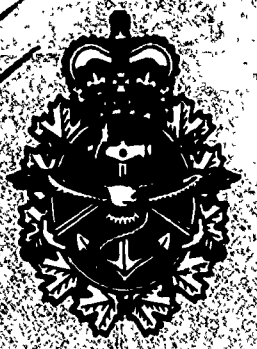
22nd ANNUAL CONFERENCE

PROCEEDINGS

MILITARY TESTING ASSOCIATION

LEVEL II



co-ordinated by

CANADIAN FORCES PERSONNEL APPLIED RESEARCH UNIT

Held in TORONTO, ONTARIO, CANADA

27-31 OCTOBER 1980

VOL. 1

50272-101

REPORT DOCUMENTATION PAGE	1. REPORT NO. MTA-22-80	2. AD-A098 677	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-22-80

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-Ended Terms

c. COSATI Field/Group

18. Availability Statement	19. Security Class (This Report)	21. No. of Pages
Release unlimited	UNCLAS	1095
	20. Security Class (This Page) UNCLAS	22. Price

(See ANSI-Z39.18) See Instructions on Reverse OPTIONAL FORM 272 (4-77) (Formerly NTIS-35) Department of Commerce

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.

(6) PROCEEDINGS of the (1)

~~22ND~~ ANNUAL CONFERENCE
of the
MILITARY TESTING ASSOCIATION (22nd),
held in Toronto, Ontario, Canada,
27-31 October 1980.
Volume I.

co-ordinated by

(11) Dec 80

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

(12) 464

(14) MTA-22-80-Vol-1

co-hosted by

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

TORONTO DOWNTOWN HOLIDAY INN

27-31 October, 1980

404408

FOREWORD

The papers presented at the Twenty-Second Annual Conference of the Military Testing Association (MTA) came from the military, government, educational and business communities of the United States, Canada, Britain, Australia, West Germany and Belgium. The views expressed in them are those of their authors, and not necessarily those of the organizations which they represent.

TABLE OF CONTENTS

OFFICERS AND COMMITTEE MEMBERS

President and Chairman	LCol Glenn Rampton	CFPARU
Secretary	Capt Ted Stenton	CFPARU

Committees

Programme	LCdr Bill Shields	CFPARU (Chairman)
	Cdr Fred Hawrysh	NDHQ/DMOS
	LCdr Ian Jackson	NDHQ/DMOS
	Maj Terry Prociuk	RMC Kingston, Ont.
	Maj Reg Ellis	NDHQ/DPSRSC
	Capt Jim McMenemy	CFPARU
	Mrs Bernie Chopra	CFPARU
Social, Logistics and Finance	Maj Frank Pinch	CFPARU (Chairman)
	Dr Ed Haltrecht	Ontario Hydro
	Mr John Joaquin	Ontario Hydro
	Capt Charles Tierney	CFPARU
	Capt Fred Wilson	CFPARU
Printing	LCdr Ernest Lourme	NDHQ/DMOS (Chairman)
	Mr RWT Maloney	NDHQ/DDDS
	Mr Henry Blaszczak	NDHQ/DDDS
	Mr Mike Evans	NDHQ/DDDS
Registration	Mr Al Doran	CFPARU (Chairman)
	Mrs Bernie Chopra	CFPARU
	Mrs Grace Humfryes	CFPARU
	Mr Rob Aylwin	CFPARU
Audiovisual	Capt Jim McMenemy	CFPARU
Publicity	Capt Philip Anido	DND Office of Information

ACKNOWLEDGEMENTS

The MTA Executive Committee wishes to express, on behalf of all association members, special appreciation to the Assistant Deputy Minister (Personnel), Lieutenant-General H.A. Carswell, for delivering the keynote address. It also wishes to thank Mr Paul Godfrey, Chairman of the Metropolitan Toronto Council, and Mayor John Sewell of the City of Toronto, for hosting the reception at City Hall. It also extends special thanks to the Vimy Band, Canadian Forces Training Systems, for playing the music at the MTA banquet. Special appreciation is also expressed to the Director of Personnel Selection, Research and Second Careers, NDHQ, and to Ontario Hydro, for providing a wide variety of administrative and other support in the planning and conduct of the conference.

(5)

TIMETABLE OF PAPERS AND ADDRESSES
(AND INDEX OF PAPERS BY SUBJECT AREA)

MONDAY, 27 OCTOBER

NOTE: Because this timetable has the presented papers segregated by subject area, a separate subject index has not been included in these proceedings. All of the papers listed here were presented at the times shown, by either their author or an alternate. The text of each paper will be found in the "Contributed Papers" section, alphabetically by principal author.

1000 Steering Committee Meeting

1330 CONFERENCE OPENING AND PLENARY SESSION

President's Opening Remarks,
LCol Glenn M. Rampton,
Commanding Officer,
Canadian Forces Personnel
Applied Research Unit,
Toronto, Canada.

Keynote Address by Lieutenant-General H.A. Carswell, Assistant Deputy Minister (Personnel), Canadian Armed Forces

Why Soldiers Enlist, Reenlist and Separate, Lawrence A. Goldman, Ph.D., and Darrell A. Worstine,
US Army Military Personnel Center,
Alexandria, Virginia.

Analysis of Motivation Towards the Federal Armed Forces, Klaus J. Puzicha, Ph.D., and Adelheid B. Roepke,
German Armed Forces Psychological Services Research Institute,
Bonn, West Germany.

The MTA Publishing Review Project
Dr Raymond O. Waldkoetter,
Army Research Institute,
Fort Sill, Oklahoma.

TUESDAY, 28 OCTOBER (AM)

	St. Patrick Room	St. David Room	The Terrace
	ACHIEVEMENT TESTING (Chairman: Capt Gordon Vandyke, CFPARU)	SIMULATION (Chairman: Dr Bob Loo, CFPARU)	RECRUITING, SELECTION AND PLACEMENT Chairman: Capt Fred Wilson, CFPARU)
0830	Computer Derived Measures for Assessing Student Training, Dr. James Boone, Federal Aviation Administration Academy, Oklahoma City, Oklahoma.	Feasibility of Low Cost Simulation for Short Range Air Defense, Richard J. Carter, Army Research Institute for the Behavioral and Social Sciences, Fort Bliss, Texas.	Computer Assisted Personnel Selection for the Royal Navy, Bernard T. Dodd, Senior Psychologist (Naval), Whitehall, London.
0900	Development of a Computer Based Skill Qualification Test, Steven Pine, Systems Research Center, Honeywell, Inc., Mols, Minnesota.	Development of Selection Simulators in the German Military Aviation Psychology, Min Rat Martin Rauch, Chief Psychologist, German Armed Forces, Bonn, West Germany.	Modelling Approaches for the Optimal Allocation of Recruiting Resources. Larry R. Looper, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas.
		PERFORMANCE EVALUATION (Chairman: Capt Jim McMenemy, CFPARU)	
0930	Objectives-Based Testing in the Air War College, A. Timothy Warnock, Air University, Maxwell Air Force Base, Montgomery, Alabama.	Criteria Definition and Measurement in the Air Force Promotion System, Major John R. Welsh, Jr., Air Force Manpower and Personnel Center, Randolph Air Force Base, Texas.	Development of Enlistment Standards - A Life Model Application, Jonathan C. Fast, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas.
1000	Coffee (third floor foyer)		
1030	Implementing A Combined Criterion and Norm-Referenced Testing System, Captain Peter Wilson, Canadian Forces School of Aerospace and Ordnance Engineering, CFB Borden, Ontario, Canada.	Evaluation Criteria for Personnel Selection, Dorothy vonK. Scanland, Ed.D., Defense Activity for Non-traditional Educational Support, Pensacola, Florida.	The Automated Guidance for Enlisted Navy Applicants (AGENA) System, William A. Sands, Navy Personnel Research and Development Center, San Diego, California.
1120	Developing Local Norms for Predicting Success on the GED Test, Joseph W. Illes, Army Education Center, Ledward Barracks, Schweinfurt, West Germany.	A New Procedure for Estimating the Standard Deviation of Job Performance, Robert C. McKenzie, US Office of Personnel Management, Alexandria, Virginia.	The Development of a Systematic Counselling Model for the Canadian Forces, Captain Fred P. Wilson, Canadian Forces Personnel Applied Research Unit, Toronto, Canada.

TUESDAY, 28 OCTOBER (PM)

	St. Patrick Room	St. David Room	The Terrace
	BENEFITS AND INCENTIVES (Chairman: Capt Brian Belec, CFPARU)	PERFORMANCE EVALUATION (cont'd)	OCCUPATIONAL ANALYSIS (Chairman: Cdr Fred Hawrysh, NDHQ)
1330	A German Model of Assessment Problems, Oberst a.D. Hans-Erich Seuberlich, Vorsitzender Heer, Bonn, West Germany.	Lateral Skill Progression, LCol D.J. Slimman, National Defence Headquarters, Ottawa, Canada.	CODAP: An Overview of the Task Factor Technology, Sgt Michael C. Thew, Air Force Human Resources Laboratory, San Antonio, Texas.
1400	Scaling the Value of Incentives for Army Enlisted Military Intelligence Personnel, George W. Lawton, US Army Research Institute, Alexandria, Virginia.	Assessment of Competency Following the Blackthorn Incident by Applying a Modified MAPL Procedure, J.R. Deloney, US Coast Guard Institute, Oklahoma City, Oklahoma.	
1430	Coffee (third floor foyer)		
	SPECIAL TOPIC		
1500	"Mind Mapping" - A Tool for Planning, Notetaking, Counselling and Interviewing, Paul L. Hollander, Paul Hollander and Associates, Inc., Toronto, Canada.	IRJ: A New Technique for Measuring the Performance Rating Process, Cristina Goggio Banks, Department of Management, University of Texas at Austin, Texas.	
		HUMAN FACTORS (Chairman: Capt Pierre Lessard, CFPARU)	
1530		Human Factors Data Collection Techniques During Field Testing Richard H. Hiss, Essex Corporation, White Sands Missile Range, New Mexico.	

WEDNESDAY, 29 OCTOBER (AM)

	St. Patrick Room	St. David Room	The Terrace
	TEST CONSTRUCTION (Chairman: Dr Bob Loo, CFPARU)	OCCUPATIONAL ANALYSIS (Chairman: Cdr Fred Hawrysh, NDHQ)	TRAINING EFFECTIVENESS (Chairman: Capt Peter Wilson, CFB Borden)
0830	Techniques for Translating Tests of Psychomotor Skills into Written Tests, Dr. Frank M. Aversano and Ms Laura A. Futrell, US Army Training Support Center, Fort Eustis, Virginia.	CODAP: Applications and their Implications for Higher Level Design, Johnny J. Weismuller, University of Texas at Austin, Texas.	Cost and Training Effectiveness Analysis: From Principle to Application, Dr. Edward L. George, Director, US Army TRADOC Systems Analysis Activity, White Sands Missile Range, New Mexico.
0900	An Appropriate Number of Multiple-Choice Item Alternatives: Swanson (1976) Revisited, R. Eric Duncan, USAF Occupational Measurement Center, Randolph Air Force Base, Texas.	A Survey of CODAP Applications in Federal Civilian Occupations, Marvin H. Trattner, US Office of Personnel Management, Washington, D.C.	Cost and Training-Effectiveness Estimation in Army Materiel Acquisition, Dr. C. Mazie Knerr, Litton Mellonics, Springfield, Virginia.
0930	Cross-Validation of a Four-Parameter Item Characteristic Curve Test Scoring Model, Lieutenant-Commander W.S. Shields, Canadian Forces Personnel Applied Research Unit, Toronto, Canada.	CODAP: Introduction and Uses in a Large Public Utility, Dr. Ed Haltrecht, Ontario Hydro, Toronto, Canada.	Post Training Surveys in IBM's Field Engineering Division, Ed Magdarz, IBM Corporation, Research Triangle Park, North Carolina.
1000	Coffee (third floor foyer)		
1030	The First Military Operational Application of Item Response Theory, Thomas A. Warm, US Coast Guard Institute, Oklahoma City, Oklahoma.	Operating and Analytic Capabilities of the New CODAP System, Douglas I. Goodgame, Texas A & M University, College Station, Texas.	Vulcan Training Subsystem Effective Analysis, John D. Tubbs, USATRASANA, White Sands Missile Range, New Mexico.
1120	Effect of Item Difficulty Information on Multiple-Choice Achievement Test Performance, Captain Ralph Kellett, National Defence Headquarters, Ottawa, Canada.	The New CODAP System's Enhanced Hierarchical Clustering Capability, Richard W. Dickinson, Texas A & M University, College Station, Texas.	Factors Limiting the Measurement of Simulator Training Effectivenes, Louis F. Cicchinelli, Ph.D., Denver Research Institute, University of Denver, Denver, Colorado.

WEDNESDAY, 29 OCTOBER (PM)

	St. Patrick Room	St. David Room	The Terrace
	TEST CONSTRUCTION (cont'd)	OCCUPATIONAL ANALYSIS (cont'd) (Chairman: LCdr Ian Jackson, NDHQ)	TRAINING EFFECTIVENESS (cont'd) (Chairman: Capt Ralph Kellett, NDHQ)
1330	Validation Studies of Written Achievement Tests Used at USAF OTS, Sydney Sako, USAF Officer Training School, Lackland Air Force Base, Texas.	Identification, Definition and Measurement of Critical Flying Skills, Edward E. Eddowes, Air Force Human Resources Laboratory, Williams Air Force Base, Arizona.	Some Problems in Evaluating Training Devices and Simulators, John A. Boldovici, Human Resources Research Organization, Fort Knox, Kentucky.
1400	Using Error Rates to Select a Cut-Score, Karen N. Jones, US Coast Guard Institute, Oklahoma City, Oklahoma.	Determining Task Commonality in Navy Training: Two Methods Examined, Douglass Davis, Chief of Naval Education and Training, Naval Air Station, Pensacola, Florida.	Longitudinal Study of Foreign Language Skills, Ronald C. Jenkins, US Department of Defense, Fort George G. Meade, Maryland.
1430	Coffee (third floor foyer)		
1500	Setting SQT Cut-Scores - Ways and Meanings, Brian Charles Davis, US Army Training Support Center, Fort Eustis, Virginia.	Cognitive Structure of Technical Knowledge: A Free Association Methodology, Brandon B. Smith, Minnesota Research and Development Center for Vocational Education, Univ. of Minnesota, Minneapolis.	Effect of Participation in the OPDP (Officer Professional Development Program) on Professional Development, Captain Ralph Kellett, National Defence Headquarters, Ottawa, Canada.
1530	MAPL Procedures - An Item Evaluation Technique for Test Constructors, Phyllis P. Voorhees, US Coast Guard Institute, Oklahoma City, Oklahoma.	Occupational Analysis for Determining Job Proficiency Requirements, Sqn Ldr Michael J. Cassidy, RAAF, AFHRL, Brooks Air Force Base, Texas.	Effects of Respiration Control on Stress and Performance of Jumpmasters, William P. Burke, Army Research Institute, Fort Benning, Georgia.
		HUMAN FACTORS (Chairman: Capt Pierre Lessard, CFPARC)	
1600	Evaluation of Job Aptitude Requirement Waivers for Retrained Airmen, Mary J. Skinner, AFHRL, Brooks Air Force Base, Texas.	Paper and Pencil Testing of Geometric Radar Symbols, Richard J. Carter, Army Research Institute, Fort Bliss, Texas.	College Major and Army Officer Performance, Dr. Arthur C.F. Gilbert, Army Research Institute, Alexandria, Virginia.

THURSDAY, 30 OCTOBER (AM)

St. Patrick Room

TRAINING DEVELOPMENT (Chairman: Capt Brian Belec, CFPARU)

0830 Instructional Systems Development, Dr. Alexander M. Gottesman, Naval Health Sciences Educational Training Command, Bethesda, Maryland.

0900 The Firefinder Radar Trainer: A Training Development Analysis Approach, Raymond O. Waldkoetter, Army Research Institute, Fort Sill Field Unit, Fort Sill, Oklahoma.

0930 Marine Safety Personnel Training Program Development, D. Todd Jones, P.E., Office of Research and Development, US Coast Guard, Washington, D.C.

1000 Coffee (third floor foyer)

OCCUPATIONAL ANALYSIS (Chairman: LCdr Ian Jackson, NDHQ)

1030 Physical Demands of Air Force Occupations: A Task Analysis Approach, Sherrie P. Gott, Ph.D., Air Force Human Resources Laboratory, Brooks Air Force Base, Texas.

1120 Validation of a Job Analysis Questionnaire through Intensive Observation, Carol A. Johnson, Ph.D., McFann, Gray and Associates, Inc., Presidio of Monterey, Monterey, California.

St. David Room

PREDICTION AND VALIDATION (Chairman: Maj Reg Ellis, NDHQ)

Validity Comparisons of Verbal and Nonverbal Measures of Vocational Interests, Joseph L. Weeks, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas.

Some Factors which Limit the Predictability of Employee Performance, LCdr Earl H. Potter London, Connecticut.

Characteristics of High Achievers in Army Officer Basic Courses, Dr. Arthur C.F. Gilbert, Army Research Institute, Alexandria, Virginia.

Impact of Stress in Air Combat: Models for Predicting Performance, Jeffrey E. Kantor, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas.

A Design for Validating Selection Procedures for Groups of Jobs, Richard Alan Lilienthal, US Office of Personnel Management, Alexandria, Virginia.

The Terrace

ORGANIZATIONAL CHOICE (Chairman: Capt Charles Tierney, CFPARU)

Reenlistment Factors for First Term Enlisted Personnel, Stephan J. Motowidlo, State University of New York, Binghamton, New York.

Self-Concept, Organizational Image and their Relationships to Organizational Choice, Lt C.D. Emerson, 21 Radar Squadron, St. Margarets, New Brunswick, Canada.

Trends in Reenlistment Intent: General vs Occupation-Specific Analysis, Lt Col Jimmy L. Mitchell, USAF Occupational Measurement Center, Randolph Air Force Base, Texas.

SPECIAL TOPICS (Chairman: Capt Jackie James, CFPARU)

Use of Audio-Visual Media in Selection, Barbara J. Price, Commissioner, Civil Service, Toledo, Ohio.

An Examination of Black Accession and Performance in Naval Aviation Training, Annette G. Baisden, Naval Aerospace Medical Research Laboratory, Pensacola, Florida.

(2)

THURSDAY, 30 OCTOBER (PM)

	St. Patrick Room	St. David Room	The Terrace
	OCCUPATIONAL ANALYSIS (cont'd)	PREDICTION AND VALIDATION (cont'd)	SPECIAL TOPICS (cont'd)
1330	The USAF Occupational Analysis Program: An Evolving Technology, Dr. Walter E. Driskill, USAF Occupational Analysis Program, Randolph Air Force Base, Texas.	College GPA and Job Performance: Application of Meta-Analysis, Dr. Brian S. O'Leary, Personnel Research and Development Center, Washington, D.C.	Adaptation to USAREUR, Richard J. Orend, Human Resources Research Organization, Mannheim, West Germany.
	UNIT EFFECTIVENESS (Chairman: Capt David Horton, CFPARU)	PSYCHOLOGICAL FACTORS (Chairman: Maj Terry J. Prociuk, RMC)	COMPUTERIZED TESTING (Chairman: Mr. Al Doran, CFPARU)
1400	A Multi-Attribute Utility Measurement Approach to Organizational Evaluation, Paul Best, Ph.D., McFann, Gray and Associates, Inc., Presidio of Monterey, Monterey, California.	Social Anxiety, Self-confidence and Military Aptitude, Arnold Bohrer, Belgian Armed Forces Psychological Research Section, Brussels, Belgium.	Human Engineering: A Computerized Testing Station, James H. Johnson and Kathy Novak, Psych Systems, Baltimore, Maryland.
1430	Coffee (third floor foyer)		
1500	The Development of Organizational Effectiveness Measures for Security Police Units, Hendrick W. Ruck, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas.	Intellectual Performance Depending on Neuroticism and Introversion, Heinz-Jurgen Ebenrett, German Armed Forces Psychological Services Research Insitute, Bonn, West Germany.	Design of an Examinee Monitoring Station, J. Stephen Prestwood, Assessment Systems Corp., St. Paul, Minnesota, and University of Minnesota, Minneapolis.
1530	Criteria Definition for Evaluation of Army Unit Tactical Training, Angelo Mirabella, Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virgina.	The Marine Corps Job Satisfaction Program, Colonel N.K. Bodnar, Director, Office of Manpower Utilization, Headquarters US Marine Corps, Quantico, Virginia.	Design of A Test Development Subsystem, C. David Vale, President, Assessment Systems Corporation, St. Paul, Minnesota.
1600	Consistency of Unit Performance Ratings by Armor Officers and NCOs, Eugene H. Drucker, Human Resources Research Organization, Fort Knox, Kentucky.		Distributed Processing Considerations in Computerized Testing System Design, Ken Hansen, Psych Systems, Baltimore, Maryland.

FRIDAY, 31 OCTOBER

PLENARY SESSION

0900 Concept, Progress and Plans for the MTA Quarter-Century Publication, Dr. Raymond O. Waldkeotter, US Army Research Institute, Fort Sill, Oklahoma.

0930 Symposium: WOMEN IN THE MILITARY, Chairman: Major Frank Pinch, CFPARU, Discussant: Dr Bruce A. MacFarlane, Carleton University, Ottawa, Canada.

The Performance Evaluation of Men and Women in the Canadian Forces, Captain Suzanne P. Simpson, National Defence Headquarters, Ottawa, Canada.

Attitudes Toward Women's Roles and Organizational Commitment within the Canadian Forces, Lieutenant Diane G. Boyce and Capt Brian E. Belec, Canadian Forces Personnel Applied Research Unit, Toronto, Canada.

Utilization of Women in the Aircraft Maintenance Career Field, M. Suzanne Lipscomb, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas.

1130 PRESIDENT'S CLOSING REMARKS

(13)

INDEX OF AUTHORS AND CO-AUTHORS

(17)

INDEX OF AUTHORS AND CO-AUTHORS (cont'd)

INDEX OF AUTHORS AND CO-AUTHORS (cont'd)

(16)

KEYNOTE ADDRESS TO THE MILITARY TESTING ASSOCIATION
LIEUTENANT-GENERAL HA CARSWELL
27 OCTOBER 1980

Mr. President and Mr. Chairman, Ladies and Gentlemen. Good afternoon, and welcome to Toronto and to Canada from the Canadian Forces Personnel Applied Research Unit, and, indeed, from the Canadian Forces. I understand that LCol Rampton is both your President and your Chairman, and I would like to express my appreciation to him for giving me the opportunity to be here today before this very distinguished group.

You have heard who I am and something of my background. You may have observed that I have spent all of my career until 1977 outside National Defence Headquarters. You may well ask, therefore, what qualifies me to give a keynote address to a meeting of the Military Testing Association? It is true that I do not have long experience in this type of work. What I do have is a responsibility for policies and decisions affecting large numbers of people as well as military effectiveness, and I also have an appreciation that those decisions need the type of advice and assistance that only you can give. That may be a marginal qualification, but it certainly is based on a sincere concern with what you do. I could be described as a vitally interested consumer of your services.

As I have been asked to set a keynote for your meeting, I would make it "cooperation between you the researchers and those of us called planners and policy-makers". Actually, that is all I was asked to do, so I could sit down now. But to justify my trip from Ottawa I would like to take a few minutes to expand on the basic theme.

During the past few years I have become acutely aware of a tendency for Western Armed Forces to emphasize equipment technology perhaps at the expense of not giving due attention to the need for personnel to operate that equipment, and, indeed, to the needs of those personnel. I except that we may soon reach a point where the human element could be the limiting factor (both quantitatively and qualitatively) in the effective operation of some of our weapons systems. It is the task of all of us to ensure that this does not occur. With that in mind, there are two themes that run through what I have to say to you today. First, we need to learn how to make better use of our behavioural science knowledge; and second, behavioural scientists have to improve their ability to translate their knowledge and the results of their research, so as to make it even more useful to policy makers.

