the CDC and STS - outlining areas of insufficient coverage in either document. Where possible, the test development SMSs also interact with the occupational analysts to provide areas of needed survey concentration; refinement of task statements; and needed updating of survey results.

FUTURE ISSUES

Timing of Control Document Development. One of the recurring issues for test developers is the timing of changes for the two fundemental control documents, the AFR 39-1 (job description) and the STS. As mentioned earlier, these documents are structured and developed to serve many purposes. The coordination involved in changing either document is extensive. Given the need for annual test revision, revision of test content to coincide with contemplated changes in the controlling documents is sometimes difficult. In general, the strategy used is to avoid including, on the newly revised test, that content which appears to be "controversial" or subject to near term change. This strategy is usually effective, but occasionally, material included on the test is made inappropriate because of changes to control documents or other policy actions. In these cases, questions dealing with the material are deleted from the scoring of the affected SKT.

A related, but more difficult timing problem exists with the uses of the Occupational Survey data. Many surveys are initiated by demands created by career field needs and need for redefinition of training requirements. In general, there can be large time intervals (in excess of one year) between availability of survey results and subsequent changes in training course curriculum, STSs, and AFR 39-1 specialty descriptions. These problems are circumvented by test developers using survey data directly, but there remains the potential problem of not having new test content properly aligned with control documents. These issues are generally avoided by the strategy mentioned above, i.e., avoiding content which as yet had not been included in the AFR 39-1 or STS.

Timing of OSR Data. As a rule of thumb, Air Force specialties are resurveyed every four years with the Occupational Survey, unless pressing needs of users of the Air Force specialties and/or functional managers or training managers require more recent survey information on which to base important decisions about the structure of a given specialty. The more dated survey results are used, but with greater care and greater reliance on control documents and SMS judgement. In many cases, the feedback provided by test development SMSs and from the field reveal that even the oldest OSRs would require little change, and hence, remain accurate reflections of the current job. It would still be desirable to have shorter cycling of surveys, but given the large increment in resources needed to realize this, it may not feasible.

Formatting of OSRs. The format of the USAF Occupational Survey is dictated in large measure by the demands of its primary user, the training community. This format is not the most useful for test development purposes. To help solve the problem of translating OSR results into a test outline, special OSR packages have been developed and prepared as described briefly above. Still, much needs to be done. Questions of outline emphasis cannot still be translated directly from OSR data to numbers of questions appearing on a final test. USAFOMC is actively engaged in efforts to develop procedures to allow a direct, objective translation of OSR data to test content and emphasis. Until that time, expert judgement must play an important role in the translation process.

CONCLUSIONS

As Guion, 1965, has claimed, the importance of of content validity is greatest for criterion measures. It has been part of the underlying thesis of this informational paper that SKTs and PFEs are, in fact, criterion measures.

They are two of the six criteria used to determine promotion eligibility of airmen under the WAPS. The complex relationship between test content and the knowledges required for performance of specific tasks on the job is established by job defining

documents and Occupational Survey results. The various training requirements, OJT and formal technical training, are also established from the same documents which drive test development. All depend on Occupational Survey data to provide a clear "picture" of the specialty, but expert judgement is used to help translate the survey data into meaningful, representative test content and emphasis. The relationship among all these standards is important in that together they serve to define what knowledge an airman must attain for what skill level achievement and level of proficiency an airman must have in order to be competitive for promotion. In all, the "picture" of the specialty is a dynamic thing. The various standards coupled with survey results, expert judgement, and annual revision of tests help assure the continuing development of specialty relevant promotion criteria in the United States Air Force.

REFERENCES

Bills, C.G., Evaluation of Computer-Driven Test Outlines Using Conventional Test Outlines as a Criterion Reference During Test Development Projects. Proceedings of the 20th Annual Conference of the Military Testing Association, 1978

Guion, R.M., Personnel Testing, New York: McGraw Hill, P. 124, 1965

Koplyay, J.B., Field Test of the Weighted Airman Promotion System: Phase I. Analysis of the Promotion Board Component in the Weighted Factors Systems. AFHRL-TR-69-101, April 1969

Koplyay, J.B., Field Test of the Weighted Airman Promotion System: Phase II. Validation of the System for Grades E-4 through E-7. AFHRL-TR-69-102, May 1969

Phalen, W. J. The development of a Technique for using Occupational Survey Data to Construct and Weight Computer - Driven Test Outlines For Air Force Specialty Knowledge Tests (SKTs). Proceedings of the 29th Annual Conference of the Military Testing Association, 1978.

USAF Occupational Measurement Center. Handbook for the Construction of the SKT and Associated Tests, 1 Nov 1979.

Vaughan, D.S., Prediction of Test Outline Weights from Occupational Survey Data. Proceedings of the 18th Annual Conference of the Military Testing Association, 1976, 435-460.

Vaughan, D.S., The Interface Between Occupational Survey and Test Construction. Proceedings of the 19th Annual Conference of the Military Testing Association, 1977, 798-800.

WILSON, Capt. Fred P., and ELLIS, Major R.T., Canadian Forces Personnel, Applied Research Unit, Willowdale, Ontario.

THE DEVELOPMENT OF A SYSTEMATIC COUNSELLING MODEL FOR THE CANADIAN FORCES (Tue A.M.)

The Canadian Forces (CF) conducts vovational counselling in a variety of settings. Primarily formal counselling occurs:

a. upon enrolment,

b. at significant career decision points during training and later service,

c. upon retirement.

Although the CF has developed technical procedures suited to each vocational counselling situation, these have been considered in isolation rather than as parts of a continuous or related process. CFPARU is attempting to integrate these requirements within a conceptual model of vocational counselling appropriate to CF needs. This has led to the development of a proposal for an automated career counselling system which is aimed primarily at supporting the recruiting/enrolment phase. The bulk of the paper describes this latter development.

The counselling model is built on the premise that vocational counselling outcomes are highly dependent upon the availability to the counsellor and enrollee of accurate and realistic personal data, hight quality information on the military environment and the efficiency of the mode(s) of comunication used to impart the information. To the extent that these three elements contribute to self knowledge and insight into the military environment, selective listening and information distortion will be minimized, vocational maturity and socialization will be accelerated, and realistic expectations will be engendered. The interactive computerized counselling model presented utilizes three modules (Search, Inform, Inquire) complemented by videotape cassetes, to maximize each of these elements.

THE DEVELOPMENT OF A SYSTEMATIC COUNSELLING MODEL FOR THE CANADIAN FORCES

Capt F.P. Wilson

and

Maj R.T. Ellis

Canadian Forces Personnel Applied Research Unit,
Toronto, Canada

INTRODUCTION

Recruiting difficulties and high attrition represent two of the most serious problems facing an all-volunteer military force. Except for conscription during the First World War and the Second World War, Canada has attempted to keep its military manpower up to strength through voluntary enlistment. However, beginning early in the 1970s, it has become increasingly more difficult to recruit and retain eligible service members. According to the authors of a recent CFPARU report (Tierney & Pinch, 1980), manning problems like those experienced by the Canadian Forces (CF) during the past 10 years will probably remain for the next two decades. Similar difficulties have been reported by other countries with all-volunteer forces, such as the U.S.A. (Sinaiko & Scheflen, 1979), Australia, and the United Kingdom.

In an attempt to refine placement procedures, behavioural scientists concerned about *selection in the military* have traditionally focused on aptitude testing. However, it has become patently obvious that there is more to selection and assignment than trade placement based mainly on the use of aptitude tests. Although there can be little dispute as to the worth of validated aptitude tests in trade assignment, overreliance on these methods has created unwarranted optimism in the user as to their purpose and power. The uninitiated have not only expected test scores to validly trade-select (their only purpose), but also to provide finite measures of human characteristics, which is clearly impossible with the kinds of measures currently being used.

A CFPARU study (Martin, 1972) reported that for samples selected in 1968 and 1969, insufficient aptitude was a minor reason for attrition in trade training - accounting for only 2.9% and 5.0%, respectively, of the total attrition. However, course failures accounted for 136 and 141 trainees out of a sample of 451 and 421, respectively. (This rate represented attrition rates of 30% in 1968 and 33% in 1969.) A similar trend of low attrition-due-to-aptitude continues to occur today (Ellis & Saudino, 1980; CFTS Statistical Review 1979/80). This relatively small percentage of attrition, as a result of low aptitude, may well be taken as evidence of the effectiveness of the aptitude-oriented selection procedures that have been in use.

To supplement the efficacy of aptitude testing, researchers have become more concerned about noncognitive assessment, i.e., measurement of such dimensions as personality, interests, values, and so forth, and the usefulness of various vocational counselling techniques. According to several CF authors (Rampton, Skinner & Keates, 1972; Skinner & Rampton, 1973; and Keates, 1975), the need for more adequate criteria coverage has been an ongoing problem for the CF. These writers have posited that such other noncognitive dimensions as mentioned above are required for a better man/job match.

1043

With increased emphasis being given to counselling theory and techniques, one area that offers good promise is the study of realistic job previews and how they affect attrition. Basically, the hypothesis here is that if realistic information about the future working environment is provided to an individual, he will have some idea of what to expect in this new environment and thus be psychologically better prepared to deal with it. However, this subject matter cannot be examined in isolation. Other collateral areas of concern that also must be looked at include vocational development (Super, 1973), socialization, selective listening, and information distortion.

SOME UNDERLYING THEORY

Research evidence has shown that if individuals have a realistic expectation of themselves, the working environment, and how they fit into this environment, significantly greater job stability can be expected (Katzell, 1968; Wanous, 1973; Ilgen & Seely, 1974; Ilgens, 1975; Horner, Mobley & Meglino, 1979). This can be maximized by effectively communicating realistic personal and environmental information. Furthermore, provided that the modes of communication have been effective, and there is movement towards greater appreciation of self and environment, some variable increase in vocational maturity will occur (Super, 1973). If properly managed, this also leads to socialization which is the extent to which an individual's attitudes can be changed, by training and counselling, to reflect the philosophy, aims, and value system of the member's organization (Van Maanen & Schein, 1978). Finally, vital considerations in achieving realistic expectations, vocational maturity, and socialization, are the phenomena of selective listening and information distortion (Katz, 1971). These are the result of an individual's not having the experience or background to give meaning to information received. Consequently, he unconsciously selects or distorts the message to agree with his preconceived notions. This problem can partly be overcome by using improved modes of communication.

CONSIDERATIONS FOR COUNSELLING IN THE CANADIAN FORCES

It is hypothesized that three conditions are necessary for an effective counselling strategy: 1. a means for gathering pertinent personal data; 2. a system for maintaining an extensive and easily accessible source of accurate environmental information; and 3. effective channels for communicating this personal and military information to the individual. Gathering personal data can be accomplished through aptitude and non-cognitive testing, biographical data forms, interviewing, and an interactive computer. A system for maintaining an extensive and easily accessible source of accurate environmental information can be accomplished through the use of printed matter, films, video cassettes, and an interactive computer. Finally, various means of communicating personal and military information include films, sound-on-slide, word-of-mouth counselling, printed matter, videotape, and an interactive computer. In each of the three elements deemed necessary for a complete counselling strategy, the interactive computer is included. However, not only does the computer satisfy the necessary requirements; if it is properly programmed and used, it can fulfil these functions in a clearly superior manner.

Likewise, videotape cassettes provide a powerful counselling tool. Videotape offers both a means of storing data and a means of communicating it. In combination, the computer and videotape offer the possibility of a quantum leap in

counselling effectiveness.

Whenever any one of these components is missing: gathering, storing or communicating information - effective counselling cannot take place. To the extent that the components contribute to self-knowledge and environmental insight, selective listening and information distortion will be minimized, vocational maturity and socialization will be accelerated, and realistic expectations will be engendered.

A COUNSELLING MODEL FOR THE CANADIAN FORCES RECRUITING CENTRES (CFRCs)

System Objectives. The goal is to design, develop and implement a systematic counselling model for the Canadian Forces which will increase the flow of recruits to CFRCs, facilitate effective trade assignment, and yield decreased attrition rates through basic training, trades training and beyond.

The system described in this report has been designed to satisfy an extensive set of criteria. Used properly by well-trained staff, the proposed system will contribute to:

1. increasing the flow of recruits while decreasing attrition rates;
2. positive impressions in its users by being appealing to potential recruits as well as to recruiting personnel and selection staffs;
3. providing realistic and accurate job previews, thus not creating or reinforcing false expectations;
4. accessing trades data via non-cognitive dimensions such as personality, interests, and values, including test data where available on such dimensions;
5. positively influencing vocational maturity and career decisivenss;
6. optimizing the retention of accurate information while minimizing the effects of selective listening and information distortion;
7. helping to ensure that the initial contact with the CF is as positive as possible, thus contributing to positive socialization;
8. facilitating the flow of information from the recruit to the Military Career Counsellor (MCC) and from the MCC to the recruit;
9. providing an up-to-date, factual and efficient data storage system;
10. maintenance of quota and waiting list data;
11. learning (by being designed in a computer-assisted learning format with information delivery, feedback, quizzing, conversational dialogue and positive reinforcement for the recruit);

12. information retention (by providing a hard-copy record of all pertinent information for the recruits' retention, and a record for the CFRC of that which the individual received); and,

13. integration of the data base within the CF structure (by being entirely compatible with other automated CF data storage systems as well as materials in other media such as hard-copy, video-cassettes and films).

Conceptual Outline. Figure 1 provides an overview of the major components of the proposed system (Jarvis & Hutt, 1980). It is followed by a discussion of each of the modules separately.