In the next few minutes, I'd like to review some evidence that, to me, foretells personnel problems that must have an impact on personnel planning for the future. Secondly, I'll provide some examples of behavioural science input that has been used in the development of special manpower programmes in the Canadian Forces. Then, I'd like to suggest how we, the policy makers and planners, and you, the behavioural science advisors, can work together to help solve our personnel problems and to learn how we can best adapt to change.

Military institutions, especially those dependent on volunteers for recruits, are extraordinarily sensitive to social and economic changes in their host societies. This connection has been demonstrated in studies that indicate that these trends, including socio-demographic change, affect the forces' ability to attract, train, retain and effectively employ its manpower. Briefly here are some of the more critical Canadian trends, which seem to be paralleled in other western nations:

- We are seeing a drastic reduction in the size of what has been our prime recruiting population, that is the 17 to 24 year old males. This trend is forecast to continue at least to the end of the century.

- There has been a significant increase in the level of education of labour force entrants. This is changing their preferences and expectations, making them less likely to pursue a military career and more likely to voluntarily OPT out of military service if they do enter.

- There is a trend towards early skill acquisition among youth through attendance at community colleges and other civilian vocational training institutions. This trends to render military trade training less attractive and, in some cases, even redundant.

- There is an increasing demand for high-skill technicians, a demand that will soon far exceed supply. As our technology becomes more complex, our technicians must become more skilled, and thus more attractive to industry.

- There is much increased demand for family stability driven by a variety of factors, not the least of which is the increasing incidence of more than one salary earner in the family.

- Another factor that has had, and will continue to have, an impact on military personnel policies is Human Rights Legislation. The Canadian Human Rights Act was brought into force in 1978, and it prohibits discrimination for a large number of factors, the most significant of which, from a military point of view, are sex, age and marital status. This social development will have great significance to our military forces.

These are just some of the factors that are bound to affect us.

In Canada, and I understand in other nations as well, the military has not always kept pace with these changing social and economic realities. For example, until very recently, we have concentrated our recruiting effort almost exclusively on the segment of the manpower pool which includes the least educated and the most turnover-prone of the Canadian labour market. This has resulted in recruiting shortfalls and in increased personnel training and replacement costs. Recently, we have started to turn this around. Following the advice of our researchers, we have taken positive steps such as to subsidize tradesmen in specialized courses, for instance marine engineering technical training which is given in civilian institutions. Also, we have started to look seriously at our recruiting practices. We are considering giving advance standing to individuals holding civilian skills that parallel those required in the forces. We do this in special cases now, but we plan to broaden the programme a good deal. These plans, the marine engineering training, and trade skill recognition are two examples of ways to hold down training costs and take advantage of labour market trends.

Obviously, military forces must evolve as society changes. At the same time the military is unique. Certain aspects of the military, such as the concept of command, and unlimited liability set them apart from the rest of society. There is much that is unattractive about military life, and one could question why anyone would join. Many of the things that lead people to enrol, and stay in, are related to those very things that set the military apart from society in general. Thus there are risks involved in making social changes too quickly. The challenge is to find the balance between changing enough to remain credible, but at the same time preserving those peculiar military institutions, values and traditions which make a positive contribution to effectiveness and morale.

These are the types of problems that we face, and it is safe to say that my colleagues and I are becoming more sensitive to the need for behavioural science help to enable us to make better decisions in solving them. My counterpart in charge of materiel, for example, has argued strongly for the need for behavioural scientists on all committees and working groups responsible for the procurement of new equipment. This includes weapons systems.

One of our personnel programmes, The Land Operations Trade Reassignment Programmes (LOTRP) is a classic case of how behavioural science research can be used for direct improvement. LOTRP was designed to regularize the supply of personnel into the combat arms trades and to reduce overall attrition in the CF. It provides combat arms personnel with the right to move to other skill areas within the military after serving for a given period. In this way they can realize both their mobility aspirations and their wish to acquire technical skills.

Researchers at the Canadian Forces Personnel Applies Research Unit (the unit commanded by LCol Rampton, your President) worked hand-in-glove with our policy makers to develop and implement this program in 1976. Its creation was based partially on the trends in society that I mentioned earlier. LOTRP appears to have worked quite well, as attrition has reduced, and further, the experienced combat arms soldier has brought maturity and stability to the young recruits in his new training group. We want to test that perception, though, so we plan to do a full-scale evaluation of LOTRP's effectiveness in the near future. Our behavioural science researchers will play a large part in that evaluation.

A more global use of our behavioural specialists is in our employment-of-women studies. Since January, we have been conducting a series of trials on the introduction of women into land combat support units; on a support ship of our fleet; to an isolated base in the high Arctic; and, as aircrew in non-combatant aircraft. Our behavioural science advisors have assisted to a very great extent in developing evaluation strategies and they will be providing analyses of data to the trials directors on a continuing basis. Our decisions regarding where and how to employ women in our force and how to overcome the problems of fully integrating women will depend on the results of this series of trials.

The programmes that I have talked about are just some examples of how behavioural science has of been help to us. I would not, however, want to leave you with the impression that the Canadian Forces have found a magic formula for making optimal use of the results of personnel research. The fact is, that in order to successfully grapple with our personnel problems of the future we must foster and extend the cooperation that now exists between the researcher and the operator, and between the researcher and the policy maker. (I'll leave the question of cooperation between the operator and the policy maker to your imagination). Our problems of manning and our problems of the interface between people and technological advances will increase. It follows from this that all of our future programmes must be evaluated in terms of their impact on our ability to maintain appropriate personnel quantity and quality levels. We must think in terms of both cost and operational effectiveness, with the latter being the driving consideration because it is our reason to be. These evaluations will provide a challenge for our behavioural researchers. There will be a special challenge for those concerned with developing better methods of human measurement.

One of the reasons that we need specially trained researchers is that laymen, like me, too easily jump to conclusions. It is of great concern to me, and I'm sure to others in positions similar to mine in our respective forces, that personnel researchers not only study and raise problems, but, that their results provide advice and guidance towards

practical solutions. Handing me a research report that concludes only that more research is required, is not very helpful. A large number of our problems beg immediate solutions and they cannot wait until "all the facts are in". Usually the decision maker does not have the luxury of time to wade through thick, inconclusive reports or to wait for the "further definitive report" which may never come. Speaking about the use of research results that are already known, we must search for ways in which this knowledge can be brought to the attention of decision makers and also to the aid of those who operate our personnel system. We are considering a number of ways to do this. One promising proposal calls for the employment of "staff implementors" who are trained in research. These advisors would collate material and would work directly with policy makers and personnel managers.

The point I want to emphasize is that we must find more direct means of putting behavioural science research findings into practice. We must learn to take full advantage of your know-how to help solve manpower problems that face us. We share a responsibility to ensure that this occurs. You, the researchers, must strive to translate your results into usable form for policy implementation, and we, the policy makers, must learn to make full use of the knowledge and experience available. In the last fifty years or so, we have come to rely on the engineers and physical scientist to provide technically sophisticated equipment to increase our battlefield effectiveness. For the next few decades, it is likely that behavioural scientists like yourselves will be equally important in helping to ensure that we have the right numbers and the right kind of skilled personnel to exploit that equipment to the full. The man/machine interface may not be the most critical topic of study in this regard. Interpersonal dynamics such as indoctrination and socialization, unit cohesion, and, leadership practices may prove to be the crucial elements in our ability to develop and maintain effective forces for defence.

It should be obvious now that I fully support your organization and I want to express a deep sense of gratitude to those of you who work for us in the behavioural science areas. Your efforts to come to grips with the complex problems facing us are now receiving a measure of the acclaim to which they are entitled.

In conclusion, I'd like now to speak as a decision-maker to you as researchers. I would think that I speak for most of my group when I direct your thoughts to the following matters which concern all military organizations.

- How can we identify and then describe the command and leadership structures required for the battlefield of the future? A battlefield where better educated and socially aware soldiers will function in smaller, more independent groups?

- Can we measure and reconcile the incompatibility between the needs of the military force, and the needs and aspirations of the individual?

- How can we measure the effect of different rotation practices on unit cohesion? To what extend do one-for-one exchanges degrade unit performance?

- What will happen to military participation as family styles change? What can be done to meet the needs of families where both spouses are employed?

- Can we measure the difference in attitudes and values between operational and support personnel? If there is a significant difference, how can it be reconciled to maintain the integrity of combat teams?

- How can we find the balance between the need to change, and the need to preserve the uniqueness of the military?

That is only a short and partial list but it should go a long way to illustrate the extent to which we planners need your help.

We all are serving in exciting times and I envy you the mental stimulation of your work. (Not that my job is dull.) As I said at first, "our key word and key note is cooperation". I wish you well in your search for the answers to our questions.

CONTRIBUTED PAPERS

This section contains copies of the papers presented, arranged alphabetically by principal author. The pages have been numbered using letters followed by numbers. The letters are the first ones of the principal author's surname, the minimum number of letters needed to distinguish him or her from other authors. The numbers are simply the page numbers of the paper. For example, the paper by Smith is numbered from SM-1 to SM-10. If a paper was not received before the publication deadline, only the abstract is included.

No attempt has been made to edit papers. Therefore, their correctness is the sole responsibility of their authors.

TECHNIQUES FOR TRANSLATING TESTS OF PSYCHOMOTOR SKILLS INTO WRITTEN TESTS[1]

BY

Frank M. Aversano, Ph. D.

and

Laura A. Futrell

A Paper Presented at the 22nd Annual Conference of the Military Testing Association, Toronto, Ontario, Canada, October 29, 1980.

1. The views, opinions, and/or findings contained in this paper are those of the authors and should not be construed as an official Department of the Army position, policy, or decision.

Introduction [2]

The belief that the reading ability of enlistees has declined, the dislike and fear of testing among enlistees, the flight from high school, and the need to produce reliable tests of high fidelity to manage training, are four reasons among many for the movement toward hands-on testing, and away from written testing. Accordingly, a large portion of US Army testing employs the hands-on approach to testing. The Hands-On Test (HOT) is that part of the Skill Qualification Test (SQT)[3] which tests the soldier's ability to perform critical tasks on actual job equipment or simulators (Guidelines, 1977)[4]. The HOT is a highly structured test which must be administered and scored according to very specific instructions. All evidence suggests that the HOT is well received by the field and is one of the best training tools available because it tells commanders and soldiers where to direct training. The major complaint about the HOT is that it requires extensive equipment, people, time and other resources to administer, although recent HOT development guidelines have changed to allow production of HOT that can be administered in a typical unit. Nevertheless, it has not always been possible to test soldiers in the hands-on mode because soldiers are assigned away from units that could support the testing, equipment is not accessible, and there is a lack of qualified personnel to administer the test. Since Army testing is critical to the management of training and the selection of individuals for promotion, and the tasks tested in the HOT are critical, an alternative method to the HOT had to be found. Furthermore, the method had to be inexpensive, easy to administer, reliable, and valid. The method used was to return to paper based testing using illustrated written items. The test is called the Alternate Hands-On Test or AHOT. The rationale behind this decision was that soldier training would be improved if some type of test feedback was available. The purpose of this study is to review the techniques of writing these items and review data on the validation of these items.

---

2. The authors wish to express their appreciation to MAJ Melvin H. Sutton of the US Army Transportation School for his help in securing the data for this study.

3. The other parts of the SQT are the Job Site Component (JSC) and Skill Component (SC). The SC is a performance-based, highly illustrated written test, while the JSC tests the soldier during performance of his job.

4. The two major sources for this paper are TRADOC Pamphlet 351-2 (Draft), Guidelines for the Development of Skill Qualification Tests, 1980, and TRADOC Pamphlet 351-2, Guidelines for the Development of Skill Qualification Tests, Dec 1977. In the text of this paper they will be referred to as (Guidelines, 1980) and (Guidelines, 1977) respectively.

Part I - AHOT Techniques

The techniques for writing AHOT are not unlike those techniques used for writing items for a paper and pencil test. One of the major differences is that the item writer has the benefit of the HOT (described above). This includes the HOT scoresheet which is a behaviorally stated checklist of task performance steps. Figure 1 is an example of a typical HOT scoresheet. The scoresheet or checklist supplies the item writer with behavioral statements on which to base written items, limits the writer to the performance steps tested in the HOT, and focuses his attention on the critical areas. This technique increases the face validity of the AHOT and increases the content validity as assessed by subject matter experts. Item writers are typically senior non-commissioned officers (NCO) with recent job experience. In addition to their own experience, item writers use field manuals, technical manuals, and a Soldier's Manual as a content source for items. The Soldier's Manual is the result of extensive job and task analysis, and contains task statements that describe some relatively small part of a soldier's job. In addition to the task statement, the Soldier's Manual contains a description of the conditions under which the task is usually performed, the standards of performance, and the steps which describe accomplishments of the task from beginning to end. These steps are referred to as performance measures or performance steps. An example of a Soldier's Manual page is provided at Figure 2. What follows is a step by step description of the techniques used for writing AHOT based on the steps provided in the Guidelines (1980), while the discussion of these steps is provided by the authors.

1. Review Task Conditions: The first step in writing AHOT is to review the task conditions. Task conditions are the "givens" in a situation, and the environment within which the task usually occurs. Task conditions must be reviewed because the item writer must simulate them as closely as possible. The assumption here is that the closer the conditions are to the actual job situation, the more valid the test will be. The item writer must first list the conditions that typically surround the task by reviewing the front-end analysis and the HOT. After this list of conditions is established, the item writer must decide what conditions are most critical and then decide how to duplicate these conditions in the written test question. Two options are available to the item writer: reproduction or simulation. The best choice usually is reproduction because once again we believe that the less removed from the actual job the more valid the test item. For example, job performance aids used on the job like checklists, schematics, and tables can be easily reproduced in a written test as can pages from technical manuals and forms.

Conditions that cannot be reproduced, such as actual pieces of equipment like tanks, landing craft, weapons, and tools, or environmental conditions, like temperature, wind velocity, sand storms and rainfall, must be simulated. The best simulations involve pictures, illustrations, drawings or some other non-verbal method of describing the job conditions. The least effective, but cheapest and most often used method is the written word. An advantage of the written word is that it allows the item writer flexibility in simulating conditions. In fact, most AHOT contain a brief written statement called "general situation" which attempts to place the soldier in a realistic job situation.

2. Write the item. The next step is to write an item that is a match to a performance measure in the HOT. When an item is a match it covers the exact same content as the performance measure. The goal is to produce a typical, symbolically presented, criterion-referenced item that states a problem in the stem and offers five alternatives from which to select. However, the item writer does have two types of written items available to him -- the performance-based item and written-performance item. A performance-based item is an item that asks questions about correct task performance. The items are based on the examinee's knowledge of correct performance of the task even though he does not have to perform the task to answer the items (Guidelines, 1977, p 1-7). Performance-based items are the most removed from the job situation and probably present the greatest threat to validity. Nevertheless, performance-based items still make a contribution to our knowledge of examinee capability. The Guidelines describe the value of this type of item very clearly: "While knowing how to do a task does not prove that a soldier has the physical skill to do it, soldiers who do not know how to do the task cannot do it (the task) in any case. Therefore, soldiers who fail this kind of alternate test probably cannot do the task. Soldiers who pass it may or may not have the skill to do the task." (Guidelines, 1980, p 149)

Three types of performance-based items exist with two modes of presentation. The two modes of presentation are written items and written items with illustrations. Performance-based items that use illustrations appear to be the most valid because they simulate stimuli that are found on the job. The three types of performance-based items are as follows:

Knowledge Item. This type of item asks the soldier if he knows what to do while performing some physical task. The item is most useful when it is accompanied by an illustration because illustrations can be used to test recognition. For example, regarding the task, "Don the Protective Mask," the soldier can be shown a picture of soldiers performing various steps of the task and asked if the performance depicted in the picture is correct or incorrect. The assumption here is that the soldier who is capable of recognizing correct task performance is more likely to be able to perform the task than a soldier not able to recognize correct task performance.

Sequence Item. The sequence item asks the soldier either when some task should be performed or the order in which it should be performed. The information is critical to correct task performance and we once again believe that the probabilities for successful task performance are greater for soldiers who have this knowledge than for soldiers who do not have it. For example, a soldier is shown a slide for six seconds and asked when he should throw the grenade. The slides show out-of-range enemies, in-range enemies, friendly personnel, and tanks.

Sub-Step Simulation Item. In this type of item the soldier is asked to simulate some sub-step of the task. For example, the soldier could be asked to mark the correct keys on a typewriter or trumpet that would produce a particular letter or note. However, the sub-step simulation is not recommended.

The other major type of item is the written-performance item. In a written performance item some tasks or parts of tasks can be performed almost exactly as they are on the job. These tasks are characterized by the fact that the correct answer cannot be recognized without performing the task (Guidelines, 1977, p 1-9). The written-performance item has excellent face validity and has good content validity as assessed by subject matter experts who review the items. Examples of written-performance items are those that involve computation of travel vouchers, or computing distances between two points on a map, or advising what rnuway to use given traffic load, wind velocity and other factors. The performance on the test is so close in most cases, or virtually exact to the job situation, that one could expect the items to be valid and predictable of performance. Unfortunately, many psychomotor tasks do not lend themselves to this type of item.

3. Review the AHOT. This review is performed by another subject matter expert to insure that the AHOT is feasible and doctrinally accurate.

4. Write Adminsitrative Instructions. This section involves the review and specification of personnel and equipment needed to administer the AHOT, test conditions, preparation before the test, administration and scoring requirements, and the training of scorers. The item writer must, of course, keep these requirements to the bare minimum or make the requirements as general as possible, if not the purpose of the AHOT is defeated.

5. Validate. The final and perhaps the most important step in the development process is to validate the test questions. The validation involves trying out the AHOT with potential populations and subject matter experts followed by interviews designed to identify problems. However, the essence of the validation involves the AHOT ability to discriminate between soldiers who can perform the task (performers) and those who cannot (non-performers). In this way concurrent validity can be established for the AHOT.

Part II - Validation

The second part of this study deals with the review of validation data and an assessment of the validation to get an indication of its effectiveness. The null hypothesis was stated as: $R_{bis} \leq 0$, which is to say that the correlations between an individual's classification as a performer or non-performer and the number of items passed would be less than or equal to zero ($p \leq .05$).

Method

Subjects

The sample used in this study was soldiers (N = 7 per task validation) in the US Army Transportation Corps located at Fort Eustis and Fort Story, Virginia. The rank ranged from E-1 through E-6 with skill levels ranging from SL1 to SL4. Data regarding the sex of the subjects was unavailable although a majority of the subjects were male.

Procedure

The Hands-On Tests (HOT) for the entire Transportation Career Management Field (CMF 64) were surveyed. Only those AHOT that were acceptable according to the Guidelines (1980), i.e., valid, were used in the study. A list of the Military Occupational Specialities (MOS) studied is given in Figure 3.

The first step was to establish whether or not a soldier was a performer or non-performer. To differentiate performers from non-performers, at least four expert observers rated a soldier on task performance using the Soldier's Manual task summary or HOT scoresheets as a checklist. Three out of four expert observers had to agree before a subject was classified as a performer or non-performer. At least seven subjects were required to validate an AHOT with at least three classified as performers and three as non-performers. If enough performers and non-performers were not available for a task, it could not be validated and was eliminated from consideration. Items passed by at least as many non-performers as performers, and items passed by more performers than non-performers were used in the study. Following the validation rules in the Guidelines (1980), any item passed by more non-performers than performers was eliminated from future consideration in this study. A sample data sheet is provided at Figure 4. Biseral correlations were computed using the performance classification as the artificial dichotomus variable and the number of items passed as the continuous variable. One directional t tests were performed to determine if the correlations were significant, ($p \leq .05$).

Items used in the study were also classified as either written performance or performance-based to determine if one type of item was more effective than the other.

Results and Discussion

Tables 1 and 2 summarize the results. A total of 89 AHOT were subjected to the validation procedure and 33 were accepted as valid measures of soldier performance (see Table 1). A valid AHOT is operationalized as one that discriminates performers from non-performers. Biseral correlations were

computed on the 33 valid AHOT. Table 2 summarizes the correlations on each one of the AHOT and indicates whether or not the correlation was significant at the .05 level. It also lists the MOS and the number of performers and non-performers. Correlations ranged from -.113 to +.867 and of the 33 computed, 9 were significant. Assuming that the AHOT are valid measures of performance, the results suggest that paper based items are of limited value in obtaining information about psychomotor performance. However, a number of cautions are in order because paper based items with illustrations have been used effectively to predict many different behaviors. The findings may be due to a number of design artifacts, for example, sample size was small and as a result the correlations may be spurious or at the very least unstable.

Nevertheless, the nine significant AHOT were analyzed in an attempt to discover why these AHOT were significant and why the others were not significant. Three major variables were analyzed: amount of task coverage by the AHOT items, use of illustrations and diagrams, and the type of item used (performance-based vs written performance).

Task coverage was operationalized as the number of items divided by the number of performance measures for a task in the Soldier's Manual times 100 percent. This number gave a percentage of task coverage. Unfortunately, the percentage is not very meaningful, because some tasks have performance measures (PM) that describe in detail how the task should be performed, while other PM have very little detail. The analysis suggested that there was no meaningful relationship between the amount of task coverage and the significance of an item, although the data is not comprehensive enough to completely eliminate task coverage as a factor.

Even though matching techniques are of dubious value, we compared the 9 significant AHOT with 9 evenly matched non-significant AHOT to discover if illustrated items improved the predictability of the AHOT. No significant difference was found between the two groups. However, the authors believe that illustrated items are more psychometrically sound than non-illustrated items.

The other variable selected for analysis was the type of item used in each group. It was found that the significant group used 8 written performance items while the non-significant group used 1 written performance item. A t test was performed to see if the difference was significant at the .05 level and revealed that the difference was not significant. Even though it was not significant, the finding may suggest where future work could be done since written performance items offer more apparent validity than performance-based items.

A major criticism of the study is that since the correlations were based on validation data rather than actual data, some shrinkage in the correlations would occur between HOT and AHOT. This would further reduce the number of significant relationships. A better study, and one planned for the future, would be to administer the AHOT to soldiers taking the HOT using larger sample sizes.

Another criticism lies in the fact that the criterion group was established in two ways -- one way was to use the HOT checklist while the other was to use the Soldier's Manual performance measures as a HOT checklist. The latter method may not have been as accurate as an actual HOT despite the fact that 3 out of 4 subject matter experts had to agree that a subject was a performer.

A final criticism of the design can be found in the number of correlations computed in the study. One would expect, by chance alone, that out of 33 correlations computed some would be significant. This is indeed a justified criticism which the authors accept. The study was based upon available data which precluded elegance in design.

Conclusion

Given the facts that the sample size is so small and that the study was based on validation data rather than actual test data, conclusions are tenuous. Nevertheless, the following conclusions are made.

1. The use of written items to test psychomotor performance should be used with caution and when no other method is available.

2. Written performance items are better than performance-based items.

3. When validating items, sample size should be increased beyond the minimum.

4. Validation criteria should be made more strict to better screen items.

TABLE 1

Alternate Hands-On Tests Validation Results

SQT	MOS	SL	Number of AHOT	Number Accepted
2	57H	1	11	3
3	57H	2	7	1
4	57H	3	8	3
5	57H	4	8	4
2	61B	1	8	1
3	61B	2	6	4
4	61B	3	7	1
5	61B	4	8	1
2	61C Trk 2	1*	3	3
2	61C Trk 3	1*	3	3
3	61C Trk 2	2*	2	2
3	61C Trk 3	2*	3	3
2	61F	1	6	1
3	61F	2	6	3
2	64C	1	2	0
3	64C	2	1	0
		TOTAL	89	33

*Actually two different tests.