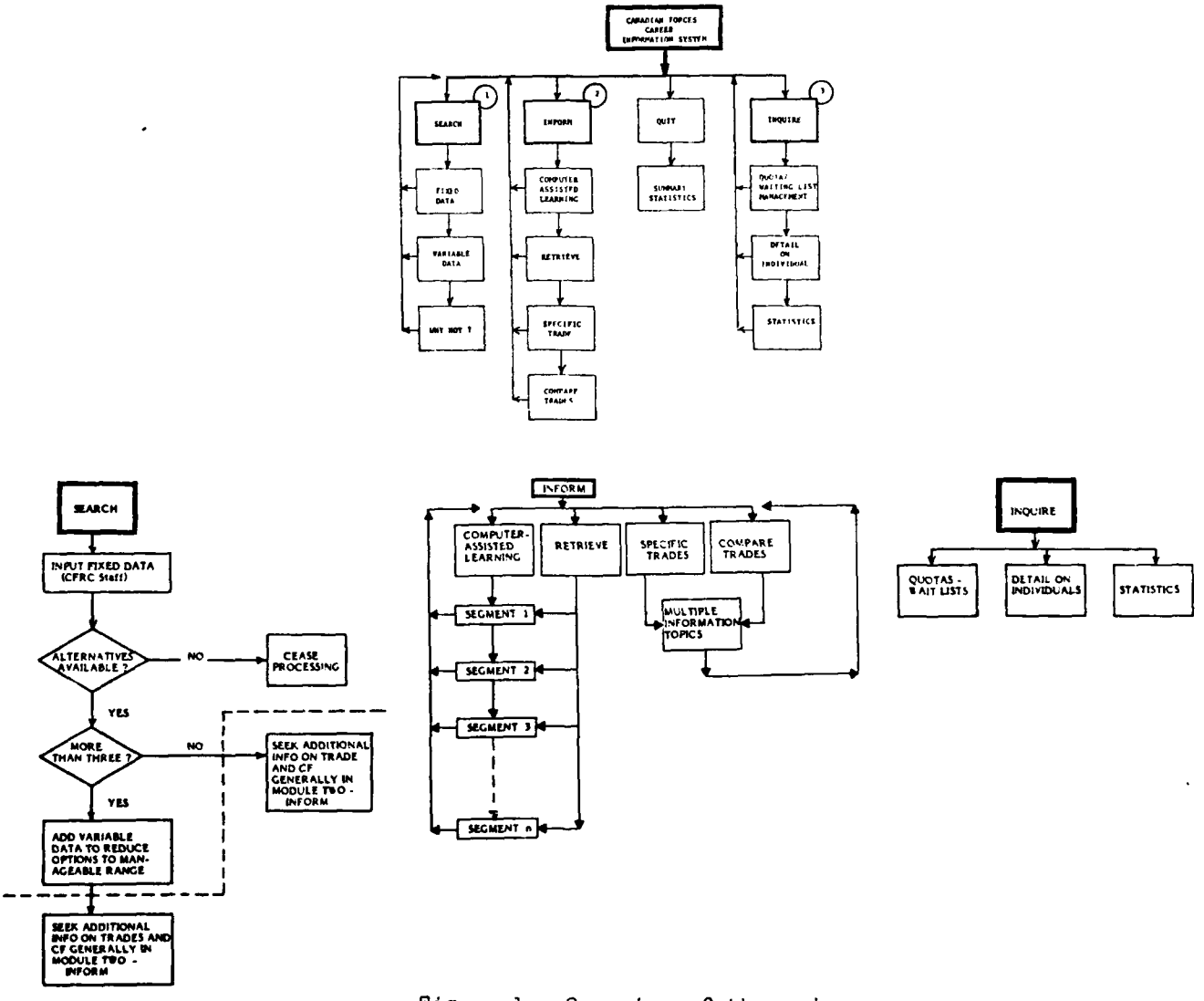


Figure 1: Overview of the major components of the proposed system

1046

Module One - SEARCH. The purpose of this module is to assist CFRC staffs to determine for which trades the applicant qualifies, and to assist applicants to determine which trades would come closest to meeting their needs. This module also serves to initiate an automated personal file for each applicant.

The SEARCH module has two principal components: (1) the Fixed Data input function, and (2) the Variable Data input function. Fixed Data is defined as data over which the client has no control. In this category are items like biographical, medical and test results. This information will be put into the computer in order to allow it to match these data to the requirements of every CF trade. In a matter of seconds the computer will be able to identify all trades for which the individual qualifies.

Data will be entered at a visual display unit (CRT) by a clerk. The computer will edit all input to ensure that errors have not occurred during this phase. Some applicants will not qualify for any trades. If such is the case processing will cease. The computer will specify the input items which caused it to reduce the number of trades available to the applicant to zero. If one, two or three trades are identified as suitable, the job for the applicant and for the MCC is not a complicated one. The MCC will check quota and waiting list data. The applicant will want to obtain information about the trades to determine whether or not one appears more satisfying than the others. If the applicant wants one of the trades, and appears to have sufficient military potential, an assignment to the trade may be made. If the computer identifies more than three trades as potentially suitable after the fixed data has been input, both the applicant and the MCC have further choices to make.

At this point, the applicant may wish to have access to the second component of SEARCH, the Variable Data input function. Variable Data includes categories of information about applicants, such as their interests, temperaments, values, and preferences for various working conditions and environments. Here individuals are encouraged to tell the computer what they would like to do, what they feel they are good at, what kind of career progression they are looking forward to, and more. In effect they are explaining to the computer, as best they can, what they feel it would take to make a "good fit" between themselves and the world of work.

Variable Data SEARCH topics could include the following: interests - self-report and test results (when validated tests for the CF are available), abilities, temperaments, educational level, environmental conditions, earnings, hours of work/travel, physical demands, inside/outside preferences, preference of element, and trade groups. As the applicant enters preferences with respect to any of these topics, the computer searches its memory to identify the military trades satisfying the parameters set by the potential recruit. Furthermore, the computer will inform the individual of the impact of each new input immediately, and allow the applicant to change his mind and see the consequence of alternative inputs.

The computer and the applicant continue to converse with each other until they have arrived at a manageable number of trades from which the applicant must eventually make a choice. The number would ideally be five or fewer. The object of the exercise is to assist the applicant to make his own choice regarding the trade, or trades, most likely to meet his long-term career needs. The computer will suggest specific trades. If it does not include in its list of suggested

trades one which looks appealing to the applicant, the applicant can ask the computer why the trade did not appear. The computer will instantly identify the specific input items which caused it to reject the trade. Perhaps the applicant communicated an aversion to specific environmental or working conditions inherent in the trade. Applicants who used the Variable Data input function of SEARCH will leave the module with a "short list" of trades which look highly promising in terms of meeting their career needs. In all likelihood they will then want to go on the INFORM module of the system.

Module Two - INFORM. As its name implies, this module informs applicants about details of the CF generally and the military lifestyle - as well as providing access to detailed, comprehensive and up-to-date information about every trade in the forces. The module is sub-divided into four components: a Computer-Assisted Learning component, a Retrieve Function, a component providing detail on Specific Trades, and, a Trades Comparison feature.

The Computer-Assisted Learning (CAL) component of INFORM actually teaches applicants, in an individualized, programmed instruction format, about most major aspects of a career in the CF. It is conversational, in the sense that it requires feedback from the applicant to ensure that learning is occurring. Applicants are led through "segments" designed to inform them about, for example: the role of the CF, organizaion of the CF (Sea/Land/Air Elements as well as Operational/Support Elements), basic training (purpose, location, duration, daily schedule, hardships, pay, etc.), trade training, language training, military lifestyle (discipline, dress, deportment, conduct, esprit de corps, physical fitness, rigours and hazards, mobility, continuous readiness), rank structure, trade progression, pay levels, specialized training plans, re-assignment/remuster options, and benefit packages (medical/dental treatment, housing and recreational facilities, leave).

Each lesson is followed by a brief quiz to assure that the applicant understands what is being communicated to him. If so, the applicant is positively reinforced and encouraged to continue. If the applicant does not answer a quiz item correctly, his error is discussed and he is then encouraged to attempt the quiz item again. This continues until the question is answered correctly, at which point the applicant is then encouraged to continue.

By the time applicants have completed the CAL component of INFORM they will have learned a good deal about the CF. To ensure that they will not distort or forget that which they have learned, a hard-copy printout of the entire "conversation", including all of the lessons, is provided to the applicant to take away. A summary of what the computer covered with the applicant, including the applicant's success rate on the post-lesson quizzes, can be retrieved by the MCC at his convenience.

The Retrieve component of INFORM allows applicants to go directly to specific "lessons" in the CAL component without having to go through the entire programmed instruction sequence. If, for example, all an applicant wants to know about is one or more of the specialized entry level training plans, he does not have to work his way through all of the previous lessons in order to get to this information. The applicant is still quizzed on the information at the end of the specific segment(s) however, only to ensure that he has understood the material. CFRC staff can access specific segments in the CAL component through Retrieve as well, but they will not be subjected to the quizzes.

The Specific Trade component of INFORM allows applicants, or CFRC staffs, to obtain detailed information regarding individual CF trades, one trade at a time. The applicant simply identifies the trade about which he desires information, then tells the computer specifically what he wants to know about the trade. He may only want to know what the educational requirements are for entry, or what the posting prospects look like, or he may want to see a complete, detailed profile on the trade. Information such as the following will be available for every CF trade: primary duties and trade functions, range of working environments, hazards, physical demands, special hardships, posting/employment/promotion opportunities, pay, secondary duties, time away from home, exercises, civilian occupational equivalencies, major equipment, tools, facilities, initial training, education requirement, and personal requirements (interests, abilities, temperaments). As in the CAL component, applicants will take with them a printed record of this segment of their conversation with the computer which they can peruse at their leisure after leaving the CFRC.

The Compare Trades component of INFORM allows comparison of specific details on up to three different trades at the same time. An applicant might want to see, for instance, a side-by-side comparison of the pay, posting and promotion opportunities of three different trades from among which a selection must be made.

By the time applicants have made tentative decisions about the trades that may suit their needs (SEARCH), and have then obtained detailed information about the ones they consider most appealing (INFORM), they will be in a good position to decide which, if any of them would be most satisfactory. Moreover, after exposure to the CAL component of INFORM, the applicant will have a much better idea of whether or not he wants to make a serious commitment to the military way of life. Virtually anything the individual might want to know about the Canadian Forces that would help him make a viable decision will be at his fingertips. Certainly no one will be able to say he didn't know what he was getting himself in for.

It should be pointed out that while the four components of INFORM will inform potential recruits like none have been informed in the past, the components could be complemented by other materials, such as video-cassettes covering the "lessons" of the CAL component and specific trades. Sound/slide, or film strip presentations are alternatives, as are micro-form materials with colour photographs. Visual images can be extremely efficacious in forming realistic perceptions and expectations.

Module Three - INQUIRE. INQUIRE is a "for office use only" module in the sense that applicants will never have access to its contents. It allows appropriate individuals to quickly retrieve data about quotas, waiting lists, individuals, and the success of the recruiting, trade assignment and training process. Inquire is therefore sub-divided into three components: a Quota Management function; Data Storage and Retrieval on individuals; and Recruiting, Selection and Training Process Statistics.

The Quota Management function of INQUIRE will maintain a totally up-to-date record of the quota status of any trade in the CF. As any change occurs in a trade's status the new data can be on-line almost immediately, anywhere in Canada. Thus, whenever this file is accessed the inquirer will see data which is completely up-to-date, not weeks or months old. Quotas can be maintained by trades, trade groups and elements, by office, district, region, command, province and nationally,

with requirement projections broken down on a daily, weekly, monthly, or annual (and beyond) basis.

Waiting list data will be stored for individuals who appear to have adequate military potential, and who satisfy all requirements for the trade for which they have been "wait-listed". The individual's position on the list will be a function of his military potential rating assessed by the MCC, as well as factors such as mother tongue, length of time on the waiting list, age, province of origin, and so forth.

At any time, CFRC staff will be able to see where they stand in terms of meeting their quotas. Furthermore, they will be able to quickly identify individuals satisfying whatever set of characteristics they desire (ie; language, birth place, height, weight, visual acuity, etc.) from an automated waiting list storing data on all potentially desirable applicants. Exchanges between CFRCs will be greatly facilitated. As well, NDHQ personnel will be able to obtain a "snap-shot" of the status of the recruiting and selection process, for the entire country, or for specific areas, or offices, which is accurate at the time the picture was taken. They will be more in control of the flow of personnel into the CF, both quantitatively and qualitatively.

The Individual Data Storage function of INQUIRE will allow qualified people to access detailed data on all individuals either in the CF, in the process of trying to enter the CF, or recently released form the CF. This function represents a merging of the automated data gathering facility provided by the preceeding SEARCH and INFORM modules with the quality control, file maintenance, data storage and retrieval capabilities already extant in CFPARU's Research Information System (RIS). Data could be stored on disk, on magnetic tape, or in fact on micro-fiche off-line (or for that matter, on all three).

The Statistics function of INQUIRE has tremendous possibilities from a research and development perspective, as well as a direct aid in the formulation of CF policy regarding recruiting, selection, training and employment. Essentially this function exploits the availability of comprehensive data on CF personnel in the Quota Management and Individual Data Storage functions of INQUIRE. Using this function, MCCs will be able to quickly obtain precise feedback on their "success rates" in making trade assignments and in accurately assessing military potential. On a more global basis, however, it will be possible for NDHQ recruiting, selection, training and other staffs to look at issues such as: (1) trade assignment vs success rates in basic and trades training by trade and region; (2) attrition rates (overtime by trade, geography province, region, CFRC, BPSO), element, sex, edcuation, language, and (3) test results vs "success" in basic and trades training and at the unit by trade, province, region, CFRC, element, sex, etc.

These types of studies are currently conducted on a regular basis by various interested directorates. The advantage, however, of doing the job via CFCIS is that data availability is optimal and all data are totally up-to-date. So rather than preparing a report today using last year's (or the year's before) data, today's report will now reflect today's data. This facility can make the recruiting, selection and training systems extremely responsive to variations in the pool of recruitable-aged people in Canada, as well as to the changing realities of the personnel pool currently within the CF.

PROPOSED COUNSELLING PROCESS FLOW AT THE CFRC

The applicant's initial suitability is determined at the counter. Unless there are obvious reasons for terminating the contact, the applicant would be encouraged to view a video-cassette (approximately 20 minutes) which would provide, in a very realistic fashion, a general introduction to the CF. After the applicant is finished watching the video-cassette, he would return to the counter. At this point, he will be in a position to decide whether he is interested in continuing processing. If appropriate, processing would then continue with the "fact-finding" interview to obtain biographical documents. This could also be conducted on-line at the computer terminal. However, this approach presents complexities which should probably be avoided, at least intially.

The applicant would then be administered the General Classification (GC) test and the full Canadian Forces Classification Battery (CFCB). The medical follows. Obviously, failure to meet the minimum cut-off on the GC or to meet acceptable medical standards would result in termination of processing with counselling. All of the fixed data obtained would then be fed into the CFCIS computer by a clerical staff person in the CFRC. This process would take approximately five minutes.

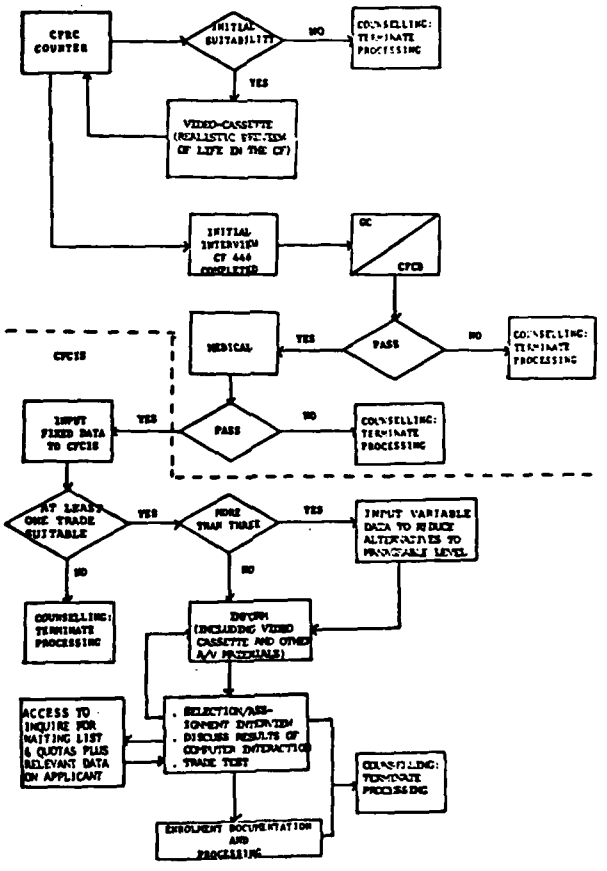


Figure 2: Proposed Counselling Process Flow at CFRCs

(105)

Once the computer has been informed that it has all the data that it is going to get regarding the applicant, actual processing begins. In a matter of seconds, the results would begin to appear on a hard-copy printout. All trades for which the applicant meets the requirements would be listed, in order of their apparent suitability.