TABLE 2

Summary of Correlations

MOS	SL	Task Number	No. Perf	No. Non-Perf	r	SIG
57H	1	1119	4	3	.73	YES
		1121	4	3	0	NO
		1122	3	4	.84	YES
57H	2	2115	3	4	.71	YES
57H	3	3011	3	4	0	NO
		3012	4	3	-.11	NO
		3013	4	3	0	NO
57H	4	4070	3	4	.62	NO
		4066	4	3	.49	NO
		4046	4	3	.47	NO
		4042	4	3	.49	NO
61B	1	1036	4	3	35	NO
61B	2	2080	4	3	0	NO
		2083	4	3	.47	NO
		2084	4	3	.76	YES
		2085	4	3	.47	NO
61B	3	1036	4	3	.35	NO
61B	4	2044	3	4	.80	YES
61C Trk 2	1	1019	3	4	.75	YES
		1024	3	4	.70	YES
		1040	3	4	.62	NO
61C Trk 3	1	1018	3	4	.56	NO
		1019	3	4	.35	NO
		1040A	3	4	.86	YES
61C Trk 2	2	2001	4	3	.42	NO
		2130	3	4	.64	NO
61C Trk 3	2	2001	4	3	.19	NO
		2067	3	4	.16	NO
		2124	3	4	.86	YES
61F	1	1016	4	3	.46	NO
61F	2	1006	4	3	.41	NO
		1014	3	4	.50	NO
		1016	4	3	.60	NO

FIGURE 1

SCORESHEET

UNIT 2. CONSTRUCT A MONKEY FIST (TASK 551-718-2080)

INSTRUCTIONS TO THE EXAMINEE: (Read these exact words out loud.) "LET ME HAVE YOUR ATTENTION. YOU WILL MAKE A MONKEY FIST. ALL MATERIAL AND EQUIPMENT NEEDED TO MAKE THE MONKEY FIST IS PRESENT AND IN USEABLE CONDITION. A SEIZING IS NOT REQUIRED FOR THE TEST. I CANNOT HELP YOU DURING THIS TEST. IF YOU COMPLETE THE TASK BEFORE THE TIME IS UP OR YOU CANNOT PERFORM THE TASK, LAY THE LINE ON THE TABLE AND TELL THE SCORER YOU ARE THROUGH. A CHAIR IS PROVIDED FOR YOU IF YOU DO NOT WISH TO STAND. (Pause) DO YOU UNDERSTAND THE INSTRUCTIONS?" (If an examinee does not understand, repeat the instructions word for word. If an examinee still does not understand, tell him, "DO THE BEST YOU CAN.") Pause at least 5 seconds, then say, "YOU WILL HAVE 10 MINUTES TO COMPLETE THIS TEST WHEN I GIVE THE COMMAND, BEGIN."
Begin timing when you give the command. After 10 minutes, announce loudly enough for the examinee to hear, "STOP".

PERFORMANCE MEASURES: (Product is scored.)

(Scorer: Refer to sample display board and compare examinee's monkey fist with model.)

	PASS	FAIL
1. There must be three series of three turns.	____	____
2. A half hitch must be tied with the bitter end on standing part.	____	____
3. Monkey fist must be oval in shape and no larger than a baseball.	____	____
4. Task must be completed in 10 minutes.	____	____

NOTE: Examinee's monkey fist must look like the sample monkey fist mounted on the display board. Seizing is not required.

GO	NO-GO
☐	☐

STANDARD. The examinee is scored GO if he passes all of the performance measures. The examinee is scored NO-GO if he fails any of the performance measures. If the examinee receives a NO-GO, tell him why and record here the performance measure(s) failed and a brief explanation to show the cause of the NO-GO.

_____

_____

_____

_____

SCORER'S SIGNATURE

FIGURE 2

Sample Soldier's Manual Page

FM 55-61B

551-61B-2080

TASK: Construct a monkey fist

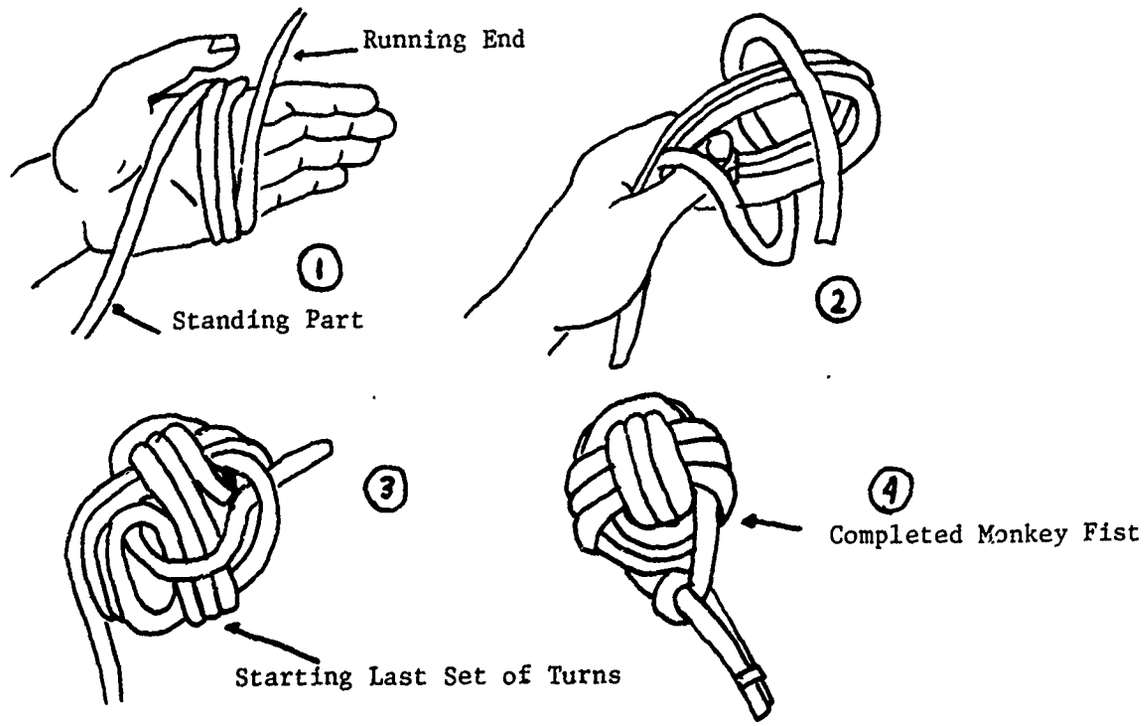
CONDITIONS: Task will be performed aboard a vessel or ashore when heaving lines are made, in all climate and weather conditions and under the general supervision of the boatswain. Cotton line, a knife, and sail twine are required.

STANDARDS: Individual must follow procedures listed below in sequence. Monkey fist must be oval in shape.

STEPS FOR ACCOMPLISHING TASK:

1. Place the standing part in front of the thumb.
2. Place three sets of turns around the hand going away from you.
3. Remove the turns from the hand.
4. Insert the forefinger in the turns, the thumb on top of the turns.
5. Place three sets of turns over the first set going toward you.
6. Place three sets of turns through the first set and over the second set.
7. Work slack back toward standing part.
8. Tie half hitch with bitter end on standing part.
9. Sew flat seizing above the half hitch.

REFERENCES: NAVTRA 10121-E, chapter 3; RT-443, chapter 1.



MONKEY FIST

FIGURE 3

MOS Codes and Titles

MOS Title	MOS Code
Terminal Operations Coordinator	57H10
	57H20
	57H30
	57H40
Watercraft Operator	61B10
	61B20
	61B30
	61B40
Watercraft Engineer	61C10 (Trk 2)
Trk 2 - Amphibians	61C10 (Trk 3)
Trk 3 - Landing Craft	61C20 (Trk 2)
	61C20 (Trk 3)
Marine Hull Repairman	61F10
	61F20

FIGURE 4

Sample Data Sheet

Test: _____

Task No: _____

_____

Item Number

	1	2	3	4	5	6	7	8	9	10
Number of Performers										
# Passing										
% Passing										
Number of Non-Performers										
# Passing										
% Passing										
Item Classification:										
Performance-based Item - J										
Written Item - K										

BAISDEN, Annette G., Naval Aerospace Medical Research Laboratory, Naval Air Station, Pensacola, Florida.

AN EXAMINATION OF BLACK ACCESSION AND PERFORMANCE IN NAVAL AVIATION TRAINING (Thu A.M.)

This paper reports on an analysis of black student accession to, and performance in, naval aviation training. The first report described in this overview involved a comparison on all black and a random sample of white civilian procured applicants for naval aviation training during 1976-1978. Comparisons were made according to performance on selection tests, recruiting area, college major, and reasons for non-selection and declination.

The second report analyzed the performance of all black students and a matched sample of white students in naval aviation training during 1973-1976, and focused on attrition and such performance indices as peer rating, officer-like-qualities, academic grades, and flight grades.

In the third report three variables, college major, grade point average, and college racial composition, were analyzed to reveal relationships to the variables in the study just described.

The fourth and final report examined black students and a matched sample of white students in naval flight officer training.

Findings of these studies are discussed in terms of relevance for recruiting programs, preparatory schools and remedial instruction.

AN EXAMINATION OF BLACK ACCESSION AND PERFORMANCE IN NAVAL AVIATION TRAINING*

Annette G. Baisden
Naval Aerospace Medical Research Laboratory
Naval Air Station
Pensacola, Florida 32508

At the request of the Naval Military Personnel Command (NMPC-06C), the Naval Aerospace Medical Research Laboratory undertook a four-part analysis of black student accession to, and performance in, naval aviation training. The following four reports were prepared and published:

1) "A Comparison of Black Civilian Procured Applicants and White Civilian Procured Applicants for Naval Aviation Training." (1)

2) "A Comparison of Black Student Performance and White Student Performance in Naval Aviation Training." (Pilots) (2)

3) "A Comparison of College Background, Pipeline Assignment and Performance in Aviation Training for Black Student Naval Aviators and White Student Naval Aviators." (3)

4) "A Comparison of College Background, Pipeline Assignment and Performance in Aviation Training for Black Student Naval Flight Officers and White Student Naval Flight Officers." (4)

A summary of each of the reports follows.

I. A Comparison of Black Civilian Procured Applicants and White Civilian Procured Applicants for Naval Aviation Training.

This report compared the performance of the two racial groups on the Academic Qualification Test (AQT) and the Flight Aptitude Rating (FAR), the two components of the U. S. Navy and Marine Corps Aviation Selection Tests. The subject population consisted of all the black civilian applicants who took the tests during calendar years 1976, 1977, and 1978 (N = 1,232). A comparative sample of white applicants for the same years was developed by random selection on the basis of the last digit of the social security number (N = 2,469). In addition to the AQT/FAR performance, comparisons were made according to recruiting area,

---

*Opinions or conclusions contained in this report are those of the author and do not necessarily reflect the views or the endorsement of the Navy Department.

college major, and reasons for nonselection and declination. Non-selection and declination data were provided by the Navy Recruiting Command on a sample of 6,352 white applicants and 189 black applicants. A summary of the findings is presented in Table I.

Table I

Summary of Findings Comparing Black Civilian Procured Applicants and White Civilian Procured Applicants for Naval Aviation Training

Variable	Finding
AVIATION SELECTION TEST	White applicants had a higher pass rate than black applicants at three cutting score levels
	White applicants scored significantly higher than black applicants
TEST REGION	Highest percentage of black applicants were from the Southeastern region
	Highest percentage of white applicants were from the Mid-Atlantic and Far West regions
	Black applicants from the Rocky Mountain and Texas regions had the highest pass rates
COLLEGE MAJOR	Applicants with engineering, technical and physical science majors had the highest AQT/FAR pass rates
	Social science and education majors had the poorest pass rates
NON-SELECTION	Reasons for non-selection were the same for both racial groups
DECLINATION	Reasons for declining were the same for both racial groups

(44)

II. A Comparison of Black Student Performance and White Student Performance in Naval Aviation Training.

The report analyzing black and white student performance in Naval aviation training focused on rates of completion and attrition, reasons and stages of attrition, and the indices of performance listed below.

- Peer Rating - A peer evaluation grade limited to officer candidate students.
- Officer-Like-Qualities (OLQ) - Aviation Officer Candidate grade based upon peer rating, instructors' observation, watch, inspection and drill grades.
- Environmental Indoctrination Final (EI) - A weighted average of Naval Aviation Schools Command grades.
- Basic Flight Grade - A composite of all Primary and Basic flight grades.
- Basic Academic Grade - A composite of all Primary and Basic academic grades.
- Advanced Flight Grade - A composite of all Advanced flight grades.
- Advanced Academic Grade - A composite of all Advanced academic grades.
- Final Overall Grade (FOAG) - A composite of all Environmental Indoctrination, Primary, Basic and Advanced academic and flight grades.

All of the black students that could be identified who entered aviation training during CY73-76 were used in the analyses (N = 99). Since attrition covaries with aviation selection test scores and procurement source, and in order to eliminate the influence of changing curricula, it was necessary to select a sample of white students matched on these variables (N = 127).

The findings of this study are presented in Table II.

Table II

Summary of Findings Comparing Black Student Performance and White Student Performance in Naval Aviation Training

Variable	Finding
TRAINING GRADES	Black student naval aviators had significantly lower grades, with the exception of OLQ
RATE OF ATTRITION	No significant difference in attrition rates by race when controlling for AQT/FAR, procurement source, and class contiguity
REASONS FOR ATTRITION	Black student naval aviators had significantly more flight failures and significantly less Drop on Request than white student naval aviators
STAGE OF ATTRITION	Black student attrition remained constant across training stages White student attrition sharply decreased during Advanced training

III. A Comparison of College Background, Pipeline Assignment and Performance in Aviation Training for Black Student Naval Aviators and White Student Naval Aviators.

The major objective of the third report in the series was the comparison of performance in naval aviation training of the black students with the matched sample of white students in pilot training during CY73-76 in the study just described. Comparisons between the two races were made by college major, grade point average (GPA), pipeline assignments, and frequency of completion/attrition from the pipeline.

The black student input was further analyzed to reveal relationships between the racial background of the college attended and the previously mentioned variables.

Table III presents the findings of this analysis.

Table III

Summary of Findings of a Comparison of College Background, Pipeline Assignment, and Performance in Aviation Training for Black Student Naval Aviators and White Student Naval Aviators

Variable	Finding	
COLLEGE MAJOR	HIGHEST INPUT RATES	
	BLACK SNAs	WHITE SNAs
	Social Science	Social Science
	Business Admin	Business Admin
	Behavioral Science	Behavioral Science
	Engineering	Engineering
	HIGHEST COMPLETION RATES	
	BLACK SNAs	WHITE SNAs
	Technical	Technical
	Engineering	Business Admin
	Social Science	Physical Education
	Physical Sciences	Engineering
GRADE POINT AVERAGE	Predictive of pre-advanced academic grades for black student naval aviators	
	No predictive validity of complete/attrite	
TRAINING PIPELINES	No difference in input rates	
	No difference in complete/attrite rates	
COLLEGE RACIAL COMPOSITION	Sixty-three percent of black student naval aviators attended predominantly white colleges	
	AQT and Basic Academic grades were higher for graduates of predominantly white colleges	

IV. A Comparison of College Background, Pipeline Assignment and Performance in Aviation Training for Black Student Naval Flight Officers and White Student Naval Flight Officers.

The fourth and final report examined black students and a matched sample of white students in naval flight officer training during CY73-76 utilizing selection test scores, training grades, complete/attrite data and college background factors. The black sample consisted of 127 students. The white sample consisted of 199 students matched on the same criteria as mentioned for the students in pilot training. Table IV presents the summary of findings.

CONCLUSIONS AND RECOMMENDATIONS

In summary, this series of studies indicates that the most qualified black college graduates are not pursuing careers in naval aviation programs. The average AQT/FAR scores of the black applicants were lower than the average AQT/FAR scores of the white applicants. Since these selection tests are predictive of success in the naval aviation training program, it is not surprising that the overall attrition rate for the black student population is significantly higher than that of the white student population.

These studies have demonstrated that when black students and white students are equated on those variables historically correlated with attrition, there are no differences in student pilot attrition rates. Although the student naval flight officers (SNFOs) were equated on the same variables as the student pilot population, the overall SNFO attrition was significantly higher for the black students. It should be noted that the current selection test battery was specifically developed for use in pilot selection and has less efficiency for predicting success in naval flight officer training.

The major problem is not one of black attrition in aviation programs, but one of failure to attract qualified black applicants. In the four years from 1973-1976, an average of only one black per day took the selection test for officer candidate and naval aviation programs. Even fewer made formal application. For the years under consideration in these studies, the black input averages into pilot training and NFO training were 25 and 32, respectively. During calendar year 1979, only 26 black students entered naval aviation pilot training.

Inasmuch as graduates of predominantly black colleges are not represented in black inputs, more emphasis should be given to recruiting in black schools, especially among the potentially more successful students who major in hard sciences. Exposing blacks to aviation through such programs as the Flight Indoctrination Program may encourage more applicants. Preparatory schools, remedical instruction, roll-back and similar programs will not significantly affect the number of black naval aviators on active duty.

Table IV

Summary of Findings of a Comparison of College Background, Pipeline Assignment and Performance in Aviation Training for Black Student Naval Flight Officers and White Student Naval Flight Officers

Variable	Finding	
TRAINING GRADES	Black students had significantly lower peer ratings, OLQs, EIs, Basic and Advanced academic grades than white students	
RATE OF ATTRITION	Black SNFO overall attrition rate was significantly higher than white SNFO overall attrition rate	
	DOR highest category of attrition for both groups	
REASON FOR ATTRITION	More black students than white students attrited for academic reasons	
	More white students than black students attrited for physical and not aeronautically adapted reasons	
	HIGHEST INPUT RATES	
	BLACK SNFOs	WHITE SNFOs
COLLEGE MAJOR	Social Science Business Admin Physical Science Natural Science	Social Science Business Admin Natural Science Behavioral Science
	HIGHEST COMPLETION RATES	
	BLACK SNFOs	WHITE SNFOs
	Hard Sciences	Hard Sciences

Table IV (Continued)

Summary of Findings of a Comparison of College Background, Pipeline Assignment and Performance in Aviation Training for Black Student Naval Flight Officers and White Student Naval Flight Officers

Variable	Finding
GRADE POINT AVERAGE	Predictive of the pre-advance training grades for white SNFOs
	No predictive validity for complete/attrite
TRAINING PIPELINES	More black SNFOs than white SNFOs were assigned to the Navigation pipeline
	More white SNFOs than black SNFOs were assigned to the Tactical Navigator pipeline
	No difference between black SNFO and white SNFO attrition rates by pipeline
COLLEGE RACIAL COMPOSITION	Fifty-seven percent of black SNFOs attended predominantly white colleges
	AQT, EI Final, and Basic Academic grades were higher for graduates of predominantly white colleges
	No difference in attrition rates

(50)

REFERENCES

1. Doll, R. E., and Baisden, A. G. A comparison of black civilian procured applicants and white civilian procured applicants for naval aviation training, CY 1976 - 1978. NAMRL Special Report 79-3. Naval Aerospace Medical Research Laboratory, Pensacola, Florida, May 1979.

2. Baisden, A. G., and Doll, R. E. A comparison of black student performance and white student performance in naval aviation training. NAMRL Special Report 78-7. Naval Aerospace Medical Research Laboratory, Pensacola, Florida, 30 November 1978.

3. Baisden, A. G., and Doll, R. E. A comparison of college background, pipeline assignment, and performance in aviation training for black student naval aviators and white student naval aviators. NAMRL Special Report 80-1. Naval Aerospace Medical Research Laboratory, Pensacola, Florida, November 1979.

4. Baisden, A. G. A comparison of college background, pipeline assignment, and performance in aviation training for black student naval flight officers and white student naval flight officers. NAMRL Special Report 80- , Naval Aerospace Medical Research Laboratory, Pensacola, Florida, April 1980.

BANKS, Cristina Goggio, Department of Management, The University of Texas at Austin, Texas.

IRJ: A NEW TECHNIQUE FOR MEASURING THE PERFORMANCE RATING PROCESS
(Tue P.M.)

Valid assessment of human work performance is critical for any organization hoping to achieve optimal utilization of human resources. Because performance ratings are the most frequent method of evaluating peformance, it is imperative that rater accuracy be maximized. Recently, several researchers have stressed the need to analyze the rating process to identify the determinants of rating accuracy. A new measurement technique, Instantaneous Report of Judgments or IRJ, which captures decision-making processes as they occur during performance rating, is presented. IRJ offers several advantages over previous attempts to study process: raters report their judgments immediately and autonomously, cue selection is unrestricted, and judgments can be linked to the behaviours of the ratee. Several rating process variables are measured: number of judgments made, variation in judgments made over time, decision latency, and type of information selected. These variables can be linked to others thought to be related to rating accuracy including cognitive style, personality, and background data. Moreover, these variables can be related directly to rating outcomes such that we may be able to identify which rating behaviors lead to accuracy and which lead to error. Research findings obtained using IRJ are briefly discussed.

IRI: A NEW TECHNIQUE FOR MEASURING THE PERFORMANCE RATING PROCESS

Cristina Goggio Banks, Ph.D.

Department of Management
The University of Texas at Austin
Austin, TX 78712

Introduction

Accurate assessment of human work performance is crucial for any organization hoping to achieve optimal utilization of human resources. Human resources are best utilized when the right people are placed in the right jobs at the right time. This requires organizations to have the ability to correctly identify and distinguish between high- and low-performing employees. When organizations have this ability through valid appraisal systems, they can make various personnel decisions to make employees productive. For example, high performers would be promoted to more demanding jobs, average performers would be developed through training or goal-setting programs, and poor performers would be trained in basic skills, transferred to less demanding jobs, or simply dismissed. Thus, accurate performance appraisal is critical for making effective decisions about individual employees.

Performance ratings are the most widely used method of appraising performance. Unfortunately, performance ratings are subject to a variety of perceptual and judgmental errors. For instance, raters tend to let their evaluation of one performance dimension influence their evaluation of other dimensions (halo error). Raters also tend to rate either too high or too low than they should (leniency error). Some raters tend to give the same rating to all ratees (central tendency or restriction of range error). Errors such as these have been well-documented (e.g. Cummings and Schwab, 1973; Dunnette and Borman, 1979); Landy and Farr, 1980), and they continue to plague those who use appraisal information for making various personnel decisions. It is clear that the value of performance ratings is diminished to the extent they are based on something other than actual job performance. Indeed, rating errors undermine the utility of ratings as decision tools.

Because performance ratings play such an important role in organizational effectiveness, researchers have developed strategies for reducing errors with the expectation that accuracy would be increased. One strategy is to design appraisal formats that help raters minimize their error. An example is the development of behaviorally-anchored rating scales (BARS; Smith and Kendall, 1963) which define each dimension in concrete, behavioral terms and provide examples of actual job behaviors as anchors on the rating scale. Another strategy is to train raters to reduce rating errors by changing their rating distributions (i.e. spreading ratings out) so that raters differentiated more

between performance dimensions for each ratee and within dimensions across ratees (e.g., Borman, 1975, 1979; Latham, Wexley, and Purcell, 1975; Bernardin and Pence, 1980). Basically, raters are taught about various rating errors and then shown they could reduce their error by changing their rating distributions to achieve greater differentiation. A third strategy was to increase raters' observational skills (e.g., Bernardin and Walter, 1977). Raters are instructed to keep notes (i.e. a diary) on ratees' performance effectiveness during the appraisal period and later use this information in subsequent appraisals. This strategy attempts to increase the amount of job-related information utilized by raters in appraisals. All three approaches are aimed at increasing accuracy by reducing potential error.

Empirical studies of the success of each of these strategies suggest that some improvement in the psychometric properties of ratings (i.e., rating error) can be achieved without corresponding increases in rating accuracy (Borman, 1979; Dunnette and Borman, 1979; Landy and Farr, 1980). Why haven't these strategies been successful? How could it be that there are instances in which errors such as halo and leniency are reduced but the raters are not more accurate? The lack of encouraging findings led researchers to take a closer look at the process of performance rating. Could it be that raters do not use BARS the way they were intended? Do they ignore the behavioral anchors and consequently eliminate one of their major advantages? Is it possible that raters cannot process all the information presented to them in BARS and thus, simplify the forms, rendering them simple graphic rating scales? A desire to answer these and similar questions led researchers to develop a new strategy, that of analyzing the performance rating process to identify determinants of rater accuracy and error. Researchers hope that a greater understanding of the rating process will provide new insights into the underlying mechanisms of accuracy.

Although many have called for an in-depth study of the rating process (e.g., Bernardin and Pence, 1980; Borman, 1979; Dunnette and Borman, 1979; Feldman, 1980), few researchers have actually undertaken such studies. A formidable problem in "process" research is employing a measurement technique that accurately reflects on-going processes. This is best accomplished by having raters report their decision-making processes verbally or nonverbally as they occur. Delay of measurement (of even a few minutes) is likely to cause distortions in processes reported because of memory decay and the tencency for decision makers to infer processes where gaps are found in memory (Ericsson and Simon, 1980). Therefore, the greater delay between "process" and "process measurement," the less useful these data are for identifying important underlying mechanisms. Unfortunately, the two techniques that have been used to measure "process" suffer from distortion problems.

One technique, regression analysis, involves the development of regression equations which reflect the weight and thus, the importance of information (cues) utilized in a decision. However, regression equations recover only the functional relationships between cues and decisions; regression equations infer "process" from outcomes (decisions). However, they do not purport to capture the actual process (Hoffman, 1960). Statistically, weights

assigned to cues can be manipulated regardless of "process." For example, cues carry more weight when they have greater variance. Also, weights are determined in part by their intercorrelation with other cues. Thus, regression techniques may not be capturing anything real (Dawes and Corrigan, 1974; Hoffman, 1960; Yntema and Torgerson, 1961).

The other technique consists of decision-makers' verbalizations of the contents of their decision processes. Most "process" studies employing this technique are based on reports given after decisions are made. In other words, reports are retrospective and thus, suffer from distortion. Ericsson and Simon (1980) argue that the time and type of verbalization can have a strong influence on the validity of verbal reports. As mentioned earlier, retrospective reports have a smaller likelihood of reflecting the actual content of these processes.

A new technique was developed by the author to overcome some of the limitations of regression analysis and retrospective reporting (Banks, 1979). The technique, "Instantaneous Report of Judgments" or "IRJ," attempts to capture raters' decision-making as it occurs. Raters are provided two recording devices, a keyboard and a tape recorder, with which to report their judgements. Specific details of the procedure and recording devices are given below.

The IRJ Technique

Rater judgments are "captured" during a performance rating task. Raters view a videotape of a manager performing in a job and rate the manager's effectiveness along one of several performance dimensions. For example, raters would view a manager conferring with a subordinate and evaluate the manager's effectiveness in "Establishing and Maintaining Rapport." Raters are asked to report their judgments as they occur during the videotape. Whenever they make a judgment about the manager's performance, they report it by pressing one of seven keys on a keyboard indicating the effectiveness level of the manager (1 = low performance, 7 = high performance), and by reporting into a microphone the basis for their judgment. Verbal reports indicate what information (e.g., manager behaviors) the rater attended to when a judgment was made.

Measurement of the rating process is made possible by linking the recorded judgments to the behaviors of the ratee. Rater judgments are linked to specific sequences of manager behavior by relating the time of each key press to elapsed time on the videotape. This can be accomplished a few ways. One can develop an equation which converts time of key press to a count on the videotape (revolutions per minute), thereby pinpointing the count in the videotape at which a key was pressed. Another possibility is to interface the keyboard (of a minicomputer system) with the videotape recorder whereby the minicomputer can read the exact frame number at the point a key is pressed. Still another possibility is to have the minicomputer read the time at which a key is pressed and to match the number to

time elapsed in the film. In all three cases, the time of the key press may not be the point at which a judgment was actually made. (It may be quite later, in fact.) Therefore, it is necessary to examine raters' verbal reports to "zero-in" on exactly what information was utilized when a judgment was made.

IRJ permits measurement of several aspects of the rating process:

1. NJ: Number of judgments made during each appraisal
2. SDJ: Variation in judgments made within each appraisal
3. $\overline{SDJ}$: Variation in mean judgments across managers
4. AVGSD: Average variation in judgments across managers
5. JL: Latency before initial judgment of manager effectiveness is made
6. INFO: Information utilized during appraisal

Thus, IRJ measures how much information raters utilize, what information they utilize in forming overall ratings, the degree to which they utilize information representing different levels of effectiveness, and the degree to which raters judge managers differentially.

Because of the way judgments are collected, IRJ offers several advantages as a technique for measuring rating process:

1. Raters are free to utilize any information they consider important from the manager performances. This allows raters to vary the number of judgments made and the type of information they consider relevant. Thus, individual differences in cue (information) selection can be measured. With regression techniques, cues are selected by the experimenter.

2. Raters report their judgments as they occur. By allowing instantaneous reports, distortions due to memory effects and inference are minimized. Also, because of the type of reports requested from raters, it can be argued that raters do have access to the content of their thought processes and thus, can give reports which provide meaningful information about these processes (cf. Ericsson and Simon, 1980).

3. Raters complete the performance appraisal task autonomously. Because of the nature of the task, raters have full control over the reporting process; therefore, the experimenter is minimally involved. As a result, factors such as experimenter bias and demand characteristics which may influence reports are absent from this paradigm. These factors have the potential for changing reports in situations where the experimenter probes or questions the rater.

4. Rater judgments are linked to the behaviors of the manager. This information helps paint a picture of what the rater actually rates during an appraisal. Only by analyzing the rating process at such a microscopic (behavioral) level can one begin to identify what raters do when they are accurate and conversely when they make errors. One needs to collect information at the behavioral level to solve the riddle of why changes in rating formats and rater training have negligible impact on rater accuracy. Regression techniques cannot provide this information because they analyze the rating process at a more general (dimensional) level.

The success of this technique, however, depends on raters' access to their thought processes and their ability to articulate these processes (Nisbett and Wilson, 1977). While articulation may still be a problem, it has been shown that because of the nature of the paradigm, raters probably do have access to their processes and can report them with less distortion. In relation to previous attempts to study the rating process, IRJ seems to hold the greatest promise for yielding important new information which may lead eventually to the discovery of the determinants of rater accuracy.

Studies Using IRJ

To date, two sets of empirical studies of rating process have been conducted. The first study was a preliminary investigation of various aspects of the rating process. The purpose of this study was to describe various rater behaviors during the appraisal process and to relate rating behaviors to cognitive abilities. The second study was a replication and extention of the first study. This time rating behaviors were related to a variety of input variables (variables which contribute to or underlie rating behavior, such as cognitive complexity and detail orientation). In addition, rating behaviors were related to various outcome measures such as accuracy and halo error. The purpose of this study was to identify the linkages between input, process, and outcome variables in performance appraisal. The relationship between these three sets of variables are characterized in Figure 1. The intent of this research is to determine what underlying traits or experiences contribute to various rating behaviors and in turn, which rating behaviors lead to accuracy. Preliminary findings are presented below.

Preliminary Study of Rating Process

Three questions were investigated in this research:

(1) What information do raters use when they rate performance?

(2) What do the decision-making processes in performance appraisal look like?

(3) What cognitive variables are these processes related to?

One hundred fifty-six raters viewed and rated six managers along one of six performance dimensions. (A different dimension was rated per manager.) Raters reported their judgments following the IRJ technique. In relation to the first

question, it was found that raters tended to utilize different information when they evaluated the same manager on the same dimension. That is, they disagreed concerning what information is relevant for evaluating the performance of the same manager. Moreover, even when raters did use the same information, they did not agree on the effectiveness level of the performance displayed. Perhaps part of the reason for low interrater agreement in ratings is due to raters attending to different information and evaluating the effectiveness of manager behaviors differently. This suggests that rater training may be more successful if raters are taught to differentiate more between relevant and irrelevant job behaviors and if they adopt similar schemes for interpreting the effectiveness level of relevant job behaviors.

Regarding the second question, several interesting patterns of rating emerged. Large individual differences were found in the amount of information utilized (judgment frequency), but raters tended to use about the same amount across managers. That is, raters differ widely in how much information they use, but this amount is consistent across rating situations. Raters also tended to make judgments within a limited range of the scale, and the range became smaller with practice. This latter finding may indicate that either raters become more discriminating with practice and therefore become better raters or they become more narrow-minded and more error-prone. It would be important to determine, then, whether differentiation moves in a more valid direction.

The third question yielded information about the role of cognitive complexity in performance rating. It was found that cognitively complex raters (ones who process information along many dimensions of meaning) tended to discriminate more within and between managers than less complex raters. Surprisingly, the amount of information utilized was not related to cognitive complexity. Therefore, it seems that cognitive complexity serves a differentiation function in performance rating.

Because of the experimental design of this study, the relationship between rating behaviors and rating outcomes (e.g. accuracy) could not be determined. A subsequent study permitted these relationships to be tested.

A Study of the Linkages Between Cognitive Abilities, Rating Behaviors, and Rating Outcomes

Several revisions in experimental design allowed rating outcome scores to be calculated for each rater. In this study, all raters viewed and rated six managers across all performance dimensions. That is, each manager was rated along each of six dimensions. To achieve this, raters attended six (rather than one) rating sessions. Each session consisted of viewing and rating each manager along one of the six dimensions, but a different dimension per manager. In each subsequent session, a manager was rated along a dimension different from previous sessions such that by the sixth session, each manager was rated along each of the six dimensions. This yielded 36 ratings per rater (6 managers x 6 dimensions) with which to determine accuracy and various error scores.

At this time, 56 raters have completed all six sessions. The sample of 56 consisted of three subsamples: 20 students, 12 managers, and 24 "expert" managers.[1] The rating behavior of subjects within each of these subsamples will be compared. Student responses will be compared to manager responses to determine the generalizability of findings using college students to those using inexperienced managers. Also, "expert" managers will be compared to the two other samples to determine whether experts exhibit different rating behaviors than nonexperts. Basically, the purpose is to identify what experts do during appraisal that makes them better raters. Whatever rating behaviors distinguish between experts and the two other samples will then be examined in terms of their relationship to accuracy. It is expected that experts' rating behaviors should be related to accuracy and inversely related to rating error.

Analyses of these relationships are currently underway. A brief look at these data suggest that many of the previous findings mentioned earlier in this paper were replicated. It is believed that these analyses will generate new information about the underlying structure of the rating process which may lead to insight into the problem of increasing rater accuracy.

Summary

The need for analyzing the rating process at a microscopic level is clear. The IRJ technique seems to hold great promise for measuring several aspects of the rating process validity. New findings regarding the rating process derived from studies using IRJ have demonstrated the efficacy of the process approach to studying performance appraisal and the usefulness of this technique. Future research along these lines will undoubtedly forward the current "state of the art."

---

[1] Expert managers were identified through extensive interviews with the author and her staff. Managers were considered expert if they had conducted appraisals for a number of years, evaluated several employees, were aware of common pitfalls in performance ratings and who consequently took action to overcome these pitfalls in their own appraisals. Basically those managers who had extensive rating experience and demonstrated "textbook" knowledge of performance appraisal techniques and issues were considered expert. Managers who did not fit this description were included in the unselected manager sample.

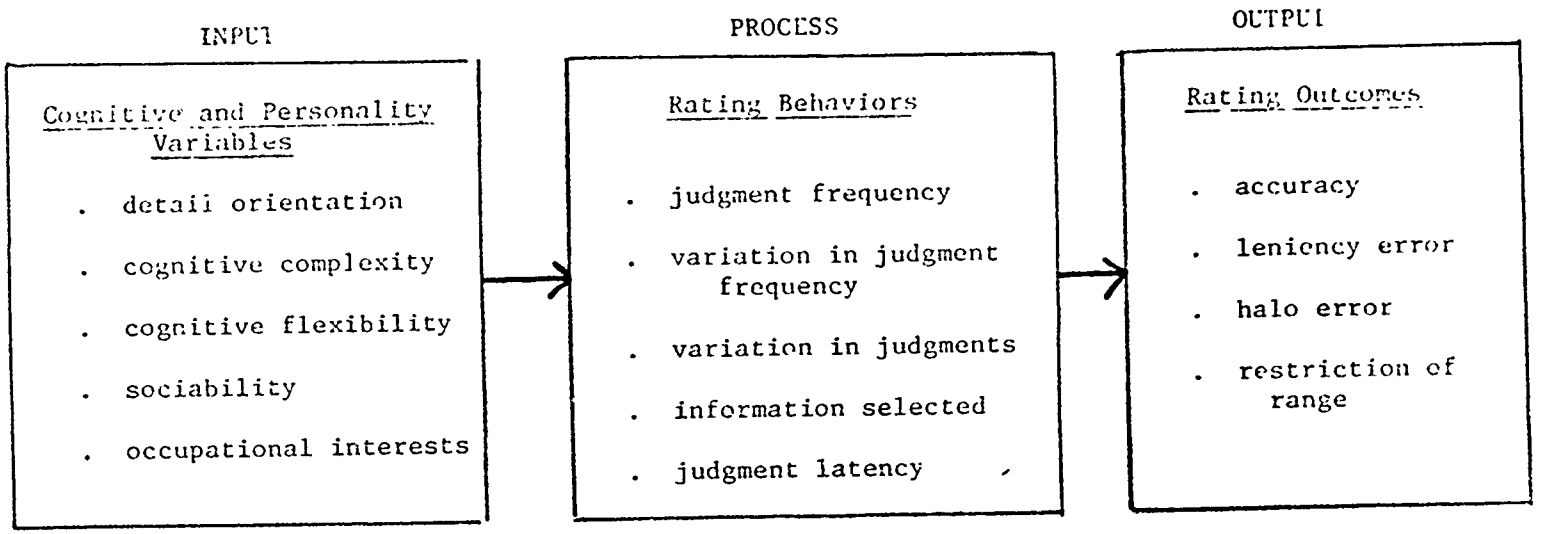


Figure 1. Input, process, and output measures to be intercorrelated to identify the determinants of rater accuracy and rating error.

References

Banks, C. G. A Laboratory Study of the Decision-Making Processes in Performance Evaluation. Unpublished doctoral dissertation, University of Minnesota, 1979.

Bernardin, H. J., and Pence, E. D. Effects of rater training: Creating new response sets and decreasing error. Journal of Applied Psychology, 1980, 65, 60-66.

Bernardin, H. J., and Walters, C. S. Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 1977, 62, 64-69.

Borman, W. C. Effects of instructions to avoid halo error on reliability and validity of performance ratings. Journal of Applied Psychology, 1975, 60, 556-560.

Borman, W. C. Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 1979, 64, 410-421.

Cummings, L. L. and Schwab, D. P. Performance in Organizations: Determinants and Appraisal. Glenview, Illinois: Scott, Foresman and Co., 1973.

Dawes, R. M., and Corrigan, B. Linear models in decision making. Psychological Bulletin, 1974, 81, 95-106.

Dunnette, M. D., and Borman, W. C. Personnel selection and classification systems. In M. Rosenzweig & L. Porter (Eds.) Annual Review of Psychology, 1979, 30, 477-525.

Ericsson, K. A., and Simon, H. A. Verbal reports as data. Psychological Review, 1980, 87, 215-251.

Feldman, J. M. Beyond attribution theory: cognitive processes in performance appraisal. Journal of Applied Psychology, 1980, 65, 550-556.

Hoffman, P. J. The paramorphic representation of clinical judgment. Psychological Bulletin, 1960, 57, 116-131.

Latham, G. P., Wexley, K. N., and Purcell, E. D. Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 1975, 60, 550-555.

Landy, F. J., and Farr, J. L. Performance ratings. Psychological Bulletin, 1980, 87, 72-107.

Nisbett, R. E., and Wilson, D. D. Telling more than we know: verbal reports on mental processes. Psychological Review, 1977, 84, 231-259.

Smith, P. C., and Kendall, L. M. Retranslation of expectations: an approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.

Yntema, D. B., and Torgerson, W. S. Man-computer cooperation in decisions requiring common sense. IRE Transactions of the Professional Group on Human Factors in Electronics, 1961, HFE-2(1), 20-26.

BEST, Paul; McFann, Gray and Associates, Inc., Monterey, California.

A MULTI-ATTRIBUTE UTILITY MEASUREMENT APPROACH TO ORGANIZATIONAL EVALUATION (Thu P.M.)

An evaluation methodology was developed for analyzing the performance of military units employing various experimental management structures.

Evaluation dimensions were established by interviewing commanders and by reviewing Army inspection procedures. A multi-attribute utility measurement approach was described by Edwards, Guttentag and Snapper (1975) was then used to have commanders establish the relative importance of the evaluation dimensions. Next, unit performance on each dimension was rated by the commanders and others in the command hierarchy. Performance ratings and importance weights were then used to establish overall unit performance scores and unit profiles.

The unit profiles were organized to reflect unit performance on each of 18 dimensions covering four areas: training, personnel, management and job performance. The performance scores provide an overall unit performance index which can be used for gross comparisons among units.

The benefits of this organizational evaluation technique for research use and management information will be covered. Future development and applications of the techniques will be discussed.

A MULTI-ATTRIBUTE UTILITY MEASUREMENT APPROACH TO ORGANIZATIONAL EVALUATION

Paul R. Best, Ph.D. Nancy A. Euske, J.T. Harden
and Howard T. Tokunaga

McFann-Gray & Associates, Inc.
2100 Garden Road, Suite J
Monterey, CA 93940

and
Jack Hiller, Ph.D.

U.S. Army Research Institute for the Behavioral
and Social Sciences
Presidio of Monterey Field Unit
P.O. Box 5787
Presidio of Monterey, CA 93940

INTRODUCTION

The long range goal of a major Army Research Institute project is to design operational management systems which will enable infantry commanders to reduce the time spent performing garrison/administrative tasks, thereby increasing the time available for readiness training. This project entails several coordinated research efforts from which new management approaches will be developed and evaluated. First, the missions and activities required of infantry companies by regulation and standard operating procedures (SOPs) were catalogued (Giesler, Harden, Best and Elliott, 1979). Second, this compendium of required missions and activities was used to develop questionnaires, structured interviews and observation guides which provided documentation of the actual missions and tasks performed by commanders and other key personnel. A discussion of these missions and activities is presented in McCluskey, Scott, Tokunaga, Giesler, and Whitmarsh, 1980).

This paper describes the development of a methodology which will contribute to the evaluation of the performance of units employing the experimental management systems. The evaluation methodology and procedures are presented along with actual examples from Army unit administration. In addition, future developments and application of the methodology are discussed.

EVALUATION ISSUES

Two key issues must be resolved before the most appropriate evaluation approach can be selected. These issues are:

1) evaluation focus--should the evaluation focus on the organizational output and its relationship to organizational goals, or should it focus on the organizational processes and their relationship to goals and performance?

2) effectiveness dimension--once the focus is selected what are the dimensions (i.e. performance areas, outputs, goals etc.) on which the organization should be evaluated?

Evaluation Focus

Organizations can be evaluated using their own goals as evaluation standards and comparing performance to the goals. Various organizational theorists, Cyert and March (1963), Etzioni (1964), Proce (1972), Simon (1964), and Steers (1977) have advocated this goal approach to studying organizational effectiveness. Examples in Army units include comparison of re-enlistment rates to unit re-enlistment goals and performance on readiness tests. Alternatively, organizations can be evaluated by determining if their activities and procedures (i.e., implementation processes) are in agreement with formally stated rules. This process orientation also has its supporters including, among others, Beckhard (1969), Likert (1961), and Taylor and Bowers (1972).

The organizational system under study and the research project itself demand that, as a minimum, the goal approach be used as part of the evaluation methodology. Military leaders determine the goals for which unit commanders are held responsible and an evaluation system must take these goals into account. Process evaluation does provide information on the mechanisms used by the organization to reach its goals, however, it does not specify the extent to which the goals are actually accomplished, and is, therefore, by itself inadequate. In addition, the goal approach is necessary to accomplish two important research objectives: 1) determination of the amount of time saved in the conduct of garrison/administrative tasks due to the experimental management systems (and any concomitant increase in time devoted to training); and 2) changes in overall organizational effectiveness as a result of the experimental management systems. Process evaluation will also be included as a part of the research project, but will not be addressed in this paper.

In their discussion of the goal oriented approach, Yuchtman and Seashore (1967) distinguish between prescribed and derived (or functional) goals. In the prescribed goal approach, the organization is evaluated by comparing performance with the organization's stated goals. The main problem with this approach is that various components of the organization are often not aware of the organizational goals or the work in a given unit is not directly related to attaining these goals. In the derived goal approach, the goal(s) is determined empirically within the organization and evaluation consists of measuring the level of goal attainment. Since military commanders have some autonomy in accomplishing their mission it is important to look at both the prescribed and derived goals. These goals may be the same for all units within a single large organization (e.g., Army Combat Brigade), but may receive different emphasis at different levels of command.

Effectiveness Dimensions

The selection of the effectiveness dimensions is based on several considerations: importance of each dimension to the evaluation of research

objectives, importance and credibility of each dimension to research leaders, and the ability to derive valid measures for each dimension. In addition, available project resources constrain the number and types of dimensions selected. The importance and credibility of the dimensions to military leaders is a primary focus of this paper. Inclusion of a controversial dimension or the exclusion of a commonly accepted dimension, however inappropriate, could damage the credibility of the research results. Credibility of the research results is very important to the eventual acceptance of any successful management approaches developed and tested during the project. These considerations led us to determine from a sample of military leaders all of the commonly used evaluation dimensions and to allow for the insertion of additional dimensions as they surface.

MULTI-ATTRIBUTE UTILITY METHODOLOGY

The plan for evaluating the experimental management systems is designed to provide decision makers with information as to what each system accomplishes and at what cost. Since the priorities and values of decision-makers as well as military leaders vary markedly, it is necessary to provide data which are comprehensive and can be presented and used in a flexible manner to meet the needs of such leaders. The multi-attribute utility approach to evaluation research (Edwards, Guttentag and Snapper, 1975) specifically provides a method to incorporate value judgments by the leaders or policy-makers in the evaluation process. The first step in the approach is to discover the important evaluation dimensions (goals) from the leaders. Next, judgments are made by them to determine the relative importance of each dimension by assigning weights. Then, unit performance is estimated (i.e., measured) on each dimension. Finally, the relative importance weights and performance indicators are combined in a weighted linear model. The result is an overall performanc evaluation score. Where it has been determined empirically that individual leaders or echelons of leadership significantly differ in the key dimensions they select or in the the importance they attach to the dimensions, then individually tailored linear models may be employed.

The use of military commanders and staff officers to determine the importance of evaluation dimensions is key to the success of this research project. On-going research and development projects notwithstanding, training and subsidiary task accomplishment are still based on judgments by commanders who have years of experience observing military unit operations, management, and training. The Army, indeed the military system, is designed to obtain maximum benefit from this experience base. Officers command platoons, then companies, serve on battalion staff, and then command battalions and so on, at each step gaining insight into what approaches work and do not work to enable units to accomplish all assigned missions and tasks effectively.

PROCEDURES

Selection of Dimensions

The initial set of evaluation dimensions was established by interviewing battalion and company commanders, personnel officers and operations

officers from infantry and artillery units of an Army Infantry division. These interviews were conducted to determine dimensions of unit effectiveness. Eighteen dimensions were identified, and then organized into four categories: training, personnel, management of equipment and facilities, and job performance/management. The individual dimensions are listed in Figure 1.

Dimension Weights

Each of the 18 dimension names was placed on an individual card. The deck of 18 cards was then sorted by the respondents into their perceived rank order of importance. Once the cards were in order, the respondent was asked to weight each dimension by arbitrarily assigning a "10" to the lowest ranking dimension and assigning higher values to higher ranked dimensions so they would bear a ratio relationship to each other. For example a dimension receiving a "20" would be subjectively twice as important as the lowest dimension. Tied scores were allowed. The dimension weights were rescaled to establish comparable values across respondents. Each assigned dimension score was divided by the total of the assigned scores resulting in a 0-100 point scale for each dimension. (This method is described by Edwards et al., 1975).

Performance Ratings

A scale was placed on the back of each card for rating unit performance on the dimension named on the front of the card. The scale ranged from 0 to 1,000, where 0 represented the worst unit performance imaginable by the respondent and 1,000 represented an ideal but attainable level of performance. The mid-score of 500 represented average performance as witnessed during the respondent's Army experience.

Unit Performance Score

The combined dimension weights and performance ratings allow the derivation of a performance score for each unit. The performance score is the sum of the product of each dimension weight and associated dimension performance rating.

EXAMPLES

The results of applying the evaluation methodology in one battalion are presented in this section. The evaluation procedures were used in nine battalions, however, the results of one are sufficient for describing the methodology. The officers who participated in this exercise expressed no difficulty in rank ordering the cards or assigning the relative weights. In fact, many of them found the procedure intriguing. The dimension importance weights for this battalion are shown in Table 1. These weights illustrate some differences among the respondents. For example, several of the officers rated "performance of physical readiness training" as the most important dimension while the battalion commander rated it lower in overall importance. Also, "individual training" received a lower importance rating from several of the officers than from the battalion commander.