If the applicant does not statisfy the requirements of any of the CF trades, counselling will occur and processing will likely be terminated. If, on the other hand, three or fewer trades are listed, the applicant will probably want to obtain information about them in order to decide his order of preference. At this stage the applicant would be encouraged to access the INFORM module of CFCIS to obtain the desired information, in a hard-copy format which he can take home. Again, ideally at this point, the computer would suggest to the applicant that he also view video-cassettes or other A/V materials on the trades being considered.

Of course, the applicant would have access to Specific, to obtain detail on a single trade, to Compare for comparison of up to three trades on selected dimensions, to CAL in order to learn considerably more about the CF, or to Retrieve to focus on particualr aspects of the military lifestyle.

If, after the fixed data regarding the applicant was entered, more than three trades were listed by CFCIS as potentially suitable, the applicant would be encouraged to enter variable data about himself to reduce the number of trade alternatives to a manageable range. Those who are in this position would be instructed to prepare themselves for this particular interaction with the computer by reading a User Guide (probably overnight). This would allow them to make some informed decisions about their interests, values, aspirations, preferences, and so on, before submitting this information to the computer. When they have read the User Guide and reached tentative decisions about what they want to say to the computer, they would have a counselling session with the MCC. The object of this session would be to ensure that the applicant has clearly understood the variables in question and that he has, in fact, given serious and reasoned consideration to the types of things that would contribute most heavily to his career satisfaction.

The end-product of this computer session, likely to last approximately 15 to 30 minutes, is a short-list of highly suitable trades for the individual. At this point, he would proceed to the INFORM module of the system, as described above, in order to learn more about each of the trades so that he could arrange them in order of personal preference, and to learn more about whether he wants to commit to a career in the CF.

The applicant and MCC would then proceed with the selection/assignment interview where both would have further access to the SEARCH or INFORM modules of CFCIS if necessary. The MCC would have rapid access to the INQUIRE module for data on the individual (including a summary of this exposure to CFCIS) and quota/waiting list data as required.

REFERENCES

Canadian Forces Training Systems Statisical Review. 2nd Quarter, 1979/80.

Ellis, R. T., & Saudino, D. A. Selection and trade assignment (Men): Status report. (CFPARU Working Paper in preparation). Toronto: Canadian Forces Personnel Applied Research Unit, 1980.

Horner, S. O., Mobley, W. H., & Maglino, B. M. An experimental evaluation of the effects of a realistic job preview on marine recruit affect, intentions and behavior. Navy All-Volunteer Manpower R&D program of the Office of Naval Research N00014-C-0938. University of South Carolina, 1979.

Ilgen, D. R., & Seely, W. Realistic expectations as an aid in reducing voluntary resignations. Journal of Applied Psychology, 1974, 59, 452-455.

Ilgen, D. R. The influence of expectations and beliefs on the motivation and adjustment of new members of military organizations. Paper presented at "Psychology of Military Service" Conference, April 23-25, 1975; Center for continuing Education, University of Chicago, Chicago, Illinois.

Jarvis, P. and Hutt, R. The Canadian Forces career information system: Functional analysis report. (Report 80-9) Toronto: Canadian Forces Personnel Applied Research Unit, 1980.

Katz, D. Psychological barriers of communication. Chapter in De Vito (1971) from Psychological Barriers to Communication, Annals of the American Academy of Political and Social Science, CCL (March, 1947), 17-25.

Katzell, M. E. Expectations and dropouts in schools of nursing. Journal of Applied Psychology, 1968, 154-157.

Martin, M. A. The reliability and validity of Canadian Forces selection interview procedures. (Report 72-4). Toronto: Canadian Forces Personnel Applied Research Unit, 1972.

Rampton, G. M., Skinner, H. A., & Keates, W. E. Selection and trade assignment (Men) project status report. Report 72-7. Toronto: Canadian Forces Personnel Applied Research Unit, 1972.

Sinaiko, H. W., & Scheflen, K. C. Personnel attrition in the U.S. armed services: Some examples of information analysis. Prepared for TTCP(U) Technical Panel UTP-3 meeting in Williamsburg, Virginia, 1979.

Skinner, H. A., & Rampton, G. M. Evaluation of the personality research form (PRI) in a military environment. (Report 73-5). Toronto: Canadian Forces Personnel Applied Research Unit, 1973.

Super, D. E. Vocational development in adolescence and early adulthood: Tasks and behaviors. In S. H. Osipow, Theories of Career Development (2nd). New York: Appleton-Century- Crofts, 1973.

Tierney, E. C., & Pinch, F. C. Military implications of socio-demographic and related changes in the 1980s and 1990s. CFPARU (Working Paper 80-4). Toronto: Canadian Forces Personnel Applied Research Unit, 1980.

Van Maanen, J., & Schein, E. Toward a theory of organizational socialization. In B. Staw (ed.) Annual Review of Research in Organizational Behavior, 1, New York: JIP Press, 1978.

Wanous, J. P. Tell it like it is at realistic job previews. Personnel, 1973.

Wilson, F.P. Towards a more systematic counselling model for the Canadian Forces. (CFPARU Working Paper 80-3.) Toronto: Canadian Forces Personnel Applied Research Unit, 1980.

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

1054

WILSON, Capt. Peter W., Canadian Forces School of Aerospace & Ordnance Engineering, CFB Borden.

IMPLEMENTING A COMBINED CRITERION AND NORM-REFERENCED TESTING SYSTEM (Tue A.M.)

In response to an institutionally perceived need for improvement in testing and also to a redesign of all the air-affiliated trades, a new exam system was developed at the Canadian Forces School of Aerospace and Ordnance Engineering. Since the school was reluctant to give up the information derived from a norm-referenced testing system and since those developing the new exam system wished to have the advantages of criterion-referencing, an attempt was made to implement a system which combined the best features of both. The norm-referenced aspect of the system accounted for the pass standard, scoring and results reporting as well as statistical analysis used as a limited feedback to question form. The criterion-referenced portion influenced exam development, exam content (as opposed to item) analysis and feedback to the instructional system. The problem of setting a pass standard which is fair to both aspects of the system is discussed.

IMPLEMENTING A COMBINED CRITERION AND NORM-REFERENCED TESTING SYSTEM

Captain Peter W. Wilson

Canadian Forces School of Aerospace And Ordnance Engineering

NTRODUCTION TO THE PROBLEM

In any large training institution the forces of change for improving rocedures seem to work very slowly and in some cases lose ground as policies nd personnel change. At the Canadian Forces School of Aerospace and Ordnance ngineering (CFSAOE), a technical training school with 550 staff, 3000 students nd 90 different courses of instruction, the concept of performance based raining had been applied differentially for several years for practical per- rmances. Because of this there was some understanding of the need for cri- erion testing of the physical skills learned on the various courses but there as little concern for applying that same philosophy to the evaluation of in- ellectual skills and the knowledge requirements measured by written examina- ions. In fact, there seemed to be a great and ineradicable resistance to the tion of criterion-referenced testing (CRT).

It was apparent that there was a need for both types of testing; the riterion-referenced, to ensure that the requirement for the retention of the upporing knowledge for practical skills was met and the norm-referenced to atisfy the military requirements for differential assessment of trainees. he question then became, "Can a workable exam system be devised which would aximize the benefits of both criterion and norm-referenced examinations nd minimize what is lost through compromise?" i.e. Could we obtain all or ost of the information we needed without double testing?

The basis for the application of criterion referencing was available n the training development system through accepted procedures of determining hich skills were necessary to accomplish specific jobs and in the orientation f the training to objective checking of skills against performance criteria. he military performance evaluation system required a comparative assessment f individuals so that rewards for superior performance could be applied. hus the need for norm-referencing.

XPLORING THE BASES FOR COMPROMISE

Clear definitions of what types of testing we were working with were ecessary and it became apparent after some searching that it was criterion- eferenced and not "objectives-referenced" testing that we wished to do. opham's (1975) definition as a test "used to ascertain an individual's status ith respect to a well-defined behavior domain." (p. 130) was acceptable and eminded us that the knowledge domain which was being tested would have to e completely rationalized with a particular behavioral performance to satisfy he definition. Robert Glaser has also given a good definition (1971) "A riterion-referenced test is one that is deliberatley construct d to yield easurements that are directly interpretable in terms of specified performance tandards." (p. 41)

1057

The norm-referenced test has been acceptably defined as one which is designed to determine a student's performance relative to others who write the same or an equivalent test.

In comparing both types of testing and reducing the bases of the difference between them, it cannot be said that the behavior criterion is not derived from a selection of relative behaviors, because it is. The important distinction is that those relative behaviors are outside the testing situation and not within it as in norm-referenced testing. In fact, the criterion behavior has already been selected on the basis of an evaluation of the behavior of several individuals, through successive approximations, until the minimum acceptable behavior is established. So the behavioral standard is essentially a norm external to the testing situation.

It has been stated in Hambleton, et al (1978) that, "A criterion-referenced test constructed by procedures especially designed to facilitate criterion-referenced measurement can and sometimes is used to make norm-referenced measurements." (p.3). So it can be seen that the possibility, though unspecified, was there for creating such a combined testing. They went on to give a caveat that the use of a criterion-referenced test for norm-referenced purposes was not particularly satisfactory because it is "not constructed specifically to maximize the variability of test scores" (p.3).

In a test which can be evaluated in two ways there is a need to specify exactly what information can be gained and what that information will tell us about the student, the instructor, the instructional environment, the material being learned and the test itself. Criterion-referenced tests can tell much about the success of the student in relation to a specified behavioral goal. They are also useful in determining how successful an instructor was in a given situation although because of the variable of student ability it is difficult to generalize to the competence of the instructor based on only one group of students or one testing. Over several groups and several tests CRTs are a useful instrument on which to base an instructor accountability program. As Smythe, et al (1973) state, "When a specified criterion or performance standard is used as the basis for determining student achievement, the efficacy of instruction is easily ascertained" (p.2) and with some reservations about the ease of this, the possibilities are good. In our use of CRTs we also wished to attempt an analysis which would give some indirect information about the suitability of the instructional environment and the nature of the material being learned. Most of the information gained from the norm-referenced analysis was to be used to "red-flag" poor questions and to help improve format.

To gain the most useful information from the criterion-referenced evaluation the tested material had to be directly related to the appropriate behavior in the task analysis stage of the course development. The test would then reflect the knowledge base of the performance required of the student. The main compromise that might seem to be required with the norm-referenced test is the measure of student "success". If the minimum knowledge required is the basis of the CRT then the pass standard must be high so that the student has enough knowledge to do the associated performance.

In our analysis we found several areas where slippages from the "minimum" knowledge occur. In the task analysis, course design and lesson development, "nice to know" or marginally critical knowledge slips in and tends to be included on the test. The requirement to provide a flexibility of technical competence in order to repair various types of equipment which have the same function shows up as a need to give more general instruction and use a specific equipment type as a training medium only. This further loosens the concept of "minimum" knowledge. The problem of information retention is also involved. If a criterion test measures the knowledge shortly after it is learned the retention ratio will be higher and the criterion more likely to be achieved. However, three months later some part of the knowledge has been lost and the technician is still required to use it as he begins working in his trade. If instruction to minimum competence is given and only that is tested then the student will know less than he is required to as he begins his work. Based on this rationale and the knowledge that non-essential material has creeped into the course content the possibility of rationalizing a lower criterion level than, say 90% of the class achieving 90%, could be allowed. Since the basis of student success would be determined with a norm-referenced standard, the standard had to be high enough to ensure a reasonable criterion was reached and low enough to give some mark spread. So 70% was selected as the minimum standard of success.

Fortunately, fine discriminations among student performances are not required by the military personnel system. Since we live with the training product there is little need to provide them. Other measures of practical skill and personal suitability to a given trade are taken and the final grade of A,B,C or F on a given course is modified by these measures. The school staff did not object to the level of the standard and felt that it was achievable by most students.

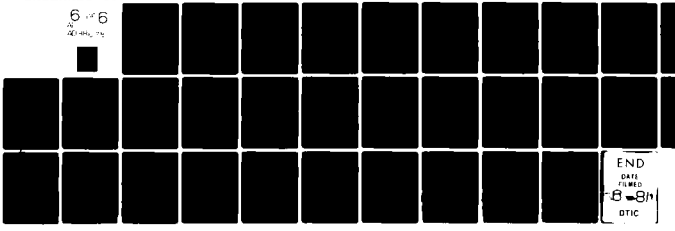
METHODOLOGY OF TEST SYSTEM DEVELOPMENT AND IMPLEMENTATIONS

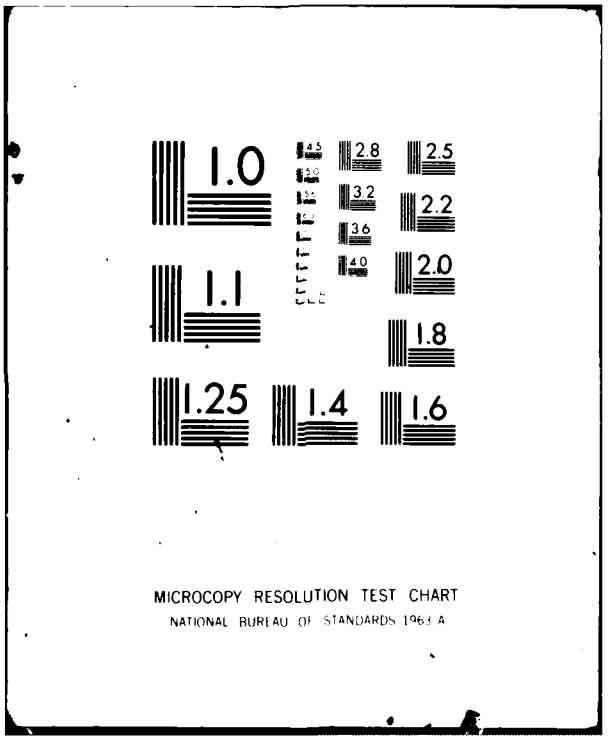
Test Development

Following is a brief description of the test development procedure. From statements of performance criteria, conditions and standard (à la Mager) which were derived from task lists, teaching points were selected by trade specialists which exhausted, to the level desired, the content of each performance. The teaching points were in turn broken into sub-teaching points of considerably greater detail and these were listed. To the greatest extent possible the compartmentalization of the task elements which these teaching points reflected was avoided by ensuring that the sum of the parts produced at least the whole from which they were derived. This ensured that critical environmental cues and other information which normally "falls through the cracks" an any analysis were retained. The specific reference which served as a basis for both lesson planning and test item construction was noted and the teaching point or sub-teaching point was specified as knowledge which was either critical or not critical to the ultimate achievement of the performance.