Once the respondents ordered and weighted the dimensions, they rated unit performance on each dimension on the back of each card. The battalion commander and staff officers rated each company's performance. Company commanders rated only their own company's performance. A sample of performance ratings produced by a battalion commander and a company commander rating the same company is illustrated in Figure 2.

The performance ratings are purely subjective evaluations of the company on each dimension. However, future development of this methodology will include use of objective measures where they exist. For example, actual AWOL and re-enlistment rates and SQT results can be obtained and included in the evaluation results. Some dimensions such as ARTEP and IG performance levels, while less objective, can perhaps be related to actual results. On the other hand, some dimensions which are routinely evaluated are subjective by nature and will remain so.

The dimension weights and performance ratings for each respondent were combined into an overall unit performance score. Each performance rating was multiplied by its dimension weight and the resulting products were then summed. The set of scores for the sample battalion officers is shown in Table 2. This set of results shows that the combat support company (CSC) received the highest total performance score based on all individual ratings. However, examination of the individual dimension performance ratings is necessary to determine how this company achieved its high total score.

DISCUSSION

The multi-attribute utility evaluation methodology provides the capability of evaluating organizations from several perspectives. Decision makers come from different organizational components and necessarily represent different viewpoints. The multi-attribute utility approach can take each perspective into account.

An additional advantage of this approach is that assigned weights and rank orders among the dimensions produced by members of an organizational unit may also be used to provide a measure of communication effectiveness within the unit. During the conduct of this effort company commanders expressed, in addition to their own priorities, their perceptions of the battalion commanders' priorities. The comparison of the "expected" battalion commanders priorities with his actual priorities served as a very useful exercise in unit communication. Organizational functioning can usually be improved when everyone understands the supervisor's priorities. The supervisor can also dramatically influence organizational behavior by clearly communicating priorities in terms of goals.

Multi-attribute utility evaluation methodology appears to be extremely promising not only for our project research needs, but also for internal organizational diagnosis and management information.

TRAINING

- ARTEP performance
- Collective training
- SQT performance
- Individual Training
- Performance of Physical Readiness Training (PRT)

PERSONNEL

- Unfavorable personnel actions (Article 15s, administrative discharges)
- Appearance of personnel
- Personnel utilization (assignments)
- AWOL rates
- Re-enlistment rates

MANAGEMENT OF EQUIPMENT AND FACILITIES

- Maintenance
- Security of weapons and documents
- Appearance of unit area
- Supply management
- Result of IG

JOB PERFORMANCE/MANAGEMENT

- Communication (administrative through chain of command)
- Execution of SOPs
- Accomplishment of assigned tasks

FIGURE 1. Evaluation Dimensions Selected

EVALUATION DIMENSIONS	BN CMDR	STAFF OFFICERS				COMPANY COMMANDERS*				
ASSIGNED TASKS	10	3	4	1	8	5	4	1	15	6
SECURITY	10	6	2	3	2	2	10	22	1	11
INDIVIDUAL TRAINING	9	2	8	7	4	1	6	1	0	4
RE-ENLISTMENT RATE	9	13	28	10	14	11	4	3	7	6
MAINTENANCE	8	5	11	8	3	6	6	12	3	11
PERFORMANCE PRT*	7	23	4	10	14	19	15	13	33	6
COMMUNICATION	6	3	2	2	2	2	1	1	1	1
APPEARANCE OF UNIT AREA	6	8	2	9	10	16	10	8	6	11
APPEARANCE OF PERSONNEL	6	7	2	9	9	13	6	4	4	5
SQT PERFORMANCE	6	3	3	7	3	1	5	1	0	4
EXECUTION OF SOP's	5	1	1	1	1	1	1	1	0	1
RESULT OF IG	4	11	8	4	9	8	10	11	21	11
COLLECTIVE TRAINING	3	2	3	7	5	1	5	1	0	4
ARTEP PERFORMANCE	3	1	1	1	1	1	9	1	1	2
AWOL RATE	2	3	1	10	6	5	1	2	4	4
SUPPLY MANAGEMENT	2	4	17	5	2	3	5	16	2	11
PERSONNEL UTILIZATION	2	2	2	2	1	1	1	1	1	1
UNFAVORABLE PERSONNEL	2	3	1	4	6	4	1	1	1	1
	100	100	100	100	100	100	100	100	100	100

* Physical Readiness Training

TABLE 1: Dimension Importance Weights for a Sample Battalion.

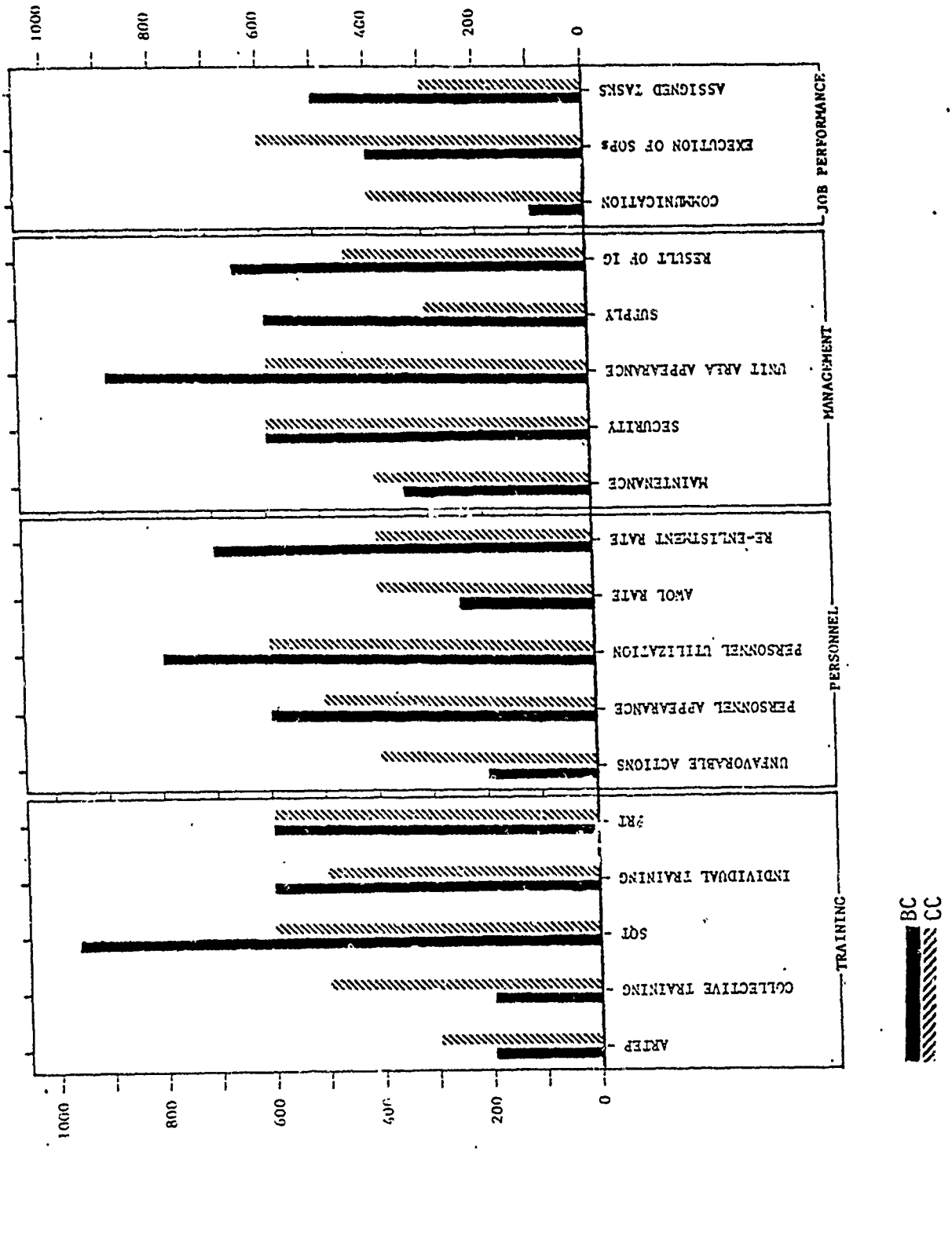


FIGURE 2: Company Performance Rating
Example for a Battalion Commander (BC)
Company Commander (CC)

	HHC	A CO.	B CO.	C CO.	CSC
BN COMMANDER	446	481	512	469	641
XO	494	600	706	664	909
S-1	487	523	531	446	600
S-3	384	468	476	480	575
S-4	359	407	436	389	477
CO COMMANDER	729	689	614	589	839

TABLE 2: Performance Scores

Beckard, R., *Organizational Development: Strategies and Models.* Reading, Mass.: Addison-Wesley, 1969.

Cyert, R.M. and March, J.G., *A Behavioral Theory of the Firm.* Englewood Cliffs, NJ: Prentice Hall, 1963.

Edwards, W., Guttentag, M. and Snapper, K., A Decision-Theoretic Approach to Evaluation Research. In Struening and Guttentag (eds.) *Handbook of Evaluation Research.* Sage Publications. Beverly Hills: 1975.

Etzioni, A., *Modern Organizations.* Englewood Cliffs, NJ: Prentice Hall, 1964.

Giesler, R.W., Harden, J.T., Best, P.R., and Elliott, M.P., *Missions, Responsibilities, Duties and Tasks of Infantry Companies and Field Artillery Batteries.* ARI Technical Report, September 1979.

Likert, R., *New Patterns of Management.* New York: McGraw-Hill, 1961.

McCluskey, M.R., Scott, A.C., Tokunaga, H., Giesler, R.W., and Whitmarsh, P.J., *Actual Missions, Activities and Job Tasks in Companies and Batteries.* ARI Technical Report, January 1980.

Price, J.L., The Study of Organizational Effectiveness. *The Sociological Quarterly.* 1972.

Simon, H.A., On the Concept of Organizational Goals. *Administrative Science Quarterly.* 1964.

Steers, R.M., *Organizational Effectiveness: A Behavioral Review.* Santa Monica, CA: Goodyear Publishing Co., Inc., 1977.

Taylor, J.C. and Bowers, D.G., *Survey of Organizations.* Ann Arbor, MI: Institute for Social Research, University of Michigan, 1972.

Yuchtman, E. and Seashore, S.E., A System Resource Approach to Organizational Effectiveness. *American Sociological Review.* 1967.

THE MARINE CORPS JOB SATISFACTION PROGRAM

COLONEL N. K. BODNAR
Director

Headquarters United States Marine Corps
Office of Manpower Utilization
Quantico, Virginia 22134

ACKNOWLEDGEMENTS

The theory and its general application employed in this work is directly derived from the scholarly research of J. Richard Hackman, PhD, Yale University; Greg Oldham, PhD, University of Illinois and their associates. Their efforts were conducted under the sponsorship of Office of Naval Research (ONR) Contract Number N00014-67A-0097-0026. The Office of Manpower Utilization made only minor adaptive changes to the methodology of these academicians. We are deeply appreciative of the cooperation and advice we received from them as well as the Office of Naval Research.

The Marine Corps Job Satisfaction Program

Introduction

One aspect of the Occupational Analysis Program conducted by Headquarters United States Marine Corps is a Job Satisfaction Survey. The survey normally is administered in person to Marines of all grades within the given study sample (occupational field or groups of related military occupational specialties).

Anonymous incumbent responses are collected and survey results are tabulated and analyzed. A master tape file of job satisfaction results for the Marine Corps population surveyed to date allows for benchmark comparisons on a sample by sample basis. Perceptions of work problems are identified, graphically presented, and addressed. A number of implementing concepts (action steps) are then related to the problem areas identified, providing a useful and well received set of management tools which can produce marked improvements in desired work outcomes.

The Theory Explained

Some explanation of the theory behind the survey is necessary to understand the analysis methodology. A number of central theory issues are addressed through utilization of a 77 question survey which is "mapped down" into 24 "combined factors". Survey responses range on a seven point scale with specific survey questions "keyed" to the 24 combined factors. For example, the average rating for the factor Task Identity is an average of the ratings for the following:

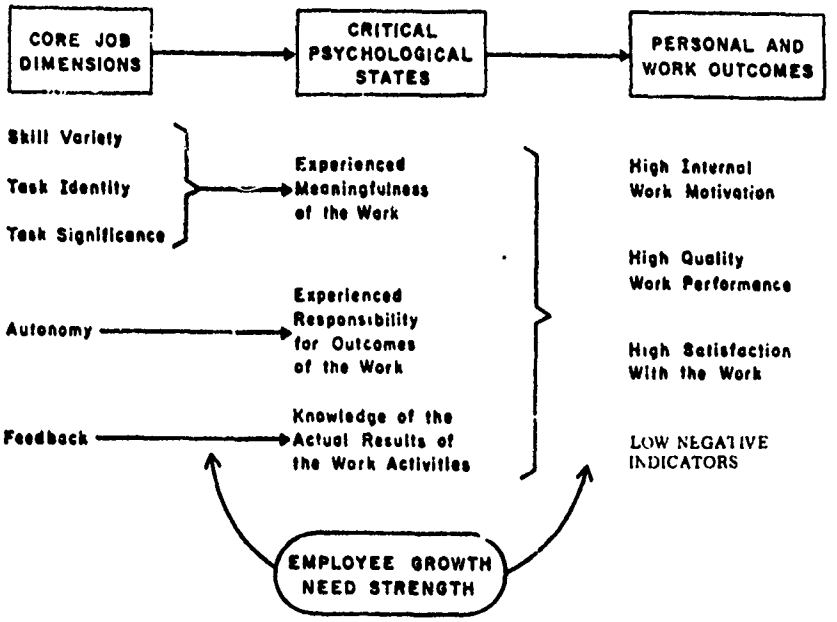
1. To what extent does your job involve doing a "whole" or "complete" piece of work?

2. The job is arranged so that I do not have the chance to do an entire piece of work from beginning to end (reverse score).

3. The job provides me the chance to completely finish the pieces of work I begin.

Response averages for all survey questions and for the combined factors are then tabulated for the entire sample and for each grade in the sample.

Figure 1 reflects the theory, which has been extensively validated.



The Relationships Among the Core Job Dimensions, The Critical Psychological States, and On-the-job Outcomes.

Figure 1

As Figure 1 illustrates, three critical psychological factors must be present if high personal and work outcomes are to be attained.

1. The work/effort must be meaningful to the individual.
2. Individuals must clearly be held responsible for their work.
3. Knowledge of results must be available to the individuals performing the work. It is best if this knowledge comes directly from the work itself rather than through pats on the back or paper reports.

Since the psychological states are a combination of emotions, feelings, and thoughts, they are extremely difficult to measure. There are in every job, however, certain core job dimensions which relate to the psychological states, and these can be accurately measured. For example, three of these relate to "meaningful work" (Figure 1):

1. Skill Variety - Performing activities that challenge skills and abilities; reduce monotony; enhance job appeal.
2. Task Identity - Arranging work into whole "modules" with an identifiable beginning and end, and a visible outcome.
3. Task Significance - Degree to which work has an important impact on the lives of individuals, other people, the unit, and the mission.

It is apparent that the greater the degree to which these factors present themselves to the worker, the more meaningful the work.

The other two core job dimensions are Autonomy, the amount of personal freedom and discretion given to individuals in scheduling and carrying out work; and then Feedback, the information received about the effectiveness of the individual work effort.

In sum, the five measurable core dimensions promote the psychological states, in turn producing desired work outcomes. The five core dimensions relate to the job as a single summary index called the Motivating Potential Score (MPS) of the job (Figure 2).

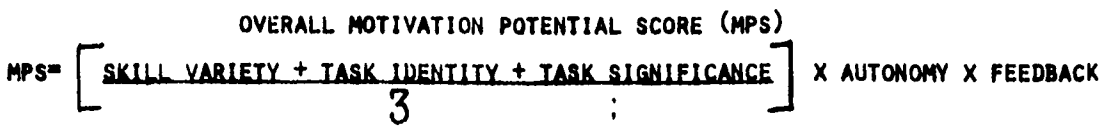


Figure 2

The MPS value allows for a macro look at the job. Specific strong and weak areas can later be identified through additional analysis steps. The MPS simply reflects how motivated the group is. It is particularly useful to make comparisons of the MPS of the given sample with the MPS of the Marine Corps population surveyed to date.

A final factor called Growth Need Strength (GNS) answers the question "Is the theory equally applicable to all?"

The GNS measure is an average of the ratings for the following survey items:

1. Stimulating and challenging work.
2. Chances to exercise independent thought and action in my job.
3. Opportunities to learn new things from my work.
4. Opportunities to be creative and imaginative in my work.
5. Opportunities for personal growth and development in my job.
6. A sense of worthwhile accomplishment in my work.

It allows managers to prioritize groups in terms of which they really desire to grow on the job and which are relatively happy where they are. Time, money, and other resources can then be put into enrichment efforts for those groups desiring further growth.

Theory To Practice

In employing the theory we identify what aspects of the job need attention, and then provide some implementing concepts aimed at improving problem areas.

(51)

The 77 question survey guages the following:

1. The current levels of General Satisfaction, Internal Work Motivation, and performance of job incumbents.
2. The motivating potential (MPS) of the job, the score on each core dimension, plus additional environmental and motivational aspects of the job.
3. The level of Growth Need Strength (GNS).

The analysis actually revolves around four sequential questions which reflect strengths and weaknesses of the job as perceived by job incumbents.

1. Are motivation and satisfaction central to the problem?
2. Is the job low in motivating potential? Compare the sample MPS with established norms.
3. What specific aspects, if any, are causing some difficulty or concern? A plot of the mean scores for the core dimensions is useful here (Figure 3).

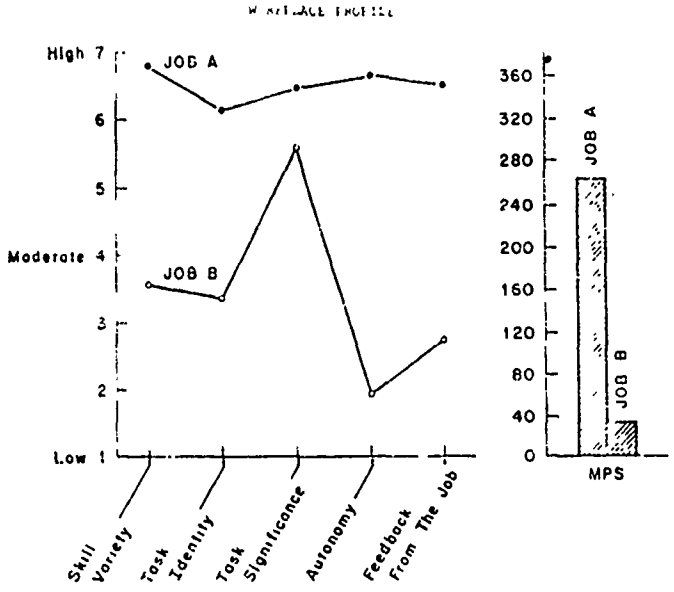


Figure 3

4. How ready for change are the Marines involved? Compare the sample GNS to established norms.

Once the questions are answered, some implementing concepts are recommended, if needed, for improving specific core job dimensions. Figure 4 reflects the relationships between the core dimensions and their associated implementing concepts, each of which is briefly described.

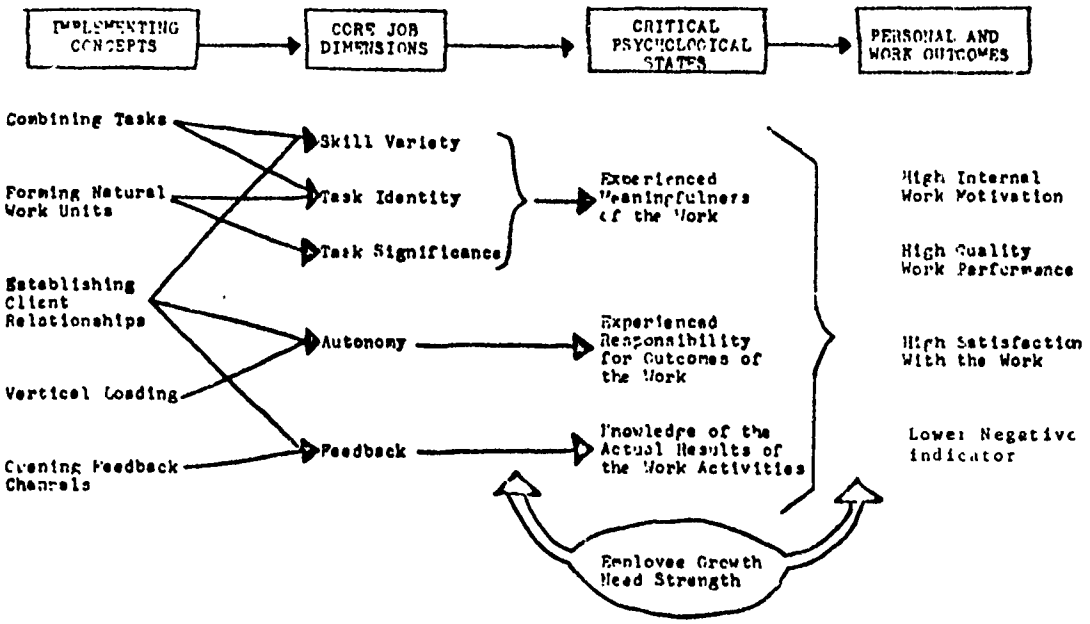


Figure 4

The Full Model: How Use of the Implementing Concepts Can Lead to Positive Outcomes

1. Combine Tasks. This increases Skill Variety, Task Identity, and eliminates fractionalizing jobs. Enlarging the work module or forming teams may be useful.

2. Form Natural Work Units. This gives incumbents a feeling of "ownership". Organize the work into meaningful modules, clustering tasks so a person can feel responsible for an identifiable body of work. This increases Task Identity and Task Significance.

3. Establish a Client Relationship With a Customer. When incumbents have a sense of personal responsibility for managing a customer relationship, Feedback, Skill Variety, and certainly Autonomy will increase.

4. Vertical Loading. This, in short, is pulling down responsibilities from above, pushing some down to lower echelons, and incorporating some pre and post planning activities (Figure 5). Vertical loading increases Autonomy.

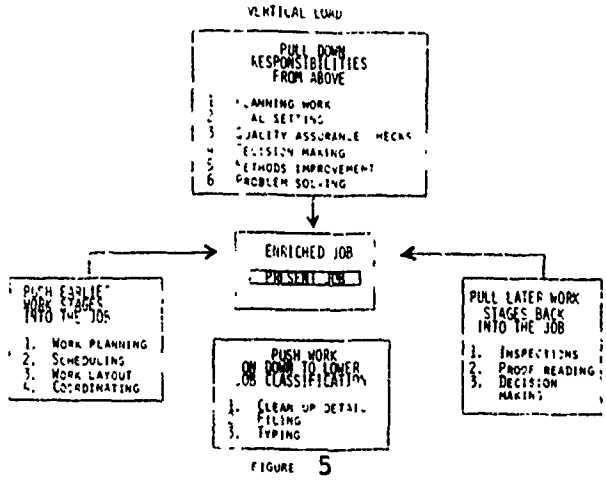


FIGURE 5

5. Opening Feedback Channels. Reports should go down as well as up when possible, and feedback should be as close to the action as possible in terms of time and place.

A Review of Job Satisfaction Survey Results For The Marine Corps Population Surveyed To Date

We will review the Marine Corps job satisfaction effort by looking at results for over 12,000 incumbents surveyed to date. We can accomplish this on a macro basis by going through the four step analysis methodology. This is the procedure normally utilized in analyzing each study sample. Additionally, in analyzing each study sample the population surveyed to date (from the master tape file) is considered the established norm for comparisons, and a more in depth analysis is made by reviewing the actual survey responses associated with the 24 combined factors for the sample and across the grades in the sample.