AD-A098 678 MILITARY TESTING ASSOCIATION F/G 5/10
PROCEEDINGS OF THE ANNUAL CONFERENCE OF THE MILITARY TESTING AS--ETC(U)
DEC 80
UNCLASSIFIED MTA-22-80-VOL-2 NL

6 of 6
AD-A098 678




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

As instructors developed lesson plans to the level specified in the detailed references, the question writers determined the required level of test questions through a Question Identification Index developed specifically for this purpose. Since multiple-choice questions would comprise the test we presumed that only three levels of intellectual activity could be effectively tested; recall, understanding and analysis. The Index related each of these levels to the verb used in describing the required performance, e.g. the verbs "state", "list", etc., would denote the need for questions which tested recall; the verbs, "explain","discriminate",etc., those which required questions to test understanding and the verbs "analyze", "troubleshoot", etc. reflected need for analysis. Since all objectives required the need to recall information, those objectives which expressed the requirement for higher levels of intellectual skill would only change the proportion of recall questions. One objective which required the understanding level might have 80% recall questions and 20% understanding questions for one subject area but the proportion might change for another. The main use of the Index was to ensure that some appropriate proportion of the higher level questions were included. Once established for a subject area that proportion was usually fixed and subsequent item selection took that proportion into account. Finally, the writer ensured that each question written reflected the content in the references and approached the subject at a level appropriate to the task which the knowledge supported. Questions were written only for that knowledge considered "critical" to the support of the associated practical performance.

System Development

Since in neither criterion nor norm-referenced testing can every possible question be asked, a sample was taken which represented an adequate test of the knowledge necessary to the task. We decided that to ensure an appropriate selection was made we would use a stratified method of selecting exam questions, the teaching points being strata, and weighted (percentage of questions on test per teaching point) by the importance of each point to supporting the overall task rather than the amount of material studied. This weighting was subjective but deliberate. The selection of questions within teaching points was random. The exam was also weighted for type of question whether recall, understanding or analysis. The demands of the criterion-referencing aspect of the system were met by ensuring that the knowledge tested was referenced directly to the practical skills necessary to do the job. Another aspect of the testing system already in place was that students were retested in areas in which they had failed until they achieved the standard or were removed from the course.

Item banks for each performance objective were developed. As of writing, tests are still constructed manually awaiting implementation of an automated item selection system on an AES C-20 word processor. In this system,variables of trade, course and performance objective are keyed in to obtain a program specific to each objective. This selection program will contain the specifics of item selection criteria and test length will be alterable. The whole system serves 12 trades and comprises 25,000 items.

System Outputs

Tests are scored using mark-sense apparatus and the results are given in percentages. The pass/fail criterion is 70% and a rewrite/removal decision for each failure depends partly on factors other than score on the test but is usually a reflection of overall performance on the course.

Study time of at least 3 days is allowed before a test rewrite. The way the test is constructed allows certain analyses to be done which have implications beyond the test itself. A spread sheet of student names by question number specifies total numbers of students making selections of each wrong choice in each question. Through a content analysis algorithm it is possible to analyze this and draw conclusions regarding which teaching points are poorly taught or not taught at all and which questions may be faulty (e.g. the whole class selects one distractor). Looking at several spread sheets of the same content area tested over a number of courses shows instructor weaknesses or a lack of cooperation in implementing recommended changes. Ongoing problems uncorrectable by the trainers would result in an investigation and perhaps direct monitoring of classes by an instructional designer who would identify weak instructors or weak design in lessons and make appropriate suggestions. Thus the test results give those who monitor instruction and student performance on a regular basis a stimulus for intervention in a specific problem area.

Test results are also analyzed statistically to give the conventional discrimination indices, difficulty level, and reliability/validity measures. As Hambleton and Novick (1973) state, "it seems clear that the classical approaches to reliability and validity estimation will need to be interpreted more cautiously (or discarded) in the analysis of criterion referenced tests" (p. 167). Because test items were not selected on their discrimination ability these measures were taken rather lightly although the difficulty level had some value in identifying poor questions.

RESULTS OF SYSTEM IMPLEMENTATION AND DISCUSSION

In addressing the problem of whether a testing system can be devised which will satisfy the need for both instructional system (external) and test system (internal) feedback as well as providing information on how well the student performs in relation to his fellow students and the specific job requirements, some compromises were required. However, the following results were achieved.

The requirement of the personnel evaluation function for a test score which discriminates was satisfied. On tests given to date, class averages range from 64% to 94% with the mean class average somewhere in the low 80% range. Individual student marks usually have a spread of at least 20 to 25 percentage points. The trainers are accepting the fairly high pass standard and it has stimulated instructional design improvements and increased student learning activity to meet the standard. Combined with pass/fail practical performance evaluations and more subjective "personal quality" evaluation, the test score is an acceptable differentiator among students. (As an aside, it is interesting to note that in the military which preaches training, assessment of a person's personal qualities and suitability for military life form an integral part of his overall "score" on a training course whereas in the field of education where the function is to educate the whole person, very little of his response to this education is reflected in the person's school grades). As a method of determining student success in relation to an external criterion the passing grade may be a little too low so that some pass through with lower grades who may have been selected out in a purely criterion-referenced system. However, lack of manpower and high attrition rates in the technical trades as a whole mitigate against using initial trades training courses as stringent means of personnel selection. To reduce Type 1 selection errors, the system seems willing to carry personnel as long as possible and, except for obvious problem cases, have them virtually select themselves out.

Of course, the true state of affairs in the actual relationship between test scores and journeyman competence is not clear. One study by Carpenter-Huffman and Rosther (1976) found that much of the theory portion of a highly technical training course in avionics was irrelevant to the actual skill needed to be performed on the job. Thus it would not matter how accurately such knowledge was measured if it told us little about performance capability. In any case, test scores are very high on the whole and provide confidence in the students' having met or come very close to a typical criterion score.

Until automation of the exam production portion of the system is accomplished we have been reduced to providing limited external feedback to the instructional system. Even with this limitation, analysis has led to the re evaluation of the level of instruction of the digital techniques package which is included on several courses. It has also led to changes in course training plans and lesson plans. Using the feedback as a measure of instructor competence is more problematical, the tendency being to blame the class for being below par in some respect. As more classes are instructed by a given instructor, the validity and reliability of decisions made regarding competence increases and poor instructor performance can be isolated. This might also be done by comparing class results as they go through various phases of a course but different levels of difficulty on each phase and any tendency of a class to be typed as poor students would make this method unreliable.

The statistical measures to be obtained will be of some value in determining atypical items among those measuring the domain. The spread of marks is not so narrow that the validity of the measures will be greatly affected. Test items will not be omitted because they do not discriminate between those doing better and those doing worse on the test overall. Whether or not the question discriminates well is only interesting for what it tells us about the nature of the material learned, the wording of the question and the method of instructing the material. Since we cannot assume that all information is homogenous with respect to the best method for presentation, neither can we assume that all material presented is most easily learned by those who score the highest on the test as a whole. A negative discriminating item may provide us valuable information about the nature of the intellectual skills required in various parts of the task. If the test uses mainly memory items and the negative discriminator tests understanding or ability to analyze, this tells us something about the nature of the intellectual basis of the task and in turn can feed back into the trade selection criteria and into the design of the instruction. This information would be discarded in a norm-referenced test. Investigation of difficulty value and other statistical measures with their implications for criterion-referenced instruction is ongoing.

Conclusion

It is possible and practical to derive both norm and criterion-referenced measures from the same test. At CFSAOE the possibility of using criterion based instructional system design techniques provided an opportunity to establish a sound criterion-referenced testing system. The further requirement for a norm-referenced determination of student success on the course led to successful implementation of a testing system which took advantage of the benefits of both criterion and norm-referencing. The compromises required to establish the system may have had some detrimental effects

on the overall quality of students who achieved mastery as compared with a purely criterion referenced system but certainly provided those used to traditional testing with the sort of test outcome they were familiar with. Besides, it gave feedback into the instructional design of the course and a reading on instructor capability which was not present before.

References

Carpenter-Huffman, P. & Rostker, B. The relevance of training for the maintenance of advanced avionics. Project Air Force Office, Washington, D.C. 1976.

Glaser, Robert. in Criterion Referenced Measurement (an introduction), James Popham (Ed.) Educational Technology Publications, Englewood Cliffs, N.J., 1971.

Hambleton, R.K. & Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement. 1973, 10, 159-170.

Hambleton, R.K., Swaminathan, H., Algina, J., Coulson, D.G. Criterion referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, Vol. 48, No. 1, Winter, 1978.

Popham, W.J. Educational evaluation. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

Smythe, M.J., Kibler, R.J., Hutchings, P.W. A comparison of norm-referenced and criterion-referenced measurement with implications for communication instruction. The Speech Teacher, Vol. XXll, No. 1, January, 1973.

1063

SYMPOSIUM

COMPUTERIZED TESTING SYSTEMS: DESIGN FROM THE USER'S PERSPECTIVE

Introduction

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

It has been nearly 10 years since the first large scale research program on on-line computerized testing began at the University of Minnesota. It has been more than five years since the first large scale computerized testing system for routine psychological assessment was implemented at the VA Hospital in Salt Lake City, Utah. In the years since computerized testing began, the state of the art of computerized testing has reached the point where the focus of research has shifted from questions of how well the technology will work to questions of how best to implement it. Many of the questions regarding implementation center on considerations of testing system design. To design an effective system, it is essential that the system designer know what system configurations are most efficient for accomplishing what the user wants. It is also essential that the user know what s/he wants the system to do. This symposium is aimed toward the user. Its intent is to suggest what it is the user may want.

This symposium consists of four papers. The first by Johnson and Novak presents a human engineering perspective on the design of a computerized testing station. This station is the element of the testing system with which the examinee must interact. It is essential that the testing station be engineered in such a way that it does not threaten the examinee, confuse him/her, or induce extraneous test bias. The first paper presents some considerations relevant to this.

The second paper by Prestwood suggests how the computerized testing system might interact with the test proctor to make his/her job easier. It is the proctor's job to see that everyone has an optimal environment in which to take the test and that no one has an unfair advantage in the testing process. Prestwood's paper suggests how a computerized system can perform some of the proctor's typical duties and how it can perform additional tasks impossible with conventional testing.

I will present the third paper. This paper will discuss design of the test development subsystem. It is through this subsystem that the individuals responsible for creating the computerized test communicate their test designs. Different approaches to item entry and maintenance and to specification of testing strategies are covered.

1064

Finally, Hansen will present a discussion of various ways computer hardware can be networked together to achieve maximal reliability and cost effectiveness. The general concept of a network will be presented and several different types of networks along with their various strengths and weaknesses will be described.

It is the intent of this symposium that all considerations discussed be approached from the user's perspective. The presenters have all been in the role of the user and have been faced with the task of how best to design a system for their specific psychological testing problems. It is this experience that we hope to communicate in this symposium.

NOTE: The papers presented in this symposium are reproduced in these proceedings under the names Johnson, Prestwood, Vale, and Hansen respectively.

SYMPOSIUM

WOMEN IN THE MILITARY

Introduction

Major Franklin C. Pinch
Canadian Forces Personnel Applied Research Unit
Willowdale, Ontario

It is almost a truism to state that changes in society at large impact on the military, and this is particularly the case of militaries that depend on voluntary systems of personnel acquisition, such as the United States, Canada and Great Britain. One of the most far-reaching of these changes in the past decade has been the changing role of women in society. In most western nations, but most especially in the United States and Canada, women have moved from their traditional roles of mother and homemaker to active participants in the labour market, and it is projected that female labour force participation rates will equal those of males by the end of the 1980s.

Increased participation rates have also been apparent within the U.S. and Canadian military: for Canada, from less than two percent in the early 1960s to around seven percent at present (the percentage for the U.S. forces is similar). Quite apart from any individual or group aspirations among the women themselves, these increases have resulted both from demographic trends, and from legislation (the Equal Rights Amendment in the United States in 1973, and the promulgation of the Human Rights Act in Canada in 1978). As regards the former, we are witnessing a dramatic decline in the number of young men available for military service; therefore, an increase in female recruitment is essential if current and projected manning levels are to be sustained.

In Canada, at least, legislation prohibits the exclusion of women from employment in the so-called "nontraditional" military occupations, unless such exclusion can be justified by a bona fide occupational reason. (The U.S. case is similiar, though not identical). Owing to the requirements both for broadening the military recruitment base and for meeting the requirements of the law, the Canadian Forces is currently conducting a series of evaluations, to determine the effects of employing women in close support (i.e., noncombatant) roles, with the Navy, Army and Air Force, and in isolated locations. Preliminary analyses of some of the data collected are presented in one of the papers of this Symposium.

It is not quite clear in what ways increased women's participation will impact on the military in the long term: there is no solid evidence to suggest that operational effectiveness - the main dependent variable - will be either increased or impaired. However, it can be expected that there will be other social consequences, some as yet to be identified, as well as unique problems to be solved. Given the perceived need to maintain current manning levels and the urgency attached to the expansion of women's roles by virtually all western militaries, the topic is considered both timely and important.

This Symposium consists of three papers, followed by a formal discussion of the issues raised. The first, by Captain Simpson, presents a model for investigating the career expectations for and of women, as well as the determination of the extent to which either a positive or negative bias exists in Canadian Forces' personnel evaluation procedures with respect to women. If women are to be given equal opportunity of employment in the military, it is important that they be neither over-rated or under-rated on the basis of stereotyping or their ascribed characteristics; rather, assessment of their actual performance of tasks should determine career progression and career outcomes. Captain Simpson plans to use previously collected data from performance evaluation files and data currently being collected in order to illuminate this important area of research.

The second paper, by Lieutenant Boyce and Captain Belec, offers a preliminary analysis on the attitudes of the roles of women in society by several groups of civilian and military men and women, using Spence and Helmeich's "Attitudes Toward Women Scale". It would appear that these attitudes are associated with gender, educational level, linguistic affiliation and a number of other background variables. Whether the attitudes held will be implicated in the Canadian Forces' evaluations of women in near combat environments remains a matter for speculation and further behavourial research.

The third paper is presented by Ms. Lipscomb, and sets the stage for data analysis not yet available. It provides the basis for examining the patterns of participation for women in technical (traditional) male jobs within the U.S. Air Force. There is some suggestion that even though women are capable of performing these technical jobs, the influences within the environment - both group and individual - may see women, over time, moving back into more traditional (e.g., clerical) jobs. The importance of understanding the factors underlying this trend, and of assessing its implications for the participation of women, cannot be overstated.

Finally, Dr. MacFarlane will provide some linkages of the issues surrounding the expansion of women's roles within the military with those of society at large. He offers information that permits a more balanced view of societal and military trends, and raises some problems not addressed in the papers presented. It is hoped that these papers and the discussion arising from them, will provide an adequate basis for what should be a useful interchange among Symposium participants and attendees.

NOTE: The papers presented in this symposium are reproduced in these proceedings under the names Simpson, Boyce, Lipscomb, and MacFarlane respectively.

PUBLISHING REVIEW GROUP (PRG) EDITORIAL BOARD

An abstract outlining PRG goals and activities will be found in the "Contributed Papers" section on page WAL-2-O. Further details are included in the Minutes of the Steering Committee Meeting. MTA members who can assist the PRG in any way should contact the appropriate board member of those listed below:

Dr. Raymond O. Waldkoetter Army Research Institute Fort Sill Field Unit Fort Sill, Oklahoma 73503	PRG Chairman and Senior Editor
Dr. Walt[illegible] E. Driskill Chief, US Air Force Occupational Analysis Program USAFOMC, Randolph AFB, Texas	Associate Editor Occupational Analysis and Research
Dr.B. Micheal Berger Chief, Occupational Survey Division US Army-Soldier Support Centre 200 Stovall St., Alexandria Va 22332	Associate Editor Occupational Analysis and Research
LCol G.M. Rampton Commanding Officer Canadian Forces Personnel Applied Research Unit 4900 Yonge Street, #600, Willowdale, Ontario, M2N 6B7 Canada	Associate Editor Personnel Measurement and Evaluation
Dr. Martin Wiskoff Program Director Navy Personnel Research and Development Center 5151 Bixel Drive, San Diego, Ca	Associate Editor Personnel Measurement and Evaluation
Dr. Hendrick W. Ruck Chief, Skill Requirements Air Force Human Resource Lab 8814 Pertshire, San Antonio Texas 78250	Associate Editor Training Methods and Programs
Maj R.T. Ellis (DPSRSC-3) National Defence Headquarters Ottawa, Ontario, K1A OK2 Canada	Associate Editor Training Methods and Programmes

(1068)

Dr. Arthur C.F. Gilbert Senior Research Psychologist US Army Research Institute 5001 Eisenhower Avenue Alexandria, Va 22333	Associate Editor Organizational Assessment and Technology
Lt. Col. Wm. Hendrix AFIT/LSS Wright Patterson AFB, Ohic 45424	Associate Editor Organizational Assessment and Technology
Col H.E. Seuberlich Member-Chairman Army Section Federal Armed Forces Assoc. (D) Sudstr 123 Bonn 2 Federal Republic of Germany	Associate Editor European (Co-ordination)
Dr Klaus J. Puzicha Streitkraefteamt. Dezermat Wehr-psycholologie Regierungadirektor Post Box 205003 D-5300 Bonn Federal Republic of Germany	Associate Editor European (Co-ordinat[illegible])

Note: Associate Editors for "Organizational Background and Functions of the MTA" and for "Military Personnel Research: Review and Projections" will be announced in due course.

MTA STEERING COMMITTEE MEMBERS

1. Naval Personnel Research and Development Center
2. Naval Education and Training Program Development Center
3. Army Research Institute
4. Air Force Human Resources Laboratory
5. Air Force Occupational Measurement Center
6. U.S. Coast Guard Institute
7. Canadian Forces Personnel Applied Research Unit
8. Canadian Forces Directorate for Military Occupational Structures
9. Royal Australian Air Force Evaluation Division
10. German Armed Forces Association
11. German Armed Forces Psychological Services Research Institute
12. Belgium Armed Forces Psychological Research Section.

(1020)

MINUTES OF MTA STEERING COMMITTEE MEETING
HELD IN TORONTO - 27, 29 OCT 80

13 November 1980

IN ATTENDANCE:

Members of the MTA Steering Committee and Representatives of the Publishing Review Group (PRG).

INTRODUCTION

1. The Steering Committee Meeting was opened with the introduction and welcoming of Dr Arnold Bohrer, The Commandant of the Psychological Research Section of the Belgium Armed Forces. Dr Bohrer's organization was accepted as a member of the MTA by a unanimous vote of the Steering Committee. The Chairman, LCol Rampton, confirmed that formal applications of additional countries would also be accepted and ruled upon by the MTA Steering Committee, but that the accepted norm is that MTA members would not solicit membership.

ITEM	DISCUSSION	ACTION BY
I	MTA MANUSCRIPTS	
	2. The Director, Military Occupational Structures (DMOS), co-hosts of this 1980 conference, submitted a standardization guide for MTA manuscripts. Committee members agreed that a guide was necessary and will review the document and provide the Chairman of the 1981 MTA conference feedback and recommendations in time for the 1981 MTA call for papers. Dr Wiskoff and Dr Burt will assist ARI in synthesizing the recommendations.	Col Hart Dr Wiskoff Dr Burt

ITEM	DISCUSSION	ACTION BY
II	HARRY GREER AWARD/FUTURE ANNUAL CONFERENCES	
	3. The Chairman reported that there had been no returns from letters sent to delegates requesting nominations for the Harry Greer Award. The nominations for 1981 are to be forwarded to Colonel Franklin A. Hart, Commander of the US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia; President of the 23rd Annual Conference. He was represented at this meeting by Dr Arthur Marcus.	Col Hart
	4. The representatives from the Federal Republic of Germany reconfirmed their offer to host the 1983 MTA Conference, which happens to fall on the MTA's Silver Anniversary. Col Seuberlich of the Psychological Department of the German Armed Forces noted the concern of committee members with respect to transportation to and from Germany. The matter will be investigated and the German delegation will obtain clarification prior to the 1982 conference.	Col Seuberlich Dr Puzicha
III	PUBLICATION REVIEW GROUP (PRG)	
	5. Dr Waldkoetter reviewed the progress of the Publication Review Group (PRG). Members will engage in the production of a scholarly scientific book, which will review the principal content areas and critically integrate materials spanning MTA from its origin, through the projected 25th MTA Conference and Silver Anniversary. The Chairman suggested a reconvening of the PRG through the Steering Committee on 29 Oct in order that text guide-lines and publication dates be laid down.	LCol Rampton
	6. The meeting was adjourned at 1130 hrs, and was reconvened at 1700 hrs, 29 Oct 80.	
IV	PRG TASK AND EDITORIAL BOARD	
	7. Dr Waldkoetter reported that subject to the approval of MTA Steering Committee, he, as Chairman and Senior Editor of the PRG, had appointed Dr Perrigo as Executive Advisor and Senior Associate Editor. Dr Burt volunteered to assist her. The remainder of the Task and Editorial Board were proposed as follows:	

(bq2)

ITEM	DISCUSSION		ACTION BY
	Section	Associate Editors	
	1. Organizational Background and Functions: The Military Testing Association	- Editors to be selected	
	2. Occupational Analysis and Research	- W. Driskill & M. Berger	
	3. Personnel Measurement and Evaluation	- G. Rampton & M. Wiskoff	
	4. Training Methods and Programmes	- H. Ruck & R. Ellis	
	5. Organizational Assessment and Technology	- A. Gilbert & H. Hendrix	
	6. Military Personnel Research: Review and Projections	- Editors to be selected	
	7. European (Coordination)	- H. Sueberlich K. Puzicha.	Dr Waldkoetter
	PUBLICATION GUIDELINES AND STANDARDS		
V	8. Publication standards for the document will be the APA Publication manual and/or the Annual Review of Psychology. In the establishment of deadline dates it was stressed that there should be sufficient time to thoroughly cover the areas for which each author is responsible. Bearing this in mind, the committee agreed that it would be judicious to have a timetable, although the dates may be amended through negotiation between authors and editors. Section abstracts are to be forwarded by Associate Editors to Dr Waldkeotter by 1 Apr 1981. Article Abstracts from MTA members to Associate Editors are requested by 15 May 1981. The first draft is requested to be forwarded to the editors five months after acceptance of the abstract. The first edit is to be completed by 1 Jan 1982, and the second draft is to be completed three months after receipt of the first editing. It was proposed to have the final draft by Jan 1983. Members agreed that this lead time was necessary if the book was to be completed for MTA 1983.		

ITEM	DISCUSSION	ACTION BY
VI	GENERAL DISCUSSION	
	9. Dr Waldkoetter requested members to assist him in locating proceedings for 1958, 1959, and 1969. The Committee agreed that a mid-year Steering Committee Meeting in 1981 may be worthwhile. Dr Burt will assist in coordinating that session.	MTA Members Dr Burt
	10. The progress and plans of the PRG were approved by a majority vote of the MTA Steering Committee.	
VII	MTA MAILING LIST	
	11. It was unanimously agreed by the assembled members of the Steering Committee that those persons who had not responded to the 1980 MTA mail-outs or indicated that they were no longer interested in being members of the Association, be dropped from the mailing list by CFPARU, before turning this list over to ARI. It was also agreed that ARI should develop a strategy for future editing of the mailing list, and that this would be proposed to the Steering Committee at the next conference.	LCol Rampton Col Hart
VIII	ADJOURNMENT	
	12. The meeting was adjourned at 1900 hrs.	
	E.L. Stenton Capt Secretary	
	G.M. Rampton Lieutenant-Colonel Chairman	

1024

BY-LAWS OF THE MILITARY TESTING ASSOCIATION*

Article I - Name

The name of this organization shall be the Military Testing Association.

Article II - Purpose

The purpose of this Association shall be to:

A. Assemble representatives of the various armed services of the United States and such other nations as might request to discuss and exchange ideas concerning assessment of military personnel.

B. Review, study, and discuss the mission, organization, operations, and research activities of the various associated organizations engaged in military personnel assessment.

C. Foster improved personnel assessment through exploration and presentation of new techniques and procedures for behavioral measurement, occupational analysis, manpower analysis, simulation models, training programs, selection methodology, survey and feedback systems.

D. Promote cooperation in the exchange of assessment procedures, techniques and instruments.

E. Promote the assessment of military personnel as a scientific adjunct to modern military personnel management within the military and professional communities.

Article III - Participation

The Following categories shall constitute membership within the MTA:

A. Primary Membership.

1. All active duty military and civilian personnel permanently assigned to an agency of the associated armed services having primary responsibility for assessment of personnel systems.

---

*As approved at the 1978 General Meeting of the Association 2 Nov 78, Oklahoma City, Oklahoma.

2. All civilian and active duty military personnel permanently assigned to an organization exercising direct command over an agency of the associated armed services holding primary responsibility for assessment of military personnel.

B. Associate Membership.

1. Membership in this category will be extended to permanent personnel of various governmental, educational, business, industrial and private organizations engaged in activities that parallel those of the primary membership. Associate members shall be entitled to all privileges of primary members with the exception of membership on the Steering Committee. This restriction may be waived by the majority vote of the Steering Committee.

Article IV - Dues

No annual dues shall be levied against the participants.

Article V - Steering Committee

A. The governing body of the Association shall be the Steering Committee. The Steering Committee shall consist of voting and nonvoting members. Voting members are primary members of the Steering Committee. Primary membership shall include:

1. The Commanding Officers of the respective agencies of the armed services exercising responsibility for personnel assessment programs.

2. The ranking civilian professional employees of the respective agencies of the armed service exercising primary responsibility for the conduct of personnel assessment systems. Each agency shall have no more than two (2) professional civilian representatives.

B. Associate membership of the Steering Committee shall be extended by majority vote of the committee to representatives of various governmental, educational, business, industrial and private organizations whose purposes parallel those of the Association.

C. The Chairman of the Steering Committee shall be appointed by the President of the Association. The term of office shall be one year and shall begin the last day of the annual conference.

1076

D. The Steering Committee shall have general supervision over the affairs of the Association and shall have the responsiblity for all activities of the Association. The Steering Commiteee shall conduct the business of the Association in the interim between annual conferences of the Association by such means of communication as deemed appropriate by the President or Chairman.

E. Meetings of the Steering Commitee shall be held during the annual conferences of the Association and at such times as requested by the President of the Association or the Chairman of the Steering Committee. Representation from the majority of the organizations of the Steering Committee shall constitute a quorum.

Article VI - Officers

A. The Officers of the Association shall consist of a President, a Chairman of the Steering Committee, and a Secretary.

B. The President of the Association shall be the Commanding Officer of the armed services agency co-ordinating the annual conference of the Association. The term of the President shall begin at the close of the annual conference of the Association and shall expire at the close of the next annual conference.

C. It shall be the duty of the President to organize and coordinate the annual conference of the Association held during his term of office, and to perform the customary duties of a president.

D. The Secretary of the Association shall be filled through appointment by the President of the Association. The term of office of the Secretary shall be the same as that of the President.

E. It shall be the duty of the Secretary of the Association to keep the records of the Association, and the Steering Committee, and to conduct official correspondence of the Association, and to issue notices for conferences. The Secretary shall solicit nominations for the Harry Greer award prior to the annual conference. The Secretary shall also perform such additional duties and take such additional responsibilities as the President may delegate to him.

Article VII - Meetings

A. The Association shall hold a conference annually.

B. The annual conference of the Association shall be co-ordinated by the agencies of the associated armed services exercising primary responsibility for military personnel assessment. The co-ordinating agencies and the order of rotation will be determined annually by the Steering Committee. The coordinating agencies for at least the following three years will be announced at the annual meeting.

C. The annual conference of the Association shall be held at a time and place determined by the co-ordinating agency. The membership of the Association shall be informed at the annual conference of the place at which the following annual conference will be held. The co-ordinating agency shall inform the Steering Committee of the time of the annual conference not less than six (6) months prior to the conference.

D. The co-ordinating agency shall exercise planning and supervision over the program of the annual conference. Final selection of program content shall be the responsibility of the co-ordinating organization.

E. Any other organization desiring to co-ordinate the conference may submit a formal request to the Chairman of the Steering Committee, no later than 18 months prior to the date they wish to serve as host.

Article VIII - Committees

A. Standing committees may be named from time to time, as required, by vote of the Steering Committee. The chairman of each standing committee shall be appointed by the Chairman of the Steering Committee. Members of standing committees shall be appointed by the Chairman of the Steering Committee in consultation with the chairman of the committee in question. Chairmen and committee members shall serve in their appointed capacities at the dsicretion of the Chairman of the Steering Committee. The Chairman of the Steering Committee shall be ex officio member of all standing committees.

B. The President, with the counsel and approval of the Steering Committee, may appoint such ad hoc committees as are needed from time to time. An ad hoc committee shall serve until its assigned task is completed or for the length of time specified by the President in consultation with the Steering Committee.

C. All standing committees shall clear their general plans of action and new policies through the Steering Committee, and no committee or committee, chairman shall enter into relationships or activities with persons or groups outside of the Association that extend beyond the approved general plan of work without the specific authorization of the Steering Committee.