The following are the combined factor averages for the Marine Corps population surveyed to date:

Combined Factors

Factor		Population Average
1.	Skill Variety	3.95
2.	Task Identity	4.63
3.	Task Significance	5.35
4.	Autonomy	4.47
5.	Feedback From Job Itself	4.71
6.	Feedback From Agents	3.98
7.	Dealing With Others	5.43
8.	General Satisfaction	4.25
9.	Internal Work Motivation	5.36

Factor		Population Average
10.	Admin Policies, Rules, Regulations	3.95
11.	Supervision	4.34
12.	Interpersonal Relations	5.02
13.	Working Conditions	3.58
14.	Salary	3.28
15.	Status	4.06
16.	Security	4.32
17.	Achievement	4.53
18.	Recognition For Achievement	4.40
19.	Work Itself	4.66
20.	Responsibility	5.27
21.	Advancement	4.59
22.	Growth	4.31
23.	Individual Growth Need Strength	5.61
24.	Motivating Potential Score	97.93

Let us now go through the four step analysis by looking at the population in general.1

Step 1. Are motivation and satisfaction a problem? Factors 8 through 22 reflect responses to this question. Factors 8 and 9 will be described later. Factors 10 through 22 are normally plotted for the given study sample as a whole and by grade for the sample. They are plotted for the population surveyed to date in Figure 6:

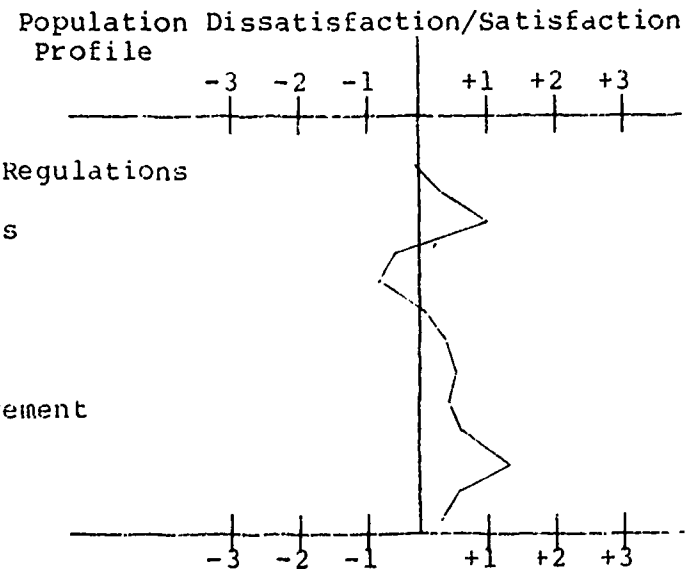


Figure 6

Factors 8 and 9 are indicators of an overall perception of the workplace as viewed by Marines and an indication of their desire to be high producers in their work. Factors 10 through 16 are environmental factors, i.e. issues external to the actual work. Factors 17 through 22 are job related. They are therefore motivational in nature.

General Satisfaction (Factor 8) and Internal Work Motivation (Factor 9) are above the 4.0 median for the population, and the plots indicate that Admin Policies, Rules, Regulations (Factor 10), Working Conditions (Factor 13), and Salary (Factor 14) are below the 4.0 median for the population. Normally, sample averages for the environmental and job related factors are compared with the population norm. The population itself appears to be reasonably motivated though expressing slight concern over the three environmental factors noted previously.

Step 2. Is the job itself a problem? Here the MPS for the given sample is examined and compared against the population MPS. The population MPS to date is 97.93, and normally remains close to 100. The average MPS across numerous occupations surveyed by Hackman and Oldham (1974) was found to be 125.

Step 3. What aspects of the work, if any, are causing the difficulty? Here it is useful to plot the core dimensions for the sample as a whole and for each grade in the sample. These are also compared to the population core dimensions. The population core dimensions are plotted as follows (Figure 7):

Population Core Job Dimensions

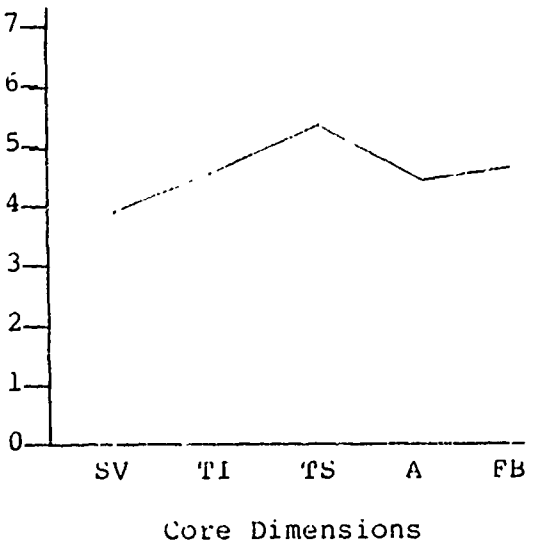


Figure 7

Figure 7 indicates that Skill Variety is only slightly below the 4.0 median and all other core dimensions are above 4.0. This may indicate some slight concern with Skill Variety.

Step 4. How ready are personnel for change? The Growth Need Strength measure (Factor 23) answers this question. It allows managers to set priorities in terms of which groups within the given study sample would respond most favorably to a job enrichment effort. Since groups with higher GNS scores would be expected to respond more favorably to a job enrichment program, managers can "key in" on these groups with their available resources.

The population GNS of 5.61 is fairly high on the seven point scale, and indicates that the incumbents in general desire further growth. Normally, the sample GNS is compared with the population norm.

Summary

The Job Satisfaction Survey provides a valuable instrument for taking a "thermometer reading" of incumbent perceptions about their jobs. A simple analysis methodology provides a vehicle for addressing job related and environmental problems, and the implementing concepts provide a set of "management tools" for job improvement.

While there is some fluctuation from study to study in terms of the various problem areas identified, data from the Marine Corps population surveyed to date provide a useful norm for comparison, and indicate that Marine Corps incumbents on the whole are a fairly motivated group desiring further growth and achievement on the job.

Additional endeavors in the area of job satisfaction will certainly be a vital part of the overall occupational analysis effort of Headquarters Marine Corps. While actual feedback from the commands themselves in terms of the viability of the program has been somewhat limited, this is not discouraging since routine reporting of job satisfaction results to the field has been ongoing only for about a year. In that sense the effort is still in its early stages, though extensive research prior to that time was conducted to formulate a Marine Corps position on job satisfaction.

Job Satisfaction Theory has been introduced to some extent through the leadership training offered with the professional schooling of Marine officers such as the Amphibious Warfare School at Quantico. In addition, the future job satisfaction effort will be expanded in terms of additional correlations of demographic information of study samples with the actual job satisfaction data, and in taking new approaches toward the subject to better meet needs in the field. One such effort is a job satisfaction survey of a cross section of women Marines which is presently being conducted. Results of that survey should be available in the near future.

In short, job satisfaction research in the Marine Corps, as with the other services, will continue. In an era of decreasing personnel with increasing requirements for technical skills in the military it is imperative that perceived incumbent needs in the workplace be met as much as possible. Though job satisfaction will not of itself totally solve the attrition problem it certainly is a step in the right direction.

---

1 These categories are the same as those employed by Dr. Frederick Herzberg in his Motivation-Hygiene Theory. The categories are used for convenience only. Information derived for this analysis did not employ the standard Herzberg Critical Incident Technique and direct interviews essential to M-H Theory.

BIBLIOGRAPHY

DDC NO = Defense Documentation Center Number

AUTHOR	ARTICLE TITLE	DATE PUBLISHED	DDC NO
Hackman, J. R.	Improving the Quality of Work Life: Work Design	June 1975	PB 255 044
Hackman, J. R. & Morris, C. G.	Group Tasks, Group Interaction Process, and Group Performance Effectiveness: a Review and Proposed Integration	Aug 1974	AD 785 287
Hackman, J. R., Weiss, J. A. & Brousseau, K. R.	Effects of Task Performance Strategies on Group Performance Effectiveness	Oct 1974	AD-A001 707
Hackman, J. R. & Oldham, G. R.	Motivation through the Design of Work: Test of a Theory	Dec 1974	AD-A009 331
Oldham, G. R.	Conditions under which Employees Respond Positively to Enriched Work	Sep 1975	AD-A016 248
Hackman, J. R. & Oldham, G. R	The Job Diagnostic Survey: An Instrument for the Diagnosis of Jobs and the Evaluation of Job Redesign Projects	May 1974	AD-779 828
Frank, L. L. & Hackman, J. R.	A Failure of Job Enrichment: The Case of the Change that Wasn't	Mar 1975	AD-A007 356
Hackman, J. R.	On the Coming Demise of Job Enrichment	Dec 1974	AD-A003 090
Hackman, J. R., Oldham, G. R., Janson, R. & Purdy, K.	A New Strategy for Job Enrichment	May 1974	AD-779 827

BOHRER, Arnold, Belgian Armed Forces Psychological Research Section, Brussels, Belgium.

SOCIAL ANXIETY, SELF-CONFIDENCE AND MILITARY APTITUDE (Thu P.M.)

All officer candidates receive at the end of each academic year of the Koninklijke Militaire School (Royal Military School) a military aptitude score, an estimate of a cadet's future officer's performance. This score has three components (1) a technical score, based on achievement in a number of military courses, (2) a physical score, based on sport-achievement, and, (3) a personality or moral score, based on evaluations made by trainers and staff members. Earlier efforts to improve the quality of the officer's selection were concentrated on personality differences between successful and not-successful reserve-officers. It was found that the selection team judged more self-confident and less social anxious aspirants more able to become good reserve-officers. The same differences appeared between successful and not successful reserve-officers. A cross-validation confirmed these findings. The two scales, Self-Confidence and Social Anxiety, constructed during this research, were applied to applicants of the Royal Military School. It was found that

(1) self-confident subjects obtain better selection scores and better Military aptitude scores than less-confident applicants,

(2) social anxiety is correlated with selection scores. It's effects on the students academic achievement depend on his ability level. Its effects on attrition from the School are related to the kind of studies followed,

(3) cadets, who state that they perform better when they have light feelings of tension (facilitating anxiety) tend to achieve better military aptitude scores than cadets who answer in the opposite way.

---

* Much of the information in this paper was previously reported during the 15th International Conference on applied Military Psychology, May 79.

SOCIAL ANXIETY, SELF-CONFIDENCE AND MILITARY APTITUDE

ARNOLD BÖHRER

Belgian Armed Forces Psychological Research Section

INTRODUCTION

In many military leadership guides it is stressed that a leader has to be competent in interpersonal skills, and has to behave in a self-confident way. Social anxiety inhibits the development of these qualities. An important aspect of social anxiety is the fact that the anxious person is afraid of negative appraisal by others, especially what his interpersonal skills and his physical appearance concern (Willems, 1973). A social anxious person will encounter great difficulties when he has to function as a leader. Officers have to take decisions in social situations, which imply social pressures from superiors, colleagues and subordinates. The quality of their decisions depends to a great extend upon the way they cope with those pressures. An officer with high social anxiety will tend to show a non-adaptive leadership-behavior pattern. It leads to nonassertive or to aggressive behavior (Rathus, 1978). The present study has been undertaken in order to demonstrate the relationship between successful leadership and self-confidence and between unsuccessful military leadership and social anxiety.

METHOD

General outline

Two independent judges (one Frenchspeaking, one Dutchspeaking) were asked to select out of 151 items of a personality questionnaire all the items which were related to social anxiety or to self-confidence. The judges did this on the basis of bipolar definitions of both qualities (cfr infra). Both judges agreed on 30 items as being possible questions for a social anxiety scale (SA) and on 42 items for a self-confidence scale (SC). The response difference between a group of successful reserve officers and a group of unsuccessful reserve officers were analysed for all items. Because only $\pm$ 27% of the SA and SC items did discriminate between the two groups, two new scales were constructed using only the most discriminating items. The validity of these new scales is studied for reserve officers and aspirant regular officers. The relations of the two scales with the academic-achievement of the cadets at the Royal Military School are defined, and finaly correlations between the two scales and other personality inventories are calculated in order to determine their concurrent validity.

Definitions

Social anxiety

Feeling at ease in social situations vs being anxious in social situations, or, enjoying to be with people, expressing himself freely, daring to say no to exagerated demands, daring to make remarks when needed, accepting a leadership-role, etc. vs, remaining in the background, being afraid of striking up a conversation with co-workers, strangers or superiors, being afraid of speaking before a group, fearing to be disapproved by others, being afraid of defending ones rights, avoiding asking questions out of fear of being stupid, etc..

Self-confidence

A self-confident person has the feeling to be an independent, valuable person, who can go his own way and can confirm himself with respect to others. He feels to be equal to his tasks, accepts responsability, behaves in an autonomous way, has a good self-image, dares to take some risks, controls his behavior, shows initiative, etc.. A person with low self-confidence feels to be inferior, feels to be uncertain, depreciates his own work, is quickly discouraged, is quickly upset, is unable to decide easily, does not understand his own behavior, has little self-control, lacks emotional stability, etc..

Successful military leadership

Since the study could only be directed to reserve officers (R.O.) and to aspirant regular officers (Reg.O.), the definition of successful military leadership is a very limited one. It is mainly based on the evaluations of the officers selection board, on the results obtained during the training period (reserve officers) and on the judgements of the staff of the Royal Military School (aspirant regular officers). The selection board tries to define the military leadership aptitude (L. score) of each applicant during a two days lasting selection procedure. The judgement is mainly founded on the behavior of the applicants during 5 leaderless group tasks (*). Each task takes 20 minutes. Groups of 6, 7 or 8 applicants have to construct the framework of a tent, to build a 3 m high double round arch, to draw up a model of a little industrial city (model town test), to design the selling space of a departmentstore (each member of the group being responsable for a department), and, to establish a group rankorder of a number of topics with regard to the importance they have in a civic spirit enhancing education program (each member having already made up a personal rankorder). The behavior of the applicants is observed by four welltrained regular officers. Each of the latter supervising one of the first four tasks, and two of them the last one.

(*) The final L score correlates strongly with the leadership scores based on the 5 group tasks (French R.O. : r=.88, n=967; Dutch R.O. : r=.95, n=873; French Reg.O. : r=.81, n=143; Dutch Reg.O. : r=.76, n=251)

The reserve officers have to follow a military and leadership training, which lasts 5 months. At the end of this training they have to function as an officer in a military unit. The training outcome is considered as a sufficient evidence for their military leadership capacities.

The training period of the aspirant regular officers lasts 4 or 5 years. Success in the Royal Military School depends to a great extend on the academic achievement of the cadets. Although, each year military training is provided, the real professional training is only given at the end of the total academic period. The military aptitude score, calculated at the end of each year, is based on three undistinguishable matters : the results obtained in technical courses (mapreading, armament, etc.) and in physical education are combined with a judgement about the moral and the leadership qualities.

RESULTS

Social anxiety, self-confidence and success as a Reserve Officer (R.O.)

Two groups of applicant reserve officers (R.O.) were constituted in order to analyse the response differences on 151 "Yes-No" items belonging to a personality inventory, derived from the Cattell 16 Personality Factors. This inventory is a part of the selection procedure. This analysis is made twice, once for the Dutchspeaking R.O.'s and once for the Frenchspeaking R.O.'s. The first group contains successful R.O.'s, the second group unsuccessfull ones, they failed to pass either the training or the selection.

The absolute mean difference between these groups was for the Frenchspeaking R.O.'s (F.R.O.) 6,87%, with a Standard Deviation (SD) of 5,48%, and for the Dutchspeaking R.O.'s (D.R.O.) 7,12%, with a SD of 4,99%. An absolute difference of $\pm$ 12 %, about I SD above the absolute mean difference, was accepted as discriminating between the two groups. The discriminating power of the 72 SA or SC items is shown in table I.

TABLE I The discriminating power of the Social anxiety and the Self-confidence items, selected out a personality inventory totalising 151 items

Amount of discrimination	Dutch Reserve Officers (n = 445)			French Reserve Officers (n = 319)		
	SA	SG	Other Items	SA	SG	Other Items
$\geq \lvert\bar{X}\rvert$ + I SD	12	11	3	8	12	2
$\geq \lvert\bar{X}\rvert < \lvert\bar{X}\rvert$ + I SD	6	15	15	11	15	12
$< \lvert\bar{X}\rvert$	12	16	61	11	15	65
Tot.	30	42	79	30	42	79
	χ^2_D = 32,09 p .001			χ^2_F = 36,4 p .001		

Only 27 % of the SA items and 29 % of the SC items exceed the fixed difference (X + I SD) for the F.R.O.'s.
For the D.R.O.'s these percentages are respectively 40 % and 26 %.
These results seem not to support the hypothesis. But, as can be seen in table I, 20 of the 22 discriminative items belong to one of the two scales what the F.R.O.'s concern and for the D.R.O.'s 23 items out of the 26.
These results made the construction of two well discriminating scales possible : a SA scale with 22 items and a SC scale with 27 items for the D.R.O.'s and 31 items for the F.R.O.'s.
In order to control the discriminative power of these two scales 4 groups of D- and F.R.O.'s were constituted:

the SHL group, based on R.O.'s who succeded (S) in their training and who received a High (H) Military Leadership score (L) from the selection board,
the SML group, with R.O.'s who succeeded (S) in their training and who received a Medium (M) L score,
the UML group, with R.O.'s who did not succed (U) in their training and who received a Medium (M) L score,
and finaly the LL group, with applicants who received a too Low (L) L score to be admitted to the training.

This control was done twice, once for the population on which the scales were constructed (P 1)*, once on a new population (P 2)*.
A clear difference between the four subgroups was found for P 1, Dutch and French R.O.'s, for the two scales (Table 2). Unfortunately the comparison between the two ML groups was obscured by a difference in the mean of the L scores in favor of the successful group. For the cross-validation (Table 3) these two groups were matched what the L scores concern in order to obtain two M groups with a same mean and SD for the L scores.
Although, the cross-validation shows approximately the same results as already obtained, it is obvious that the two scales discriminate better between High and Low L groups than between the two S groups and the UML group (Table 3).
The close relationship between the L scores and the SA and SC scores appears clearly in table 4.

* Composition of the two populations (P1 and P2), number of R.O.'s in each subgroup

	Dutch R.O.		French R.O.	
	P1	P2	P1	P2
SHL	124	30	50	40
SML	55	37	35	13
UML	108	53	76	34
LL	158	30	[illegible]	[illegible]
Tot	445	130	[illegible]	[illegible]

TABLE 2. Social anxiety and self-confidence : means of four reserve officers groups (n=50 in each subgroup, taken at random from the P1 subgroups).

Subgroups P1	Social Anxiety (22 items)		Self-confidence	
	Dutch	French	Dutch (27 items)	French (31 items)
Successful, High leadership	17.38	16.84	22.32	24.64
Successful, Medium leadership	15.98	15.20	22.22	21.36
Unsuccessful, Medium leadership	14.48	14.76	20.40	20.52
Low leadership (not accepted applicants)	12.88	11.98	18.35	18.14

The SA scores are inversely calculated for practical reasons; in table 2 and 3 a high score means low social anxiety, a low score high social anxiety.

TABLE 3. Social anxiety and Self-confidence : means of four reserve officers groups (Cross-validation).

Subgroups P2		Social Anxiety (22 items)		Self-confidence	
		Dutch	French	Dutch (27 items)	French (31 items)
Successful, High Leadership (SHL)		17.57 (n=30)	16.30 (n=40)	22.57 (n=30)	23.20 (n=40)
Successful, Medium Leadership (SML) (*)		15.45 (n=20)	15.50 (n=20)	20.75 (n=20)	22.00 (n=20)
Unsuccessful, Medium Leadership (UML) (*)		15.05 (n=20)	14.70 (n=20)	19.50 (n=20)	21.15 (n=20)
Low Leadership (LL)		13.03 (n=30)	13.00 (n=40)	19.00 (n=30)	19.00 (n=40)
t	SHL vs LL; UML; SML	5.26;2.76; 1.93	5.00;2.12; n.s.	4.72;3.33; 2.15	4.20;1.95; n.s.
	SML vs LL; UML	1.89;n.s.	3.36;n.s.	1.95;n.s.	2.70;n.s.
	UML vs LL	1.79	1.79	n.s.	1.73
	SHL+SML vs UML	1.74	1.80	2.63	n.s.

(*) To compose these subgroups, leadership scores were matched in order to obtain two identical subgroups concerning these scores (Dutchspeaking $\bar{X}$=11.1; Frenchspeaking $\bar{X}$=10.35).

Table 4 Correlations (Spearman) between Leadership scores and Social Anxiety and Self-confidence scores.

	Dutch R.O. (P 1) (n = 200)		French R.O. (P 1) (n = 200)	
	SA	SC	SA	SC
Leadership	-.475	.426	-.527	.526
SA		-.447		-.609
	Dutch R.O. (P 2) (n = 130)		French R.O. (P 2) (n = 155)	
	SA	SC	SA	SC
Leadership	-.402	.450	-.389	.328
SA		-.483		-.610

All correlations are stat. significant at the 1% level.

Self-confidence, social anxiety and military aptitude of aspirant Regular Officers (Reg.O.)

The Reg.O. population differs in many aspects from the R.O. population. The latter is more heterogeneous what age and education of the candidates concern. About 65 % of the R.O.'s have a university or college degree, and about 35 % a highschool degree. This latter part of the group stopped studying. Almost all Reg.O.'s have a highschool degree and intend to go on with college education (Royal Military School).
The admittance of the R.O.'s is based on the L score only, these of the Reg.O.'s on two scores, one is calculated on the results of an examination in mathematics and in native language, the other is the L score. The classification of the candidates is made upon the first. For the second a cutting score is fixed. About 15 % of all candidates fall below this score. (*)

As can be seen in table 5, the same relationship between L scores and SA and SC scores exists in this population as in the R.O. population.

Table 5 Correlations (Spearman) between Leadership scores and Social Anxiety and Self-confidence scores.

	Dutch Reg.O. (n = 381)		French Reg.O. (n = 158)	
	SA	SC	SA	SC
Leadership	-.30	.28	-.36	.31
SA		-.40		-.40

All correlations are significant at the 1 % level.

Before examining the relationship between the Military Aptitude score (M.A. score) given at the end of the first year at the Royal Military School and the SA and SC scores, the correlation between the two military leadership evaluations (L score and M.A. score) needs some attention.

(*)Abstraction is made of the medical and physical admittance conditions

This correlation is statisticaly significant for both cultural groups but, is not so high : for the Dutch group r=.22 (n=363), for the French group r=.34 (n=266). After correction for restriction of range and for attenuation in the criterion only (M.A. score) these r's become respectively .33 and .48.(*)
On the other hand, considering the fact that both scores are based on completely different observations (cfr supra), and the fact that during the first year at the military school the leadership training is very limited, these correlations can be considered as high.

The relations between the M.A. scores and the SA and SC scores seemed promising (Table 6) and for that reason the two scales were introduced in the selection procedure. But, as can be seen in table 6, the promising results could not be repeated the following years.

A second social anxiety scale, constructed by Willems (1973) (SA-W), and used in order to determine the concurrent validity of the SA scale did show about the same results as the SA scale (table 6).

The relations between the SA scale, the SC scale and the Military Aptitude score of the first training year remain unclear. The relations between the two scales and the leadership score hold as well for the R.O.'s as for the Reg.O.'s. In conclusion it can be stated that the selection team judges more self-confident and less social anxious candidates more able to become good reserve or regular officers than less self-confident and more social anxious ones.

Table 6 Correlations (Spearman) between Social Anxiety, Self-confidence and Military aptitude.

	Dutch Reg.O.			French Reg.O.		
	SA	SA-W	SC	SA	SA-W	SC
Military aptitude first year 1977	-.10	.02 (n = 97)	.16	-.20	-.28 ^{x} (n = 59)	.47^{xx}
Military aptitude first year 1978	.14	.14 (n = 100)	.07	.02	.11 (n = 65)	-.09
Military aptitude first year 1979	.07	.00 (n=121)	-.03	.00	-.02 (n = 83)	.07

xx significant at the 1% level
x significant at the 5% level

(*) The reliability of the M.A. score is .70 (n=527), this is the correlation between the M.A. scores of the first and the second training year.

The correlations for the R.O. population are : for the D.R.O.'s, r=.32 (n=173); for the F.R.O.'s, r=.31 (n=125), not corrected coefficients.