1071

D. In the interest of continuity, if any officer or member has any duty elected or appointed placed on him, and is unable to perform the designated duty, he should decline and notify at once the officers of the Association that he cannot accept or continue said duty.

Article IX - Amendments

A. Amendments of these By-laws may be made at any annual conference of the Association.

B. Amendments of the By-laws may be made by majority vote of the assembled membership of the Association provided that the proposed amendments shall have been approved by a majority vote of the Steering Committee.

C. Proposed amendments not approved by a majority vote of the Steering Committee shall require a two-thirds vote of the assembled membership of the association.

Article X - Voting

All members in attendance shall be voting members.

Article XI - Enactment

These By-laws shall be in force immediately upon acceptance by a majority of the assembled membership of the Association and/or amended (in force 2 November 1973).

HARRY H. GREER AWARD

The Military Testing Association is an outgrowth of an informal meeting of representatives of the various armed forces testing agencies in 1958. The meeting was held at the suggestion (and through the personal co-ordination) of Capt Harry H. Greer, USN, Commanding Officer of the Naval Examining Center. Thus, Capt Greer was the "founder" of the Military Testing Association. In 1962, an award in his name was created to recognize significant lasting contributions to the Association while exemplifying the ideals of the Association and its founder.

The six recipients of the award since 1962 are:

1962	CAPT Harry H. GREER, USN
1970	COL J.M. McLANATHAN, USAF
1974	MR. C.J. MacALUSO, Naval Examining Center
1977	DR. W.J. MOONAN, Naval Personnel Research and Development Center
1977	MR. J.A. BURT, U.S. Coast Guard Institute
1979	DR. Raymond E. CHRISTAL, Air Force Human Resources Laboratory

1040

NAMES AND ADDRESSES OF AUTHORS AND CONFEREES

MR. JACK L. AINSWORTH
US ARMY ENGINEERING SCHOOL
ALEXANDRIA, VA
USA 22309

LT. MOHAMMED A. AKOODIE
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE ST, SUITE 600
WILLOWDALE, ONT., CANADA
M2N 6B7

1ST LT JOHN R. ALLEN
MARINE CORPS INSTITUTE
MARINE BARRACKS
WASHINGTON D.C., USA
20013

CMDR NICHOLAS H. ALLEN
US COAST GUARD HDQTRS
G-PO/OPES-TP-42
2100 2ND STREET S.W.
WASHINGTON D.C., USA
20593

MR. CHARLES ANDERSON
EDUCATIONAL TESTING SERVICE
1 AMERICAN PLAZA
EVANSTON, IL, USA
60201

A.P. ANDERSON
US ARMY INFANTRY SCHOOL
ATSH-EV
FT. BENNING, GA, USA
31905

MR. THOMAS M. ANSBRO
TAEG (ADDU) CNET N-5
BLDG. 679NAS
PENSACOLA, FL, USA
32508

MAJ BRIAN J. ARMOUR
AUSTRALIAN ARMY SENIOR
STANDARDIZATION REPRESENTATIVE
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, ONT., CANADA
K1A 0K2

MR. PAUL ASA-DORIAN
2375 GERANIUM ST.
SAN DIEGO, CA, USA
92109

DR. FRANK M. AVERSANO
US ARMY TRAINING SUPPORT CENTER
ATIC-SMD-TA
FT. EUSTIS, VA, USA 23604

MS. ANNETTE G. BAISDEN
PSYCHOLOGY DEPT.
NAVAL AEROSPACE MED RESCH LAB
NAVAL AIR STATION
PENSACOLA, FL, USA 32508

CRISTINA G. BANKS
COLLEGE OF BUSINESS ADMIN.
DEPT. OF MANAGEMENT
BEB 500, UNIV OF TEXAS
AUSTIN, TX, USA, 78712

MR. HARRY A. BARAN
AIR FORCE HUMAN
RESOURCES LABORATORY
LOGISTICS RESEARCH DIVISION
AFHRL/LRA
WRIGHT-PATTERSON AFB, OH, USA 45433

CAPT B.E. BELEC
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE ST. # 600
WILLOWDALE, ONT., CANADA M2N 6B7

MR. B. MICHAEL BERGER
CHIEF OCCUPATIONAL SURVEY BRANCH
DA MILPERCEN DAPC-MSP-S
200 STOVALL STREET
ALEXANDRIA, VA, USA 22332

PAUL BEST, PHD
MCFANN, GRAY AND ASSOCIATES, INC.
P.O. BOX 5705
PRESIDIO OF MONTEREY
MONTEREY, CA, USA 93940

CAPT M.K. BITTEN
CFTDC
CFB BORDEN
BORDEN, ONT., CANADA LOM 1C0

JAMES D. BLACKSHER
DEPARTMENT OF THE NAVY
DAM NECK
VIRGINIA BEACH, VA, USA 23461

(1081)

GEORGE W. BOAK
DEFENSE INTELLIGENCE AGENCY
UP-A2
PENTAGON
WASHINGTON, D.C., USA
20301

COL N.K. BODNAR, DIRECTOR
OFFICE OF MANPOWER UTILIZATION
HEADQUARTERS US MARINE CORPS
(CODE MPU)
QUANTICO VA, USA
22134

ARNOLD BOHRER
BELGIAN ARMED FORCES
PSYCHOLOGICAL RESEARCH SECTION
REKRUTERINGS-EN SELECTIECENRUM
SECTIE PSYCHOLOGISCH ONDERZOEK
KAZERNE KLEIN KASTEELTJE
1000 BRUSSEL BELGIUM

DR. JOHN A. BOLDOVICI
HUMAN RESOURCES
RESEARCH ORGANIZATION
BOX 283,
FORT KNOX, KY, USA
40272

DR. JAMES O. BOONE
FEDERAL AVIATION
ADMINISTRATION ACADEMY
OKLAHOMA CITY
OKLAHOMA, USA

CAPT D. BOUDREAU
CFTDC
CFB BORDEN
BORDEN, ONTARIO
L0M 1C0

LT D. BOYCE
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE ST.
SUITE 600
WILLOWDALE, ONTARIO, CANADA
M2N 6B7

MAJ J.E. BOZYNSKI
RESERVE PSO
N. ALTA. MILITIA DIST. H.Q.
GRIESBACH BARRACKS
CFB EDMONTON
LANCASTER PARK, ALTA, CANADA
T0A 2H0

LT. D. BREWER
G-T-1/TP42
WASHINGTON, D.C., USA
20593

MR. L. BRIERLEY
MINISTRY OF COLLEGES & UNIVERSITIES
16TH FLOOR MOWAT BLOCK
900 BAY ST.
TORONTO, ONT., CANADA
M7A 2B5

ROBERT D. BRODY
ACADEMIC INSTRUCTION
FOREIGN OFFICER SCHOOL
MAXWELL AFB ALABAMA, USA
36112

MARCEL BUJARSKI
77 ATLAS RD.
BASKING RIDGE, NJ, USA

DR. WILLIAM P. BURKE
ARMY RESEARCH INSTITUTE FIELD UNIT
PO BOX 2086
FT. BENNING GA, USA
31905

DOUGLAS J. BURROWS,
EDUCATION DEVELOPMENT DIVISION DPCA
UNITED STATES ARMY INFANTRY CENTER
BLDG 35
FT. BENNING GA, USA 31905

J.A. BURT
USCG INSTITUTE
PO SUBSTATION 18
OKLAHOMA CITY, OK, USA
73169

MR. R. BURTON
US COAST GUARD
OKLAHOMA CITY, OK, USA
73169

1012

MR. FRANK D. CAMPBELL
DIRECTORATE OF TRAINING DEV
ATTN: TRAINING ANALYSIS BRANCH
FT RUCKER, AL, USA
36362

MR. RICHARD J. CARTER
USARI-FT. BLISS
PO BOX 6057
FT. BLISS, TX, USA
79916

MICHAEL J. CASSIDY, SQNLDR RAAF
AUSTRALIAN EXCHANGE OFFICER
AFHRL/MODS
BROOKS AFB TX, USA
78235

CDR. A. CATTALINI
TRAINING OFFICER
U.S COAST GUARD TRAINING CENTER
PETALUMA, CA. 94952

DR. LOUIS F. CICCHINELLI
DENVER RESEARCH INSTITUTE/SSRE
UNIVERSITY OF DENVER
DENVER, CO, USA
80208

JOAN T. CHIPPENDALE
T.D.I. - OFFICER TRG SYS DIV
BLDG. 10 - ROOM 310
FORT MONROE, VA, USA
23651

DONALD C. COLEMAN
CODE 70A00
NAS WHITING FIELD
MILTON, FL, USA
32570

MR. F. L. COMSTOCK, JR.
LINK DIV, SINGER CO.
BOX 426-1 GORMAN RD., KIRKWOOD
N.Y., USA
13795

MAJ J.Y.L. COTE
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, ONTARIO, CANADA
K1A OK2
ATTN: CPD

MR. THOMAS CURRAN
LAWTON, OK, USA
73501

COMMANDER, US ATSC
ATTN: ATIC-SMD-SE
(MR. BRIAN.C. DAVIS)
FT. EUSTIS VA, USA
23604

MR. D.D. DAVIS
CHIEF OF NAVAL EDUCATION
AND TRAINING
CODE N-511, NAVAL AIR STATION
PENSACOLA, FL, USA 32508

JULIA R. DELONEY
US COAST GUARD INSTITUTE
PO SUBSTATION 18
OKLAHOMA CITY, OK, USA 73169

MR. JACK DEMPSEY
ENVIRONMENTAL SYSTEMS INC.
12402 ALEXANDRIA
SAN ANTONIO, TX, USA

RICHARD W. DICKINSON
OCCUPATIONAL RESEARCH PROGRAM
INDUSTRIAL ENGINEERING DEPT.
TEXAS A&M UNIVERSITY
COLLEGE STATION, TX, USA 77843

BERNARD T. DODD
SENIOR PSYCHOLOGIST (NAVAL), MOD
ARCHWAY BLOCK SOUTH IV
OLD ADMIRALTY BLDG.
WHITEHALL, LONDON, UK SW1A2BE

MR. J.A. DORAN
RESEARCH INFORMATION MANAGER
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE ST.
SUITE 600
WILLOWDALE, ONT., CANADA
M2N 6B7

DR. WALTER E. DRISKILL
USAF OCCUPATIONAL ANALYSIS PROGRAM
USAFOMC/OMY
RANDOLPH AFB TX, USA 78148

EUGENE H. DRUCKER
HUMAN RESOURCES RESEARCH
ORGANIZATION
FORT KNOX OFFICE
P.O. BOX 293
FORT KNOX, KY, USA 40121

MR. DAN DULL
US NAVY NTTC CORRY STATION
8324 LYRIC DR.,
PENSACOLA, FL, USA 32504

LT. ERIC DUNCAN
USAFOMC/OMDRA
RANDOLPH AFB, TX, USA 78148

HEINZ-JURGEN EBENRETT
GERMAN ARMED FORCES PSYCHOLOGICAL
SERVICES RESEARCH INSTITUTE
STREITKRAFTEAMT
ABT. I, DEZ. WEHRPSYCHOLOGIE
POST BOX 20 50 03
D-5300
BONN 2, FEDERAL REPUBLIC OF GERMANY

DR. EDWARD E. EDDOWES
AFHRL/OT
WILLIAMS AFB, AZ, USA 85224

LT GREGORY J. EDGE
US COAST GUARD INSTITUTE
PO SUBSTATION 18
OKLAHOMA CITY, OK, USA 73169

LCOL J.C. EGGENBERGER
DIRECTOR PERSONNEL SELECTION
RESEARCH AND SECOND CAREERS
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, ONTARIO, CANADA K1A OK2

MAJ R.T. ELLIS
ATTN: DPSRSC-3
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, ONTARIO, CANADA K1A OK2

R. ESPERSON
NAVEDTRAPRODEVCEN
CODE PD-7
PENSACOLA, FL, USA 42509

CAPT IAN E. FALLE
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, ONT., CANADA K1A 0K2
ATTN: DMOS

JONATHAN C. FAST
AIR FORCE HUMAN RESOURCES
LABORATORY (MPMD)
BROOKS AIR FORCE BASE
TEXAS, USA
78235

LT HONEY R. FISHER
CANADIAN FORCES PERS APPLIED RSCH UNIT
4900 YONGE STREET # 600
WILLOWDALE, ONT., CANADA M2N 6B7

ELI FLYER
10 SIERRA VISTA DRIVE
MONTEREY, CA, USA 93940

DR. FLYNN
US ARMY INTELLEGENCE SCHOOL
FT. DEVON, MASS, USA

MAJ. J.R. FOWLIE
NATIONAL DEFENCE HQ/CPD
OTTAWA, ONT., CANADA K1A OK2

LT. STEPHEN H. FRANCIS
TRAINING OFFICER
U.S. COAST GUARD TRAINING CENTER
PETULUMA, CA, USA 94952

TERRY FRANUS
NAVAL HEALTH SCIENCES TRAINING COMMAND
BLDG 141, CODE 212 NNMC
BETHESDA, MARYLAND, USA
20014

DR. E. WAYNE FREDERICKSON
APPLIED SCIENCES ASSOCIATES LTD.
P.O. BOX 6057
FT. BLISS TX, USA 79916

LT(N) D. FREEMAN
CANADIAN FORCES TRAINING
SYSTEMS HEADQUARTERS
CFB TRENTON
ASTRA, ONT., CANADA
KOK 1B0

ROBERT L. FREY, JR.
HEADQUARTERS, U.S. COAST GUARD
G-P-1/2/TP42
WASHINGTON, D.C., USA 20593

DIRECTOR, US ARMY TRADOC
SYSTEMS ANALYSIS ACTIVITY
ATTN: ATAA-TH (DR.EDWARD L. GEORGE)
WHITE SANDS MISSILE RANGE, NM, USA
88002

DR. ARTHUR C.F. GILBERT
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA, VA, USA 22333

J.P. GODBOUT
CFTSHQ
CFB ST JEAN
RICHELAIN, P.Q., CANADA

DR. LAWRENCE A. GOLDMAN
DAPC-MSP-D
200 STOVALL STREET
ALEXANDRIA, VA, USA
22332

DR. DOUG GOODGAME
OCCUPATIONAL RESEARCH PROGRAM
INDUSTRIAL ENGINEERING DEPT
TEXAS A&M UNIVERSITY
COLLEGE STATION, TX, USA
77801