Intelligence, Anxiety and academic achievement

No significant correlations are observed between the SA and SC scale and academic achievement. From what is known from the test-anxiety literature, anxiety has a different effect on performance for persons with differing intellectual endowments. The effect of anxiety depends on the subject's level of ability. For the superior students anxiety facilitates academic performance, while anxious students in the middle ranges of ability tend to obtain lower grades than non anxious students of comparable ability. Students of low ability tend to earn low grades irrespective of their anxiety level (Gaudry § Spielberger, 1971; Sieber, 1977). This interaction seems also to exist in this Reg.O. population. Analyses of variance do show a significant interaction for SA by intelligence, for test-anxiety (F-, see infra) by intelligence, but not for SA-W by intelligence (Reg.O. 1977). Such an interaction does not appear when M.A. scores are taken into consideration instead of academic performance. It's worthwhile to note that there is no correlation between academic achievement and M.A. scores. Does the SA scale measure about the same quality as the test-anxiety scales? In the following paragraph a comparison will be made between these two characteristics.

The concurrent validity of the SA and SC scale

A number of personality inventories were used to describe the concurrent validity of the two scales (SA, SC). The social anxiety scale of Willems (1973) (SA-W) confronts the subject with 24 different possible stressful social situations and by means of a 5 point scale he can express the amount of anxiety provoked by each situation. A personality questionnaire of H. Hermans (1970), including Test-anxiety or Fear of failure (F-), Facilitating anxiety (F+) and Achievement motivation (P). The test-anxiety scale measures to what degree a test-like situation arouses tension or anxiety. It contains 26 items. Weighted average correlations were calculated between all these scales. The results appear in table 7. In the same table the correlations of the different scales with the L score are printed. In this table no distinction is made between R.O., Reg.O., Dutch or French groups. As expected, the SA-W and the SA scales are highly correlated (.58), both scales are measuring the same kind of anxiety. But, the correlation between SA-W and F- is even higher (.69). The explanation of this fact lays probably in the distinction between feelings of discomfort and behavior. SA-W and F- try to sure the degree of anxiety or discomfort in relation to specific social or 'test-like' situations, while SA emphasises the engagement in assertive behavior.
A person might feel high discomfort but engage in an assertive behavior in spite of this, or, such discomfort could be coupled with an avoidance behavior. This can perhaps explain the difference in relationship between the two SA scales and the L score. The L score expresses more what the subject is doing during a stressful social situation and not what he is feeling. And this is more central in the SA scale than in the SA-W scale. For this reason Eileen D. Gambrill and Cheryl A. Richey (1975) developed an Assertion Inventory which permits respondents to note for each item their degree of discomfort and their probability of engaging in the behavior.

TABLE 7. Weighted average correlations between anxiety, self-confidence, achievement motivation and leadership.

	SA-W	SA	SC	F-	F+	P
L	-.23 (3)	-.38 (9)	.35 (9)	-.33 (2)	.10 (2)	.10 (2)
SA-W		.58 (7)	-.44 (7)	.69 (2)	-.33 (2)	-.24 (2)
SA			-.47 (16)	.50 (8)	-.24 (8)	-.16 (2)
SC				-.47 (8)	.25 (8)	.29 (2)
F-					-.38 (8)	-.16 (2)
F+						.19 (2)

The number between brackets indicates the number of correlations. The number of subjects in each subgroup varies from 50 up to 200.

Anyway, social anxiety seems to be strongly related with test-anxiety. The SC scale is clearly negatively related with F-, the SA-W and the SA scale. This fits in with the description of the self-image of anxious subjects (Sarason § Spielberger, 1979). The correlations between ach. mot. (P) and the other scales are in the expected direction but are low. This is also the case for F+, what seems to support the idea of Hermans (1970), who, like Alpert and Haber (1960), tries to demonstrate that facilitating anxiety is not just a mirror image of test-anxiety. Depreeuw (1978), like Herron (1964), stresses an unidimensional theory concerning fear failure. He argues that the low correlation between F- and F+ can be explained by the formulation of the F+ items. The majority of these items refer to low level anxiety. "The F+ scale measures the degree in which light feelings of tension facilitate performance, while the F- scale measures the degree of anxiety (or tension) test-like situations evoke"

The reliability of the SA and SC scales

With an interval of one day the subjects answered first the original questionnaire (151 items) and second the SA + SC items only. The retest reliability coefficients are for the SA scale .78 (n=454) for the D.R.O. + D.Reg.O., .78 (n=483) for the F.R.O. + F.Reg.O., for the SC scale they are respectively .67 and .70. These coefficients are low. For this reason it was decided to maintain the original questionnaire and to calculate the SA and SC scores with appropriate keys. All calculations in this study are based on these scores.

CONCLUSIONS

There exists a stable relationship between military leadership, as defined by the officer's Selection Board, and social anxiety and self-confidence. There are some indications that this two qualities are related with the officer's training outcome, but, the facts are either weak (reserve officers) or too unstable (regular officers). Social anxiety is closely related with test-anxiety and a distinction needs to be made between feelings of discomfort and behavior.

Although, the SC items do not explicitely refer either to social stressful situations or to test-like situations, the relations of the SC scale with

the SA scales and the test-anxiety scale are substancial. This expresses the fact that self-confidence depends to an important extend on the way one copes with stress.

ALPERT, R. § HABER, R. : Anxiety in achievement situations, J. abnorm.soc.Psychol., 1960,61,p.207-15.

DEPREEUW,E. : Faalangst bij intelectuele prestaties. Theorie en behandeling, LEUVENS Bulletin L.A.P.P., 1978,27,365-75

DEPREEUW,E. : Faalangst als cognitief en emotionneel proces Dienst voor studie-advies K.U.L., LEUVEN, 1978

GAMBRILL EILEEN § RICHEY CHERYL. : An assertion inven. for use in assessement and research. Behavior Therapy, 1975,5,550-561

GAUDRY, E. § SPIELBERGER, C.D. : Anxiety and educational achievement, NEW YORK, 1971.

HERMANS, H.J. : A questionnaire measure of achievement motivation. Journal of Applied Psychology, 1970,54,353-363.

HERRON, E.W. : Relationship of experimentaly aroused achievement motivation to academic achievement anxiety, J.abnorm.soc.Psychol., 1964,69,690-94

RATHUS G. § NEVID J. : BT Behavior Therapy. Signet, New American Library, NEW YORK, 1978

SARASON, I.G. § C.D. SPIELBERGER (cds.). : Stress § Anxiety. 6 Vols Hemisphere Publ Co, Vol 1+2,1975;3,1976;4,1977;5,1978; 6,1979.

SIEBER, J.E., H.F. O'NEIL Jr, S. TOBIAS. : Anxiety, Learning § Instruction. Lawrence Erlbaum, John Wiley, 1977

WILLEMS L. et al. : Een schaal om sociale angst te meten. Nederlands Tijdschrift voor Psychologie, 1973,28,415-422

BOLDOVICI, John A., and HARRIS, James C., HumRRO, Fort Knox, Kentucky.

SOME PROBLEMS IN EVALUATING TRAINING DEVICES AND SIMULATORS
(Wed P.M.)

The purpose of this paper is to discuss a few of the problems that one encounters in the design of evaluations for training devices and simulators. The problems involve:

1. Defining Device Effectiveness. Various levels of the device-evaluation question are discussed, from, "Does practice on the device produce transfer?" to "What does the price/transfer curve look like?"

2. Selecting Dependent Variables. Choices frequently are made, not on grounds of relevance, but on other grounds-feasibility, for example, and anticipated use of data.

3. Training Devices as Test Media. Nearly all training devices are used as test media. The criteria for evaluating test media differ from the criteria for evaluating teaching devices.

4. Ceiling and Floor Effects. Avoidance of ceiling and floor effects requires considering measurement reliability, task difficulty (or entry-level proficiency), and amounts and kinds of practice.

5. Measuring Training and Other Weak Effects. Of the many problems assoicated with device evaluation, the most difficulty by far is the measurement of cause and effect. Suggestions are made for diminishing the effects of extraneous variables.

SOME PROBLEMS IN EVALUATING TRAINING DEVICES AND SIMULATORS

John A. Boldovici

and

James H. Harris

HumRRO, Fort Knox, Kentucky

INTRODUCTION

Several new devices are being developed as media for teaching soldiers to operate and maintain the XM1 tank. We are providing evaluation-planning assistance for several of these new devices, including the Unit Conduct-of-Fire Trainer, the XM1 Driver Trainer, the Turret Organizational Maintenance Trainer, and two targets systems--the Combat Training Theater, and the BT 41.

Guidance for evaluating training devices and simulators can be found in several TRADOC and Army Regulations and Circulars; 71-9, for example, and 70-1. The regulations share with most attempts to prescribe or "can" training development and evaluation matters, the characteristic of telling us a lot about what we already know. For solving the more difficult problems, however, the canned prescriptions are of little help.

The purpose of this paper is to discuss just a few of the difficult problems that one encounters in the design of evaluations for training devices and simulators. The problems involve:

1. Defining device effectiveness.
2. Selecting dependent variables.
3. Training devices as test media.
4. Ceiling and floor effects.
5. Measuring training and other weak effects.

Defining Device Effectiveness

The device evaluation question can be thought of as occurring at three levels. The most basic level is, "Does practice using the device produce transfer?" Answering this question requires only a device and some method of measuring Task B learning or performance. (Task B as used here means criterion performance--or as close to criterion performance as is possible to get without going to war.) The device need not provide for reliable measurement if it is used only for practice on Task A (the practice Task) and not for measurement on Task B. Whether analyses separate from those used to measure transfer are necessary for establishing the reliability of Task B measurement is equivocal. On the one hand, the assumption can be made that if differences in transfer are shown as a function of practice with the device and no practice with the device, then (barring Type I error) measurement must have been reliable. There are, however, at least two points to be made about this line of thinking: If no differences in transfer are found as a function of practice and no practice, then one cannot be sure whether the treatments were equally ineffective or

the measures of Task B performance were unreliable. Possible operation of ceiling and floor effects also is relevant here, and will be discussed later. The second point is that the possibility of Type I error should not be ignored where two conditions coexist. The conditions are (a) that new ground is being broken in the device evaluation (that is, a directly related evaluation of the device has not been proviously performed), and (b) that the probability of replicating the device evaluation is low. Since both of these conditions frequently apply to transfer studies with new devices, consideration should be given to designing studies which incorporate examinations of measurement reliability as well as transfer. This is easily accomplished by designing studies which require the use of repeated measures. As will be seen later, there are compelling reasons for using repeated measures designs irrespective of the reliability issue. Addressing reliability should therefore present little problem, aside from the fact that correlation typically is used to examine reliability, and is never appropriate for measuring transfer. (A high correlation between scores on Task A and scores on Task B simply suggests that the same skills are involved in performing the two tasks, and has no bearing on the transfer question.) This problem is easily solved by the use of analysis of variance designs for both transfer and reliability analyses, or by reserving the use of correlation for reliability analyses only.

A level above the simple evaluation question, "Does practice on the device produce transfer?" is the question of how much; that is, "How much practice on the device produces how much transfer?" Designs appropriate for answering this question provide data not only on transfer, but also permit avoiding ceiling and floor effects. They also address the pragmatic issue of diminishing returns from simulator practice, by providing data on optimal points for shifting trainees from Task A to Task B. Notice though, that if proficiency on Task A is used as a treatment in studies of the relation between amount of practice and amount of transfer, then unlike the situation discussed earlier (Does practice produce transfer?), the reliability of device-mediated measurement is unequivocally an issue. If the study design requires Group I to be trained to mastery level x, Group II to mastery level y, and so forth, then analyses should be performed to answer questions about the reliability of measuring mastery levels x and y.

In addition to the questions of whether and how much, one usually would like to know something about effectiveness of practice on the device relative to treatments other than practice on the device, including no treatment, conventional treatment, or both. This of course is the simplest version of the cost-effectiveness question, which can be variously stated as, "What is the least expensive means for achieving x transfer?" or "What is the most transfer that can be achieved for x dollars?" Or in its most elusive form--the form which most cost-effectiveness studies seem to address--"What is the optimal cost: benefit ratio?"

One would do well in our view, to discard from consideration the

first and last levels of the evaluation question mentioned above. The first is simply inefficient; with slight additional effort and little or no added cost, one can obtain answers to device evaluation questions that are more enlightening than, "Does practice on the device produce transfer?" The last-mentioned level of the device-evaluation question--"What is the optimal cost: benefit ratio?"--is probably unanswerable, since the requirements for widespread agreement on cost-accounting methods and on "units" of transfer are never likely to be met.

Studies should therefore be designed to address the intermediate levels of the device-evaluation questions posed above; namely, "How much practice on the device will produce how much transfer?" with particular concern for measurement reliability, and for identifying optimal "shift points" from Task A to Task B.

Selecting Dependent Variables

The selection of dependent variables and measures must begin with an examination of all sources of information about "what the device is supposed to do." The ideal device specification would be in terms of desired amount of transfer and money saved as results of practice using the device. As implied earlier, getting bogged down in after-the-fact cost analysis seems unwise. The impetus for developing devices in the first place usually is monetary. If the case for savings due to device use is not compelling on rational (prospective) grounds, then little will be gained by conducting after-the-fact cost analyses of the device and alternatives. It seems advisable to proceed from the assumption, "Given that use of the device can save money, how can additional savings, additional proficiency, or both be achieved?"

Even if one knows, however, what the device is supposed to do, the fact remains that there is no "science" of selecting dependent variables and measures, just as in test development there is no "science" of selecting criteria and measures (factor analysis notwithstanding). From a validity or relevance standpoint, one has absolutely no basis, as Gagné noted in 1954, for choosing between the criteria of hitting targets and scaring opponents away. Similarly, how does one choose between the measures of distance from target center and hit/miss? Such choices typically are made, not on the basis of validity or relevance, but on other bases--feasibility, for example, and anticipated use of data. Given the unavailability of a science for selecting dependent variables and measures, the main implication for designing transfer studies is that whatever measures are selected be justified. A modified Method of Rationales (Flanagan, 1951) seems appropriate, in which all reasonable-appearing dependent variables are considered, and the rationales for selecting some and rejecting others are stated in writing.

An issue that inevitably arises in device evaluations is whether it is acquisition of Task B or performance of Task B which is of interest. Our thinking at present is that it makes little difference in terms of study design: repeated measures of performance on Task B are desirable in any event--for two reasons: as a means of examining the

reliability of Task B measurement, and because performance on early Task B trials is not necessarily predictive of performance on later Task B trials. Given repeated measures on Task B then, the issue of acquisition vs. performance becomes a matter of data treatment: one can average over all (or later) trials to obtain performance measures. Or one can count number of trials to reach given proficiency levels to obtain acquisition measures. Or one can do both. The strengths and weaknesses of alternative designs and measures for transfer and studies are treated in full by Gagné and Crowley (1948).

Training Devices as Test Media

A training device which is not used in some testing capacity is a rarity. If scores or judgments are used to estimate trainees' mastery of less difficult objectives before proceeding to more difficult objectives, then the device is a test medium. If diagnosis and remediation are based on trainees' performance using a device, then the device is a test medium. If malfunctions or problems are presented to students in device-centered instruction in order to examine reactions to emergencies, then the device is a test medium. And if transfer studies are designed in which proficiency on Task A is a treatment, then the device is a test medium.

The criteria for evaluating devices as test media obviously are not the same as for evaluating devices as teaching machines. The minimal requirement for training effectiveness is simply that practice on the device yields positive transfer. The minimum requirements for test "effectiveness" are for reliability and validity. "Fidelity" in device-centered testing may be desirable from the standpoint of validity, but only to the extent it does not interfere with reliability. (The extent to which "fidelity requirements" conflict in their relation to training and testing is unknown--and notably non-existent as an object of scientific inquiry.)

Fitzpatrick and Morrison (1971) imply that the trade-off between fidelity and measurement reliability is direct:

> Good measurement is possible only when each examinee can be observed under similar circumstances; that is, when it is possible to control and hence standardize the displays, the surround, and the responses on which evaluation of performance will be based. Such control is characteristic of tests and is reflected in the high reliability of measurement that can be achieved with a good test. But as the test situation simulates reality more closely, control becomes more difficult. It generally would be agreed by those with experience in the matter that the more closely one tries to simulate a real criterion situation, the less reliable will be one's measurement of the performance (p. 240).

This view seems at odds with consideration of the characteristics of modern devices, and especially electronic devices. Modern visual systems provide a case in point. As our ability to approximate visual "reality" increases, there is no commensurate loss in stimulus control (i.e., control of displays and surround). Any improvement in the ability to generate visual reality in fact requires improvements in stimulus control. Thus it seems that the increased use of high-speed electronics should minimize the need for trading off realism and measurement reliability.

The increased use of electronics in device construction also has implications for improved diagnostic testing. Electronics are well-suited for counting and timing. Among the implications for testing is that the occurence and latency of nearly any overt response are easily recorded electronically with attendant potential for trouble-shooting human performance sequences.[1] This is especially relevant for improving the performance of time-constrained tasks, with which tank crewmen's jobs are replete. The limits on measuring overt responses no longer are given by available means of measurement, but by limits on the response rates themselves. Here again, the increased measurement capability is accompanied by no loss in measurement reliability: machines not only count and record faster than humans do, they also do so more reliably.

Finally, it is axiomatic that reliability is well served by standardization, which in turn can be achieved by instructor- or administrator-independence of tests. To the extent that device-centered tests are "canned," measurement reliability should improve.

The considerations presented above are not suggested as criteria to supplant the main functions of training devices; namely providing practice which yields positive transfer, saves money, or both. If, however, a training device is to double as a test medium, then the considerations mentioned here should not be ignored.

> Any test or measure of performance should aim for high validity, i.e., some sort of assurance that the test measures what it is supposed to measure. When one mentions measurement, though, reliability as well as validity is implied, since it must also be determined that the test yields a performance score which differentiates between a superior and inferior individual with some degree of dependability. To the psychologist, these are well-known concepts. They appear to be applicable without change or reservation to the measurement of performance by means of a training device (Gagné, 1954).

Ceiling and Floor Effects

Caution must be exercised in interpreting results showing no difference in learning or performance as a function of training treatments. One occasionally reads in the Armor "trade" literature

[1]See, for example, Moore (1977) for descriptions of improving Olympic athletes' performance using computer-based process measurement.

for example, reports of no difference in performance of a live-fire gunnery test as a function of simulator-based and conventional training. Only if scores are reported, is interpretation of the absence of differences possible. If scores are not reported, one can only wonder about possible floor effects. That is, all scores--those of experimentals and controls--might be abysmally low. In such cases, simulators might save money, but not yield desired proficiency. Questions also arise as to how the scores of the simulator and conventional groups might compare with the scores of crews that received no or extremely "watered-down" training (a ten-minute lecture on how to hit targets, for example).

Ceiling effects are of course the opposite of the situation hypothesized above: Both groups are so proficient on Task B before learning Task A, that learning Task A produces no measurable increment in proficiency on Task B. (The problem is analagous, at least partly, to retention-study designs in which all Ss are trained to 100 percent proficiency at the outset, then tested at a later date, when their scores have nowhere to go but down.)

Analysis and avoidance of ceiling and floor effects require consideration of at least four factors: measurement reliability, task difficulty (or baseline proficiency), amount of practice, and kind of practice. Assuming reliable measurement, ceiling and floor effects can be viewed as special cases of violating the maxim, "Begin where the trainee's performance is at" (Malot, 1972); that is, ceiling and floor effects are the result of amounts or kinds of practice that are inappropriate to Ss' pre-training proficiency levels.

The implications of ceiling and floor effects for simulator and experimental design seem not to be widely appreciated. A finding of no difference between conventional and simulator groups should not automatically be viewed as a decisive blow for the cause of simulation. Four implications of ceiling and floor effects warrant particular mention. The first implication is that modern instructional technology, including state-of-the-art devices and simulators, can be expected to have little or no effect on the acquisition or retention of easy tasks. If a task is so easy that it can be mastered in one or two trials, or on the basis of simple verbal instructions without rehearsal, then learning is unlikely to be improved by the use of devices. Learning may be made less expensive in such cases by substituting devices for operational equipment, but proficiency maintenance or improvement is likely to be unaffected.

The second implication is related to the first: if a task is so difficult that it is seldom performed to standard or reliably in operational situations, then performance may or may not be improved by simulator-based practice. The key here is in identifying correlates of satisfactory and unsatisfactory learning and performance in the operational setting. If, for example, one can determine that poor performance is due simply to lack of opportunity to practice, then devices, appropriately used, can have beneficial effects. If, on the other hand, the poor performance is due to equipment unreliability, or limits on the sensory or motor capabilities of operators, or any sources of variance other than Ss' performance, then practice with devices may have no reliable effect on performance of the criterion task. Performance on tank gunnery tests may well be a case in point.

The third implication is that device-evaluation studies should examine interactions between amounts and kinds of practice on the one hand, and Ss' proficiency levels on the other. One way to do this is to select Ss at the extremes of proficiency, administer equivalent amounts and kinds of training to the high- and low-proficiency groups, and compare learning or performance on Task B. Selecting people at the extremes of proficiency is, however, both risky (from a reliability standpoint) and expensive. A better approach is to select random samples from the population of interest, and to train various groups to various levels of proficiency in the simulator before examining learning or performance of Task B. (The results of such designs also would provide clues about optimal "shift points" from simulation to criterion tasks.) This general paradigm, in which proficiency on Task A is a treatment, should receive primary consideration for the design of transfer studies for new devices.

The fourth implication is related to the third: Ss must be carefully selected to be as representative as possible of the proficiency levels of the trainee population for whom the devices are intended. If this is not done, then a device might be incorrectly judged effective or ineffective when in fact the sample was at fault. In its extreme form, this problem would manifest itself in transfer effects that were opposite for sample Ss and population trainees.

Measuring Training and Other Weak Effects

The main point to be made here is simply that interest in the retention of tasks by military personnel should not be allowed to confound transfer issues. Of the many problems associated with training assessment, the most difficult by far is the measurement of cause and effect.[1] In any system, the further removed that the outcome being assessed (in this case, performance on Task B) is from the inputs that are to be evaluated (in this case, practicing Task A using a device), the greater is the number of extraneous variables that obscure the relationship being explored. What is measured in an assessment of "typical" on-the-job performance is inevitably the result not only of practice on the device, but also of adequacy of selection, of individual differences in working habits and personal characteristics, of the effects of supervision, and of the error impact of numerous temporal and idiosyncratic conditions that affect performance of Task B. The results also are affected by how much (and what) the trainee learns between Tasks A and B. As a result, teasing out the specific impact of practice using a device on the ultimate criterion of on-the-job performance--if it can be done at all--is a very costly operation. Each extraneous variable that must be controlled imposes additional requirements for measures and for sample size, and a truly responsive design quickly assumes astronomical proportions.

Temporal effects on transfer obviously are important, in that tracking performance decay provides information necessary for scheduling "refresher" training, and may reveal differences associa-

[1]This and other points in this section were made by Paul A. Schwarz (1971) in a proposal written with the senior author.

ted with methods, rates, and amounts of original learning. But in studies of device effectiveness it makes little more sense to measure performance of Task B long after Task A has been learned than to search for analgesic effects long after an aspirin has been administered.

For the reasons noted above and others, transfer studies typically require Ss to learn Task B immediately or almost immediately after learning Task A. The rationale here is that immediate measurement reveals the maximum performance capability attributable to the treatment (practice on the device in the present case), and this is really the most appropriate criterion for assessing its specific effects.

REFERENCES

Fitzpatrick, R. and Morrison, E.S. "Performance and Product Evaluation" in Educational Measurement (Thorndike, R.L., Ed.) p. 240, 1971.

Flanagan, J.C. The Use of Comprehensive Rationales in Test Development. Educational and Psychological Measurement, 1951, 11, 151-155.

Gagné, R.M. Training Devices and Simulators: Some Research Issues. American Psychologist, 1954, 9, 95-107.

Gagné, R.M. and Crowley, M.E. The Measurement of Transfer of Training. Psychological Bulletin, 1948, 45, 97-130.

Malot, R.E. Contingency Management in Education, Kalamazoo, Michigan: Behaviordelia, 1972.