DR. SHERRIE P. GOTT
AFHRL/MODS
BROOKS AFB, TX, USA 78235

DR. ALEXANDER M. GOTTESMAN
HEAD, CURRICULUM BRANCH
HSETC, CODE 211
NATIONAL NAVAL MEDICAL CENTER
BETHESDA, MD, USA 20014

LTC DAVID L. GROETKEN
CHIEF, ANALYSIS DIV
DIRECTORATE OF EVAL.
US ARMY FIELD ARTILLERY
FT. SILL, OK, USA 73503

MR. JOSEPH GUERREIN
SFTD, DOT
USAIS
FT. BENNING, GA, USA
31905

MR. J. HALE
INSTITUTE OF NUCLEAR
POWER OPERATIONS
ATLANTA, GEORGIA, USA

DR. E.J. HALTRECHT
ONTARIO HYDRO H2-A13
700 UNIVERSITY AVE
TORONTO, ONT., CANADA
M5G 1X6

DIRECTOR OF RECRUITING AND SELECTION
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, ONT., CANADA K1A 0K2
ATTN: MAJ M.P. HANKES-DRIELSMA

KEN HANSEN
PSYCH SYSTEMS, INC.
600 REISTERSTOWN RD, SUITE 300A
BALTIMORE, MD, USA 21208

LCOL D.A. HARRIS
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, ONT., CANADA
K1A 0K2
ATTN: CPD

MR. C. HARVEY
US ARMY INFANTRY SCHOOL
ATTN: ATSH-EV
FT. BENNING, GA, USA
31905

JOHN E. HASSEN
102 SEVERIN DRIVE
PENSACOLA, FL, USA 32503

CDR F.J. HAWRYSH
DMOS 3, NATIONAL DEFENCE HEADQUARTERS
101 COL BY DRIVE
OTTAWA, ONTARIO, CANADA
K1A 0K2

1086

CAPT CLIFF L. HEARNDEN
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, ONTARIO, CANADA
K1A 0K2
ATTN: DMOS

MR. O.E. HENLEY
MISSIL
ATSK-TD-AD-M
USAMMCS
REDSTONE ARSENAL, AL, USA
35809

MR. R. HESLEGRAVE
DEFENCE AND CIVIL INSTITUTE
OF ENVIRONMENTAL MEDICINE
BOX 2000
DOWNSVIEW, ONTARIO, CANADA

LCOL T.W.H. HEWER
ONTARIO MINISTRY OF
COLLEGES & UNIVERSITIES
16TH FLOOR MOWAT BLOCK
900 BAY ST., QUEENS PARK
TORONTO, ONT., CANADA
M7A 2B5

RICHARD H. HISS
ESSEX CORPORATION
P.O. BOX 147
WHITE SANDS MISSILE RANGE, NM
USA 88002

EDWARD N. HOBSON
11713 LARIAT LANE
OAKTON, VIRGINIA, USA
22124

CDR J.D. HOLLAND
OIC NODAC BLDG 150
WASHINGTON NAVY YARD ANACOSTIA
WASHINGTON, D.C., USA
20374

PAUL L. HOLLANDER
15 LOTUS COURT
WILLOWDALE, ONT., CANADA
M2H 1J6

CAPT D.S. HORTON
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE STREET
SUITE 600
WILLOWDALE, ONT., CANADA
M2N 6B7

LTC C. HOUSTON
DEPARTMENT OF THE ARMY
5001 EISENHOWER AVE
ALEXANDRIA, VA, USA
22333

MR. JAMES B. HOWARD
US ARMY AVIATION SCHOOL
DIRECTORITE OF EVAL/STANDARDIZATION
FORT RUCKER, AL, USA
36362

JAMES C. HUMPHLETT
FEDERAL LAW ENFORCEMENT
TRAINING CENTRE
GLYNCO, GA, USA 31524

MR. JOSEPH W. ILLES
ARMY EDUCATION CENTER
LEDWARD BKS
APO, NY, USA
09033

LCDR I.L. JACKSON
NATIONAL DEFENCE HEADQUARTERS
ATTN: DMOS3-2
OTTAWA, ONTARIO, CANADA
K1A 0K2

WILLIAM L. JACKSON
US ARMY AVIATION CENTER
DEPUTY FOR TRAINING DEVELOPMENT
FT. RUCKER AL, USA 36362

CAPT J.A. JAMES
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE ST. # 600
WILLOWDALE, ONT., CANADA M2N 6B7

DEPARTMENT OF DEFENSE
9800 SAVAGE ROAD
FT. GEORGE G. MEADE, MD, USA
20755
ATTN: E7, R. JENKINS

(1086)

MR. JOHN JOAQUIN
HUMAN RESOURCE PLANNING
AND ASSESSMENT DEPT.
LOCATION H2B13, ONTARIO HYDRO
700 UNIVERSITY AVE.
TORONTO, ONTARIO, CANADA M5G 1X6

CAROL A. JOHNSON, PHD
MCFANN, GRAY AND ASSOCIATES, INC.
P.O. BOX 5705
PRESIDIO OF MONTEREY
MONTEREY, CA, USA 93940

JAMES H. JOHNSON
PSYCH SYSTEMS, INC.
600 REISTERSTOWN RD., SUITE 300A
BALTIMORE, MD, USA 21208

ROBERT N. JOHNSON
DIRECTORATE OF TNG DEV
US ARMY ADMINCEN
FT BENJAMIN HARRISON, IN, USA 46216

D. TODD JONES, P.E.
OFFICER OF RESEARCH AND DEVELOPMENT
G - DMT-1/54
US COAST GUARD
2100 V St. S.W.
WASHINGTON, D.C., USA 20593

KAREN N. JONES
U.S. COAST GUARD INSTITUTE
PO SUBSTATION 18
OKLAHOMA CITY, OK, USA 73169

TAFT M. JOSEPH, JR
703 SW CHAUCER CIRCLE
LAWTON, OK, USA 73505

DR. ARTHUR KAHN
2 COVE CORK LANE
ANNAPOLIS, MD, USA 21401

JEFFREY E. KANTOR
AIR FORCE HUMAN RESOURCES
LABORATORY (MODF)
BROOKS AIR FORCE BASE
TEXAS, USA 78235

CAPT WAYNE KEATES
STAFF OFFICER ANALYSIS
AIR COMMAND HEADQUARTERS
WESTWIN, MANITOBA, CANADA R2R OTO

MR. JAMES B. KEETH
USAFOMC/OMY
RANDOLPH AFB, TX, USA 78148

CAPT RALPH KELLETT
NATIONAL DEFENCE HEADQUARTERS
L'ESPLANADE LAURIER EAST TOWER
140 O'CONNOR ST., STATION 7-4
OTTAWA, ONT., CANADA
K1A OK2

WILLIAM J.C. KELLOWAY
ACADEMIC ADVISOR
CANADIAN POLICE COLLEGE
P.O. BOX 8900
OTTAWA, ONT., CANADA K1G 3J2

LCDR HARRY KELLNER
NODAC
BLDG 150 WASHINGTON NAVY YARD (ANA)
WASHINGTON D.C., USA 20374

LCDR ROBERT H. KERR
DIVISION CDR,
TRAINING DEVELOPMENT DIVISION
CANADIAN FORCES FLEET SCHOOL HALIFAX
FMO, HALIFAX, NOVA SCOTIA, CANADA

MARYANN J. KICINSKI
LANGLEY AFB, VA, USA
23665

DR. FRANK KINRADE
CANADIAN EMPLOYMENT AND
IMMIGRATION COMMISSION
OTTAWA, ONTARIO, CANADA

MR. DENNIS KIRBY - ATD
TRANSPORT CANADA
TOWER B - 4TH FLOOR
PLACE DE VILLE
OTTAWA, ONTARIO, CANADA K1A ON5

MAJ EMIL K. KLUEVER
US ARMY OFFICE OF ARMOR
FORCE MANAGEMENT
FORT KNOX KY, USA
40121

DR. C. MAZIE KNERR
P.O. BOX 1286
SPRINGFIELD
VIRGINIA, USA
22151

DEFENSE LANGUAGE INSTITUTE
ATTN: ATFL-TD-T(MAJ KNOX)
PRESIDIO OF MONTEREY
MONTEREY, CALIFORNIA, USA 93940

CHRISTOPHER G. KOCH
HONEYWELL INC.
2345 WOODBRIDGE #122
ROSEVILLE, MN, USA
55113

MR. HANS KOTIESEN
ONTARIO MIN OF COLLEGES
& UNIV, QUEENS PARK
MOWAT BLOCK 16TH FLOOR
900 BAY ST
TORONTO, ONT., CANADA
M7A 2B5

S. KUMAGAI
CONTROL DATA CORPORATION
8100 34TH AVE S.
BLOOMINGTON
MINNEAPOLIS, MN, USA
55420

2LT GILLES LABARRE
1870 BOUL ST. JOSEPH
APT 3
MONTREAL, P.Q., CANADA

MAJ JR LALONDE
SENIOR PSO SE (M)
CP 6666
MONTREAL, QUE, CANADA
H3C 3L9

LT. C.D. LAMERSON
21 RADAR SQUADRON
ST. MARGARETS
JAMES PARK
NEW BRUNSWICK, CANADA
E0C 1J0

MR. T. LANDVOGT
US COAST GUARD
OKLAHOMA CITY, OK, USA
73169

RICHARD LANTERMAN (G-T-1/42)
US COAST GUARD
400 SEVENTH STREET S.W.
WASHINGTON, D.C., USA 20590

GEORGE W. LAWTON
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA, VA, USA
22333

CAPT J.M. LEBLANC
BASE PERSONNEL SELECTION OFFICER
CFB LAHR
CFPO 5000
BELLEVILLE, ONTARIO, CANADA
K0K 3R0

LCOL D.A. LEFROY
MILITARY LEADERSHIP AND MANAGEMENT DEPT.
ROYAL MILITARY COLLEGE
KINGSTON, ONTARIO, CANADA
K7L 2W3

CAPT P. LESSARD
CANADIAN FORCES PERS APPLIED RSCH UNIT
4900 YONGE ST # 600
WILLOWDALE, ONTARIO, CANADA M2N 6B7

RICHARD A. LILIENTHAL
US OFFICE OF PERSONNEL MANAGEMENT
5324 TANEY AVE.
ALEXANDRIA, VA, USA 22304

M. SUZANNE LIPSCOMB
AFHRL/MODF
BROOKS AFB TX, USA
78235

E.J. LLOYD
US MARINE CORPS INSTITUTE
1209 SOUTH FREDERICK
ARLINGTON, VA, USA
22204

1088

DR. R. LOO
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE STREET
SUITE 600
WILLOWDALE, ONT., CANADA
M2N 6B7

MR. LARRY T. LOOPER
AIR FORCE HUMAN RESOURCES
LABORATORY (MOMD)
BROOKS AFB TX, USA
78235

MAJ J.R. MACDOUGALL
STAFF OFFICER PERSONNEL SELECTION
CANADIAN FORCES TRAINING
SYSTEM HEADQUARTERS
CANADIAN FORCES BASE TRENTON
ASTRA, ONT., CANADA
K0K 1B0

EDWARD F. MAGDARZ
ADVISORY EDUCATION SPECIALIST
IBM CORP. DEPT 817/634-3
BOX 12195
RESEARCH TRIANGLE PARK, NC, USA
27709

MR. L. MAGEE
DEFENCE AND CIVIL INSTITUTE
OF ENVIRONMENTAL MEDICINE
BOX 2000
DOWNSVIEW, ONT., CANADA

ARTHUR MARCUS
8020 LAKENHEATH WAY
POTOMAC, MARYLAND, USA
20854

COMMANDANT, U.S. COAST GUARD
OFFICE OF BOATING SAFETY
AUXILIARY AND EDUCATION DIVISION
WASHINGTON, D.C., USA 20593
ATTN: JERROLD MARKOWITZ

JOANNE MARSHALL-MIES
INSTITUTE FOR BEHAVIOURAL RESEARCH
2429 LINDEN LANE
SILVER SPRING, MD, USA 20910

CAPT J. MCCUTCHEON
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, ONT., CANADA
K1A 0K2
ATTN: DPCAOR 2-4

DR. B.A. MACFARLANE
PROF OF SOC AND INT AFF
DEPT. OF SOCIOLOGY & ANTHROPOLOGY
CARLETON UNIVERSITY
OTTAWA, ONTARIO, CANADA
K1S 5B6

LTCOL W.W. MCIVER
ASST DIRECTOR
OFFICE OF MANPOWER UTILIZATION
HEADQUARTERS, US MARINE CORPS
QUANTICO, VA, USA 22134

ROBERT C. MCKENZIE
U.S. OFFICE OF PERSONNEL MANAGEMENT
4511 GAGE ROAD
ALEXANDRIA, VA, USA
22309

CAPT J.P. MCMENEMY
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE STREET,
SUITE 600
WILLOWDALE, ONTARIO, CANADA M2N 6B7

MR. JOHN B. MEREDITH
SENIOR EVALUATOR
DATA DESIGN LABORATORIES
PO 12773, 15 KOGER
NORFOLK, VA, USA 23501

NATHAN MEWHINNEY
DEPARTMENT OF THE NAVY
DAM NECK
VIRGINIA BEACH, VA, USA 23461

DR. ANGELO MIRABELLA
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA, VA, USA
22333

LTCOL JIMMY L. MITCHELL
USAF OCCUPATIONAL MEASUREMENT
CENTER (ATC)/OMY
RANDOLPH AIR FORCE BASE
TEXAS, USA 78148

AMELIA E. MOBLEY
US COAST GUARD
400 7TH STREET
WASHINGTON, D.C., USA
20590

10M

DR. JOHN B. MOCHARNUK
HARRIS CORPORATION
GOVT. ELECTRONIC SYS DIV.
BLDG 22, RM 2414
PO BOX 37
MELBOURNE, FL, USA
32901

MAJOR MORKEM
CANADIAN FORCES TRAINING
SYSTEM HEADQUARTERS
CFB TRENTON
ASTRA, ONT., CANADA
K0K 1B0

STEPHAN J. MOTOWIDLO
STATE UNIVERSITY OF NEW YORK
BINGHAMTON, NY, USA

D.J. MUELLER
PD-10, NETPDC
NAVEDTRAPRODEVCEN
PENSACOLA, FL, USA
42509

JOHN W. MURPHY
TEST DESIGN COORDINATOR
USAIA ROOM 103, BLDG 400
FT. BENJAMIN HARRISON, IND, USA
46216

MS. JEAN NEWTON
OFFICE OF PERSONNEL MANAGEMENT
1900 E. ST NW (CODE 3609)
UNITED STATES OF AMERICA
WASHINGTON, D.C., USA
20415

MR. BILL NORDBROCK
US NAVY EDUCATION AND
TRAINING PROGRAM DEVELOPMENT
CENTRE DETACHMENT
BLDG 90 - GREAT LAKES
ILLINOIS, USA
60096

MR. IAN NOY
DEFENCE AND CIVIL INSTITUTE
OF ENVIRONMENTAL MEDICINE
BOX 2000
DOWNSVIEW, ONT., CANADA

MR. ROBERT F. OBRIEN
DEFENSE LANGUAGE INSTITUTE
PENSACOLA, FL, USA

LTJG F.X. O'BYRNE, JR
US COAST GUARD TRACEN
GOVERNORS ISLAND, N.Y., USA
10004

DR. BRIAN S. O'LEARY
PERSONNEL RESEARCH AND
DEVELOPMENT CENTER
US OFFICE OF PERSONNEL MANAGEMENT
1900 E ST. N.W.
WASHINGTON, D.C., USA 20415

LCDR DAN M. OGLE
NATIONAL DEFENCE HEADQUARTERS
EAST TOWER L'ESPLANADE LAURIER
140 O'CONNOR STREET (DPED)
OTTAWA, ONTARIO, CANADA K1A 0K2

CHARLES OLSON
2303 COLERIDGE DR.
SILVER SPRING, MD, USA 20910

RICHARD J. OREND
HUMAN RESOURCES RESEARCH ORGANIZATION
HQ, USAREUR
ODCSPER (ARI) BOX 127
APO NEW YORK, USA 09403

WILLIAM C. OSBORN
HUMRRO-FT. KNOX OFFICE
PO BOX 293
FT. KNOX, KY, USA 40121

T. DIANE OWENS
NAVAL EDUCATION & TRAINING PROG
NAVEDTRAPRODEVCEN
PENSACOLA, FL, USA 32504

DAVID J. OWEN
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, ONTARIO, CANADA
K1A 0K2
ATTN: DMOS

DR. ROBERT P. PALESE,
PERSONNEL PSYCHOLOGIST
U.S.C.G. TRAINING CENTRE
GOVERNORS ISLAND N.Y., USA
10004

CAPT R.E. PARK
ATTN: BPSO
CANADIAN FORCES BASE BORDEN
BORDEN, ONTARIO, CANADA L0M 1C0

DR. JOHN J. PASS
NAVAL PERSONNEL RESEARCH
AND DEVELOPMENT CENTER
CODE 310
SAN DIEGO, CA, USA
92152

NANCY A. PERRIGO
AFHRL/TSR
BROOKS AFB TX, USA
78235

MR RICHARD PIGEON
NDHQ L'ESP EAST TOWER
140 O'CONNOR STN 7-4
OTTAWA, ONT., CANADA
K1A 0K2

LTC MARK PILGRIM
HQ TRADOC
ATTN: ATTG-DOR
FT. MONROE VA, USA
23651

MAJ F.C. PINCH
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE STREET
SUITE 600
WILLOWDALE, ONT., CANADA
M2N 6B7

STEVEN PINE
SYSTEMS RESEARCH CENTER
HONEYWELL, INC.
2600 RIDGWAY PARKWAY
MOLS, MN, USA
55413

C.M. PIPKIN
NAVEDTRAPRODEVCEN
CODE PD-7
PENSACOLA, FL, USA
32509

WILLIAM POPYUK
19 BLANCHET ST.
ST LUC, QUEBEC, CANADA
J0J 2A0

LCDR EARL H. POTTER III
DEPARTMENT OF HUMANITIES
US COAST GUARD ACADEMY
NEW LONDON CT, USA
06320

J. STEPHEN PRESTWOOD
ASSESSMENT SYSTEMS CORPORATION
2395 UNIVERSITY AVE., SUITE 306
ST. PAUL, MINNESOTA, USA 55114

BARBARA J. PRICE
1100 JACKSON STREET
TOLEDO, OH, USA

MAJ TERRY J. PROCIUK
MLM DEPT
ROYAL MILITARY COLLEGE
KINGSTON, ONTARIO, CANADA
K7L 2W3

SQN LDR BRIAN N PURRY, RAF, (RETD)
WINDYRIDGE 51 THRAPSTON RD
BRAMPTON HUNTINGDON CAMBS.
UNITED KINGDOM

KLAUS J. PUZICHA, PHD
HEAD, GERMAN ARMED FORCES PSYCHOLOGICAL
SERVICES RESEARCH INSTITUTE
STREITKRAFTEAMT
ABT. I DEZ WEHRPSYCHOLOGIE
POST BOX 20 50 03
D-5300
BONN 2, FEDERAL REPUBLIC OF GERMANY

LCOL GLENN M. RAMPTON
COMMANDING OFFICER
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE ST., SUITE 600
WILLOWDALE, ONTARIO, CANADA M2N 6B7

CARMELA RAPILLO
NAVAL HEALTH SCIENCE TRAINING COMMAND
BLDG 141 CODE 212 NNMC
BETHESDA, MARYLAND, USA 20014

JOHN H. RATHKAMP
ARMY EDUCATION CENTER
BLDG 112
FT. MCPHERSON, GA, USA 30330

MARTIN L. RAUCH
CHIEF PSYCHOLOGIST-MOD-BONN
MINISTRY OF DEFENCE
POST BOX 1328
53 BONN 1,
FEDERAL REPUBLIC OF GERMANY

(1091)

LT FRANK RIPKIN
USA CAREER PROGRAMS SEC
ARLINGTON ANNEX
WASHINGTON, D.C., USA
200350

MR. W.M. RITCHIE
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, ONT., CANADA
K1A OK2
ATTN: DGPRD

MR. MICHAEL N. RODGERS
PERSONNEL PSYCHOLOGY CENTRE
L'ESPLANADE 300 LAURIER
WEST TOWER
OTTAWA, ONT., CANADA
K1A OM7

KENDALL ROOSE
NAVEDTRAPRODEVCEN
PENSACOLA, FLORIDA, USA
32509

MR. ROBERT ROSS
RESEARCH PSYCHOLOGIST
ARMY RESEARCH INSTITUTE
5009 ALTA VISTA CT.
BETHESDA, MD, USA

CAPT P. ROSSITER
NATIONAL DEFENCE HQ PROGRAMMER
ATTN: DMOS 3-4
OTTAWA, ONT., CANADA
K1A OK2

DR. HENDRICK W. RUCK
AFHRL/MODS
BROOKS AFB TX, USA
78235

M. RUMSEY
US ARMY RESEARCH INST
5001 EINSEHOWER AVENUE
ALEXANDRIA VA, USA
22333

MR. SYDNEY SAKO
CHIEF, MEASUREMENT BRANCH
OFFICER TRAINING SCHOOL (OTS/MTCM)
LACKLAND AFB TX, USA
78236

WILLIAM A. SANDS
NAVY PERSONNEL RESEARCH AND
DEVELOPMENT CENTER (CODE P310)
SAN DIEGO, CA, USA
92152

MILDRED E. SARGENT
1019 STILLBROOK ROAD
PENSACOLA, FL, USA
32504

CAPT DEBORAH A. SAUDINO
BASE PERSONNEL SELECTION OFFICER
FLEET SCHOOL
CFB HALIFAX, HALIFAX, NS, CANADA

DR. DOROTHY SCANLAND
DEFENCE ACTIVITY FOR NONTRADITIONAL
EDUCATIONAL SUPPORT
PENSACOLA, FL, USA
32508

DR. WORTH SCANLAND
CHIEF OF NAVAL EDUCATION
AND TRAINING (CODE N-5)
PENSOCOLA, FL, USA
32508

COL H.E. SEUBERLICH
CHAIRMAN ARMY SERV
(DEUTSCHER BUNDERSWEHR-VERBAND)
VORSITZENDER HEER
5300 BONN 2, SUBSTRASE 123
FEDERAL REPLUBIC OF GERMANY

MR. BRADFORD P. SHARP
G-PO/OPES-TP-42
US COAST GUARD HDQTRS
2100 2ND ST. S.W.
WASHINGTON, D.C., USA 20593

LCDR W.S. SHIELDS
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE ST. #600
WILLOWDALE, ONTARIO, CANADA
M2N 6B7

(1092)

DR. NORMAN M. SHUMATE
DEPUTY DIRECTOR OF
TRAINING DEVELOPMENTS
US ARMY ARMOR CENTER
FORT KNOX, KY, USA
40121

CAPT. SUZANNE P. SIMPSON
STAFF OFFICER, PERS DEV STUDIES
CHIEF PERSONNEL DEVELOPMENT
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, ONT., CANADA
K1A 0K2
ATTN: DPDS 4

MARY J. SKINNER
AIR FORCE HUMAN RESOURCES
LABORATORY (MODF)
MANPOWER AND PERSONNEL DIVISION
BROOKS AFB TX, USA
78235

LCOL D.J. SLIMMAN
DIRECTOR PERSONNEL
DEVELOPMENT STUDIES
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, ONT., CANADA
K1A 0K2

DR. BRANDON B. SMITH
MINNESOTA RESEARCH AND DEVELOPMENT
CENTER FOR VOCATIONAL EDUCATION
B-12 FRASER HALL
UNIV OF MINNESOTA
MINNEAPOLIS, MN, USA
55455

MR. DONALD G. SMITH
ONTARIO GOVERNMENT MCU
55 CHARLES ST. E.
TORONTO, ONT., CANADA

MABRON H. SMITH
US ARMY MISSILE/MUNITIONS CEN/SCH
ATTN: ATSK-EMU
REDSTONE ARSENAL, AL, USA
35809

DANIEL E. SPECTOR, PHD
US ARMY MILITARY POLICE SCHOOL
ATTN: ATZN-MP-TDE-S
FT. MCCLELLAN, AL, USA
36205

INSP. W.R. SPRING
OIC/POLICY, PLANNING & EVAL SEC.
STAFFING & PERSONNEL BRANCH
R.C.M. POLICE, 1200 ALTA VISTA DRIVE
OTTAWA, ONTARIO, CANADA
K1A 0R2

DR. JAMES A. SQUYRES
DEPT OF TREASURY
FEDERAL LAW ENFORCEMENT TRAINING CTR
GLYNCO, GA, USA
31524

CAPT E.L. STENTON
CFPARU
4900 YONGE ST., SUITE 600
WILLOWDALE, ONTARIO, CANADA
M2N 6B7

DALE W. STEWART
OFFICE OF ARMOR FORCE MANAGEMENT
HQ US ARMY ARMOR CENTER
FT. KNOX, KY, USA
40121

MR. J. STOKES
STATISTICIAN
US COAST GUARD HEADQUARTERS
G-T-1/TP42
WASHINGTON, D.C., USA
20593

CAPT F.G. STRICKLAND
ATTN: BPSO
CANADIAN FORCES BASE GREENWOOD
GREENWOOD, NOVA SCOTIA, CANADA B0P 1N0

DENIS J. SULLIVAN, JR.
HQ USAISD
ATTN: ATSIE-TD-AD-D
FORT DEVENS, MA, USA
01433

CAPT L.M. SURA
ATTN: DPSRSC-3-2
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, ONTARIO, CANADA
K1A 0K2

CAPT LISA A. TALMAGE
OFFICE OF MANPOWER UTILIZATION
BLDG 2009
QUANTICO, VA., USA
22193

MAJ J.F. TAYLOR
ARMY SCHOOL OF TRAINING SUPPORT
MILTON PARK
BEACONSFIELD
BUCKS, ENGLAND

SGT. MICHAEL C. THEW
AIR FORCE HUMAN RESOURCES LABORATORY
AFHRL/TSPZ
BROOKS AFB, TX, USA
78235

MR. HARVEY L. THORSTAD
CNET N5, US NAVY
NAS, PENSACOLA, FL, USA
32508

CAPT E.C. TIERNEY
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE STREET
SUITE 600
WILLOWDALE, ONT., CANADA
M2N 6B7

DR. MARVIN H. TRATTNER
U.S. OFFICE OF PERSONNEL MANAGEMENT
PRDC ROOM 3H17
WASHINGTON, D.C., USA
20415

JOHN D. TUBBS
USATRASANA ATAA-THC
WHITE SANDS MISSILE RANGE
NM, USA
88002

MARY TURK
US ARMY ORDANCE CTR & SCH
ABERDEEN PROVING GROUND
MD, USA
21005

MR. DUANE TYERMAN
CANADIAN FORCES TRAINING SYSTEMS
CFB TRENTON
TRENTON, ONT., CANADA

C. DAVID VALE
ASSESSMENT SYSTEMS CORP.
306 SECURITY BUILDING
2395 UNIVERSITY AVE.
ST. PAUL, MN, USA 55114

CAPT. GORDON VANDYKE
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE ST. #600
WILLOWDALE, ONT., CANADA M2N 6B7

PHYLLIS P. VOORHEES
US COAST GUARD INSTITUTE
PO SUBSTATION 18
OKLAHOMA CITY, OK, USA
73169

DR. R.O. WALDKOETTER
AIR FIELD UNIT
PO BOX 33066
FT. SILL, OK, USA 73503

MAJ R.W. WALKER
STAFF OFFICER PERSONNEL SELECTION
AIR COMMAND HEADQUARTERS
WESTWIN, MANITOBA, CANADA
R2F 0T0

MR. G. WILLIAM WALSTER
CONTROL DATA CORP
1956 PORTLAND AVE.
ST. PAUL, MN, USA

THOMAS A. WARM
US COAST GUARD INSTITUTE
PO SUBSTATION 18
OKLAHOMA CITY, OK, USA
73169

MR. A. TIMOTHY WARNOCK
AIR WAR COLLEGE ASSOCIATE PROGRAMS
AIR UNIVERSITY, BLDG 1401
MAXWELL AFB, AL, USA
36112

DR. BRIAN WATERS
HUMAN RESOURCES RESEARCH
300 N. WASHINGTON ST.
ALEXANDRIA, VA, USA 22314

(109K)

W. WATT
US ARMY INFANTRY SCHOOL
FORT BENNING, GA, USA
31905

JOSEPH L. WEEKS
AFHRL/MODS/WEEKS
BROOKS AFB, TX, USA
78235

JOHNNY J. WEISSMULLER
CENTER FRO CYBERNETIC STUDIES
UNIVERSITY OF TEXAS AT AUSTIN
AUSTIN, TX, USA
78712

MAJ JOHN R. WELSH, JR.
AIR FORCE MANPOWER AND
PERSONNEL CENTRE
RANDOLPH AFB, TX, USA
78148

RAYBURN A. WILLIAMS CODE N-53
CHIEF OF NAVAL EDUCATION AND TRAINING
PENSACOLA, FL, USA 32508

RICHARD C. WILLING
US COAST GUARD INST.
PO SUBSTATION 18 (MVP)
OKLAHOMA CITY, OK, USA
63169

CAPT F.P. WILSON
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE STREET, #600
WILLOWDALE, ONTARIO, CANADA
M2N 6B7

CAPT P.W. WILSON
TRAINING DEVELOPMENT OFFICER
CFSAOE
CFB BORDEN
BORDEN, ONTARIO, CANADA
LOM 1CO

DR. MARTIN F. WISKOFF
5151 BIXEL DRIVE
SAN DIEGO, CA, USA
92115

STAFF SGT YARD
C/O TRAINING & DEVELOPMENT BRANCH
RCM POLICE
1200 ALTA VISTA DRIVE
OTTAWA, ONTARIO, CANADA K1A 0R2

BRIAN J. YORE
HUMRRO
P.O. BOX 293
FT. KNOX, KY, USA
40121

(1095)


DAT
FILM
6