Moore, K. Gideon Ariel and His Magic Machine. Sports Illustrated, 1977, 52-60.

BOONE, Dr. James O., Federal Aviation Administration Academy, Oklahoma City, Oklahoma.

THE FEDERAL AVIATION ADMINSTRATION RADAR TRAINING FACILITY EMPLOYEE SELECTION AND TRAINING (Tue A.M.)

The Federal Aviation Adminstration (FAA) has recently constructed a radar training facility (RTF) in Oklahoma City, Oklahoma, to aid in screening appropriate personnel for work in radar air traffic control (ATC). The approach is based on the idea that limited exposure to simulated radar ATC in a controllled and measured environment will lead to the identification of persons who possess the skills and attributes necessary for success in this type of work. This report describes the results of a study performed at the FAA National Aviation Facility Experimental Center. (Now FAA Technical Center) comparing an over-the-shoulder method of scoring student performance with scoring by computer derived measures for use in screening at the RTF.

Results indicate that the computer-derived measures are far more reliable than over-the-shoulder scoring and the computer-derived measures predict a global rating of potential succcess in radar ATC at least as well as over-the-shoulder scoring. The implications of the results are discussed in relation to other automated training systems.

THE FEDERAL AVIATION ADMINISTRATION'S RADAR TRAINING FACILITY AND EMPLOYEE SELECTION AND TRAINING

James O. Boone, Linda Van Buskirk, and JoAnn Steen
Aviation Psychology Laboratory, Civil Aeromedical Institute
Federal Aviation Administration, Oklahoma City, Oklahoma 73125 USA

Successful air traffic control specialists (ATCSs) who have made a transition from manual to automated air traffic control (ATC) appear to prefer the advantages in the automated environment. However, some prospective ATCSs do not perform successfully in radar ATC. Successful employment in the radar environment requires that a person possess certain aptitudes. It is in the interest of the Federal Aviation Administration (FAA) and the prospective ATCSs to determine as soon as possible if the prospective ATCS possesses the aptitude necessary to successfully operate in the radar ATC environment. The philosophy of the FAA in regard to this selection process is that the best way to measure aptitude is to place the prospective ATCS in a radar simulation laboratory and perform a systematic, objective appraisal of the person's potential. To this end the FAA has constructed a Radar Training Facility (RTF) at the FAA Academy in Oklahoma City, Oklahoma. The training/screening process involves a mini-radar training program with rigorous assessment which occurs over a 4- to 5-week period. During this period, the trainee receives basic radar training sufficient to allow systematic evaluation of his or her performance. Those who demonstrate potential to become successful ATCSs are retained and those who do not are screened from the program. To explain this system, the RTF background, RTF positions, system operation, and the evaluation process are described in detail below.

The original simulators used in FAA ATC training were "patches" developed for the operational automated field systems. The "patches" permitted flexible training at designated positions without interfering significantly with the operational positions. Experiences with these prototype simulators resulted in at least two major notions related to using simulation for radar training. First, the value of computer-driven simulation for training purposes was firmly established. Second, several problems associated with using operational field systems in a training mode were identified. A 1965 Institute for Defense Analysis (IDA) study on the training of air traffic controllers discussed some of these problems and suggested that a standardized computer-driven program should be established by the FAA to provide basic radar training. The IDA study further suggested that the radar training should be pass/fail to select out those persons who did not demonstrate the potential to perform proficiently in a radar environment.

In July 1976, engineering requirements were completed by the FAA for a radar training system. During that same month the FAA Administrator approved the procurement and construction of the RTF to be located at the FAA Academy in Oklahoma City.

In October 1977, the FAA completed a program implementation plan that outlined the development and implementation of the RTF. The contract for the development of the computer-driven simulator training system was awarded to Logicon, Tactical and Training System Division, San Diego, California, in January 1978. Groundbreaking for the construction of the new RTF at the FAA Academy was held on December 22, 1977. The new facility was built and accepted by the FAA in January 1980, and the training system developed by Logicon Corporation was accepted in April 1980.

(115)

RTF Training System and Laboratory Configuration.

The primary objective of the RTF is to closely duplicate the specialized operational environment existing at automated terminal and en route facilities as well as have the capability of synthesizing and presenting a wide variety of air traffic control situations. These situations would be based on a reference data base created through scenario programs with a full range of control necessary to establish a realistic simulation of actual aircraft traffic under a variety of conditions.

To accomplish this objective, four independent laboratories are utilized. Figure 1 describes how the laboratories are configured.

Positions. There are Trainee positions and Supervisory and Support positions/stations corresponding to each radar training sector. At a "position," the operating personnel have input/output (I/O) equipment access to the system with associated voice communications for monitoring, instructing, and supervisory functions.

Trainee Position.

1. Radar Control Position (R). The R controller positions (six in each lab) have a display console, (PVD) for en route and (DEDS) for terminal. They have associated voice communications. The displays and voice communications are similar to those at field facilities. Displays include maps, weather, aircraft position symbols, alphanumeric readouts, and other digital and symbolic data.

2. Nonradar Controller Position (HO/D). The D controller for en route and the HO position for terminal (six in each lab) have the capability of making and accepting handoffs. This position also permits training for manual or nonradar control by using flight progress strips generated by the flight strip printers.

3. Pilot Position (P). Three pilot positions are associated with each sector (18 in each lab). These positions are in a separate room. Each position operator performs at a console with a tabular display and keyboard for data entry with associated voice communications. These operators simulate aircraft pilots during the exercise by actual responses to ATC clearances/instructions.

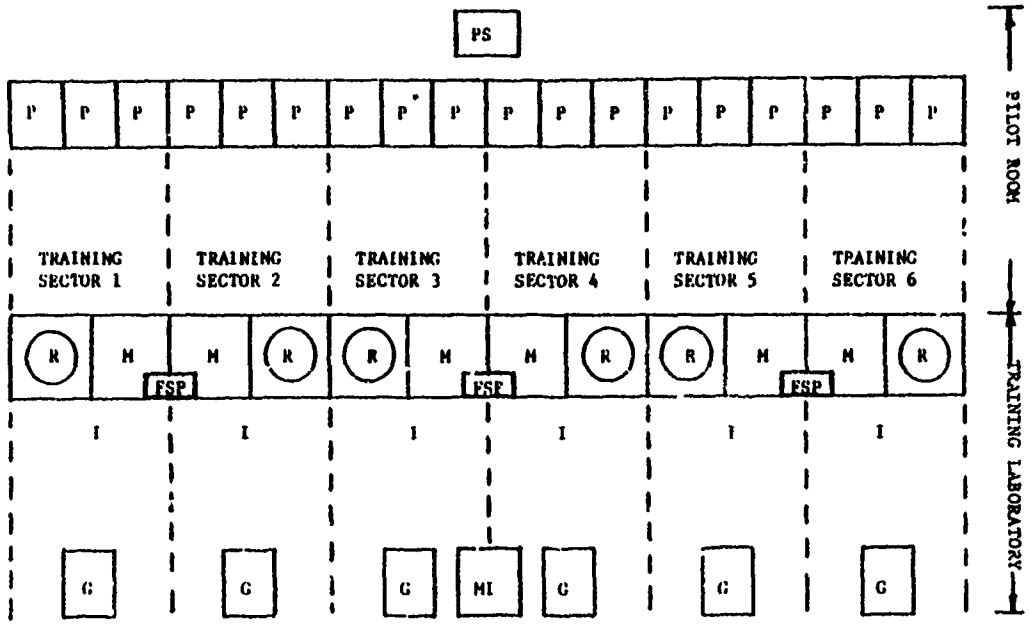
4. Ghost Position (G). This position is associated with each R and/or HO/D position. There are six ghost positions in each lab. The position console and display are identical to those of the pilot position. The ghost position operator adds realism to the exercise by performing related functions of adjacent centers, terminals, flight service stations, and positions/sectors. Functions include initiating handoffs, accepting handoffs, and generally ghosting functions of other facilities/sectors.

Supervisory and Support Positions/Stations.

1. Instructor Station (I). An instructor station is provided at each sector (six in each lab). The instructor has voice communication with each student and monitors the overall exercise from behind the trainee positions.

2. Pilot Supervisory Station (PS). This position (one in each pilot room) has voice communications for supervising, monitoring, and instructing operation of pilot positions as well as for coordinating activities with the master instructor station and the system monitor position.

(116)



P - Pilot Console
G - Ghost Console
R - Radar Trainee Console
M - Manual Controller Trainee Position
Enroute "D" Controller or
Terminal "HO" Controller

MI - Master Instructor Position
PS - Pilot Supervisor Station
I - Instructor Station
FSP Flight Strip Printer

Figure 1. Laboratory configuration.

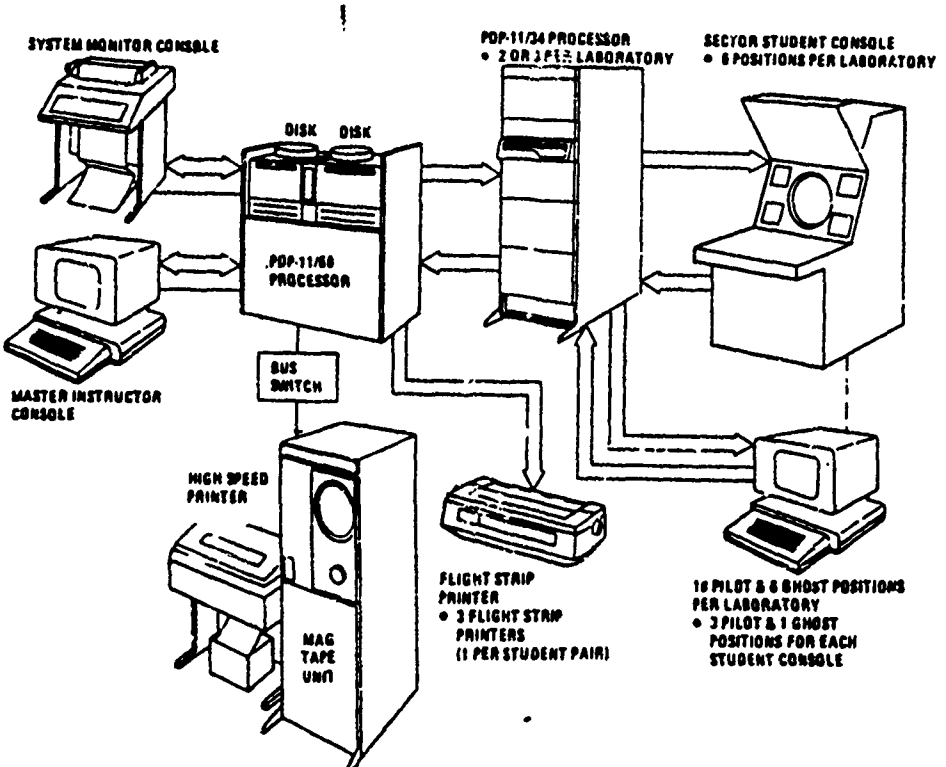


Figure 2. Computer system configuration.

3. Master Instructor Station (MI). This position (one in each lab) controls the exercise within the lab. The position has a tabular display, a data entry keyboard, and associated voice communications with each trainee and with each operator of ghost, instructor, and pilot positions in the lab. The master instructor station will permit setting clock time, starting, monitoring, freezing, backing up, replaying, and restarting the exercise as necessary. The position also provides for data recording and analysis of the exercise.

4. System Monitor Position (SM). One position is provided for each lab. The position will have voice communications with two master instructor positions and two pilot supervisor positions. The position will permit computer operation and operational and maintenance monitoring.

Figure 2 describes the system configuration for operating the positions and stations in each laboratory. The training sectors are controlled by a Digital Equipment Corporation (DEC) PDP 11/60 computer with a PDP 11/34 computer serving as an interface between the PDP 11/60 and the operating positions.

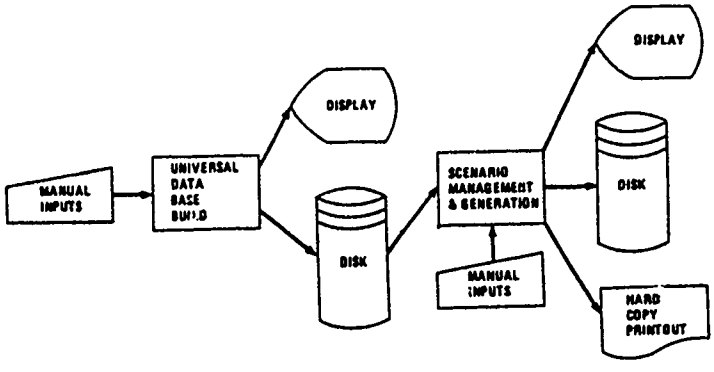


Figure 3. Components of scenario generation.

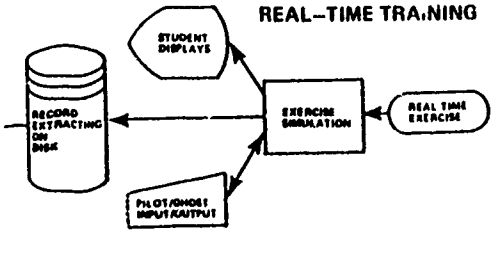


Figure 4. Components of the real-time training system.

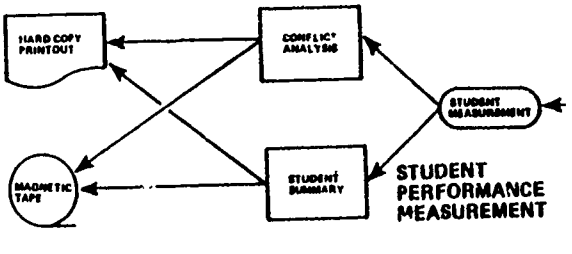


Figure 5. Components of the student performance measurement.

The training process involves three sequential systems of operation: (1) SCENARIO GENERATION --→ (2) REAL-TIME --→ (3) PERFORMANCE MEASUREMENT. Scenario generation, illustrated in Figure 3, is the non-real-time process of building exercises and evaluation problems for the system. Aircraft characteristics, flight plans, and other essential information of this type are stored in the Universal Data Files (UDF). The exercise is built by first selectively retrieving intermediate files and then creating other intermediate data files from the universal data base through the scenario management program.

The real-time component, illustrated in Figure 4, utilizes the scenario management files to generate the actual radar simulation exercise. The real-time component drives the display at the radar position. Aircraft movement is controlled through the pilot and ghost positions according to the instructions the operators of those positions receive from the controller trainee or, in some cases, from a scenario prompt which appears on the cathode-ray tube (CRT) at the pilot or ghost positions. During the operation of the real-time training exercise, all actions taken during the exercise are recorded.

At completion of the exercise, the computer will analyze the recorded actions to determine violations of separation standards and to quantify other pertinent performance information, such as delay times, in order to evaluate the student's ability to move air traffic "safely and expeditiously." The process of student performance measurement is illustrated in Figure 5.

Table 1 contains a list of the computer-derived measures to be employed in evaluating the students' performance on a given problem. The primary focus of this paper is the student performance component in the training system.

TABLE 1. List of RTF Measures

1. Number of aircraft in the sample
2. Ideal aircraft time-in-system (based on filed flight plan)
3. Ratio of the ideal aircraft time in system and the number of aircraft in the sample
4. Number of completable flights
5. Data period duration
6. Number of arrivals
7. Number of departures
8. Arrival/departure ratio
9. Arrival rate scheduled per hour and departure rate scheduled per hour
10. Conflicts--terminal (3 nautical miles (NMI))
11. Conflicts--en route (5 NMI)
12. Number of delays (start time)
13. Delay time (start time)
14. Number of delays (hold and turn)
15. Delay time (hold and turn)
16. Number of delays (arrival)
17. Delay time (arrival)
18. Number of delays (departure)
19. Delay time (departure)
20. Number of delays (total)

(TABLE 1 continued on next page)--

21. Delay time (total)
22. Aircraft time-in-system (real)
23. Number of aircraft handled
24. Number of completed flights (total)
25. Number of arrivals achieved
26. Arrival rate achieved per hour
27. Number of departures achieved
28. Departure rate achieved per hour
29. Number of air-ground contacts
30. Air-ground communications time
31. Number of altitude changes
32. Number of heading changes
33. Number of speed changes
34. Number of path changes (altitude, heading, and speed)
35. Number of hardoffs

Background in Performance Measurement

The earliest studies in air traffic control which used some form of automated measurement were conducted by a Civil Aeronautics Administration (CAA) group in Indianapolis, Indiana, at the Technical Development Center (TDC) with support from the Franklin Institute of Philadelphia (18,20,51,65). The "dynamic simulator" used at the TDC consisted of a translucent screen on which maps could be projected with motor-driven light projectors capable of projecting a spot of light and moving it across the screen to simulate radar echoes from an aircraft. Personnel acted as pilots by moving the aircraft across the screens according to the control messages they received over a telephone line. The setup resembled the radar Plan Position Indicators (PPI) used in air traffic control (15). Research at this facility spanned from 1950 to 1959, at which point it was moved to the National Aviation Facilities Experimental Center (NAFEC; renamed FAA Technical Center in May 1980).

The research at TDC covered topics in air traffic control such as (i) airport design, (ii) approach systems, (iii) ATCS workload, (iv) data acquisition, and (v) decision making. Reports on the studies contained quantitative data on (i) number of separation violations, (ii) number of aircraft delayed, (iii) average delay per aircraft, (iv) altitude changes, (v) number and length of communications, and (vi) number of missed approaches (10,11,12,13,14,67,68,69).

Concurrent with the TDC studies, a series of 19 simulation-based experiments were conducted in air traffic control at the Ohio State University's Aviation Psychology Laboratory under the direction of Paul M. Fitts (27). The studies were performed between 1954 and 1961 and involved measurement of controller performance. In 1954 Hixson et al. (34) developed an electronic radar target simulator for air traffic control studies. As a part of the development, Hixson made performance measurements on the accuracy of "headings," "airspeed," "turn rate," and "attitude" for each target generated. A camera was mounted on the display indicator and the path of the aircraft was recorded. Calculations were then computed from the recordings to measure the accuracy of the simulation. These measures were used to determine the accuracy of the system operation.

Later studies at Ohio State University involved more direct measurement of ATCS performance and were conducted on a variety of topics such as (i) pattern-feeder controllers, (ii) individual differences among subjects, (iii) display variables, (iv) workload variables, and (v) procedural variables (9,35,37,39,40,41,42,43,44,

45,46,47,48,49,52,57,58,59,60,61,66). Several different types of measures were used to assess subject and system performance. These include measures of (i) overall flight time, (ii) percent delay time, (iii) fuel consumed, (iv) missed approaches, (v) separation errors, (vi) time intervals between landings and departures, (vii) time and frequency of communications, (viii) delay time in responding to emergency situations, and (ix) traffic load, i.e., number of aircraft in the problem and number of aircraft handled.

During the 1960's research involving ATCS performance was done by at least three groups: (i) the MITRE Corporation, (ii) the Systems Development Corporation, and (iii) NAFEC. Between 1961 and 1963 the MITRE Corporation conducted six studies in air traffic control. The six studies covered topics in (i) high altitude air traffic control, (ii) beacons and automatic tracking, (iii) display clutter on the CRT, (iv) multisector coordination, (v) handoff procedures between en route and terminal, and (vi) conflict resolutions (33,36). The studies were performed by computer-generated simulation where "canned" scenarios were constructed and then run in real time. Automated measures taken in the studies included: (i) traffic load, (ii) teletype usage, (iii) frequency of various displays, (iv) flight plan deviations, and (v) conflicts.

In 1961 the System Development Corporation began a series of studies in air traffic control. The studies were performed by computer-generated simulation and sufficient information from each program to subsequently reproduce the problem was stored on mag tape. Studies were conducted on topics such as: (i) spacing of aircraft, (ii) geographic point of aircraft entry, (iii) heterogeneity of aircraft, and (iv) procedural variations (2,3,4,5,6,7,8,25,29,30,54,55,56). The stored data from the problems made possible an extensive list of postexercise measures. These included (i) safety violations, (ii) percent of time aircraft in holding pattern, (iii) percent of aircraft held, (iv) difference between actual flight time and time by the shortest available path, (v) the ratio of iii and iv, (vi) mean time spacing between successive aircraft, (vii) aircraft waiting time before departure, (viii) delay time holding, (ix) fuel consumption, (x) variability in aircraft arrival time, (xi) number of radio communications, (xii) average communication time, (xiii) average number of communications per aircraft, (xiv) total communication time, (xv) number of controller data entries, and (xvi) number of clearance points an aircraft passed.

Perhaps the most extensive research during the 1960's involving performance measurement occurred at the FAA NAFEC facility. As previously mentioned, the simulator at the TDC was moved to NAFEC and used until about 1962. Between 1960 and 1962 NAFEC also had a Model A and Model B simulator installed. The simulators generated radar echoes on a CRT. Pilots were also employed to move the echoes around on the CRT. Later, a Sigma 5 computer was introduced which extended NAFEC's simulation capabilities. A sampling of the research topics covered included: (i) dual approaches, (ii) combining approach facilities, (iii) equipment arrangements, (iv) traffic flow patterns, (v) final approach spacing, (vi) display usage, (vii) airspace jurisdiction, (viii) helicopter movement, (ix) supersonic control procedures, (x) airport site selection, and many more (1,20,26,32,38,50,53,64,65,66,70,71). Measures employed in the studies consisted of (i) delay time, (ii) number of vectors, (iii) number of holds, (iv) conflictions, (v) aircraft time in the system, (vi) interval between arrivals, (vii) communication workload, (viii) number of departures and arrivals, (ix) the ratio of departures and arrivals, (x) missed approaches, (xi) total aircraft handled, and several others.

During the 1960's and into the 1970's, there was a shift in emphasis in performance measurement at NAFEC. While most of the prior NAFEC research had employed these measures to evaluate various equipment, procedures, or configurations, research interests shifted to using automated performance measurement to evaluate how well the ATCS was performing. The basic notion at that time was to construct norms on ATCS proficiency through Controller Performance Measurement (CPM) (19,22). A 1969 study on controller aging (20) set the groundwork for the NAFEC effort in CPM, and later two experiments termed PROBE tests further supported the basic performance measurement rationale (21). The tests demonstrated the possibility of developing parallel problems and of identifying a consistent ATCS profile across different sectors. The early tests demonstrated that a larger random sample of representative ATCSs was needed in order to proceed.

With the introduction of the RTF, it was decided that the feasibility of using computer-derived measures to evaluate ATCS student performance should be studied. The present system of student evaluation consists of an over-the-shoulder observation of students by expert air traffic controllers with recent field experience. Scores are comprised on the basis of a composite of instructor ratings (Instructor Assessment) and a count of errors committed while controlling simulated aircraft (Problem Average). A study was designed, employing the computer-driven ATCS simulation lab at NAFEC, to study the possibility of using automated measuring devices as a substitute for the Problem Average portion of the composite score. The purpose of the study was twofold: (i) to make a preliminary assessment of the feasibility of using computer-derived measures (CDM) to evaluate student laboratory performance and (ii) to improve the over-the-shoulder evaluation procedure for student laboratory evaluation.

Methods. To accomplish these goals, 48 students, 24 en route and 24 terminal, were transported to NAFEC to receive radar training and evaluation at the Dynamic Simulation Facility. The students were evaluated over-the-shoulder by an instructor and the problems were recorded by computer on mag tape and later reduced to a set of computer-derived measures (see Table 2 for a listing of the measures used).

Five problems in increasing complexity were administered to each student. Each instructor had an opportunity to observe each student at least once. On problems 4 and 5, randomly selected students were evaluated over-the-shoulder simultaneously and independently by two instructors. An index of agreement (reliability) was computed on the simultaneous evaluations by forming a ratio of the number of agreements over the total number of error conditions recorded by the two instructors. An initial laboratory evaluation manual and a laboratory evaluation form were developed by consensus of the instructors in each option prior to the study; the manual and lab form were modified during the study based on new agreements formed after reviewing the disagreements on the laboratory evaluation forms.

After each student was evaluated on the individual problems, each instructor provided an overall, global rating, stating the student's potential to become a full performance level (FPL) radar controller. The rating was a 5-point global scale, (1) definitely will not become FPL, (2) maybe (doubtful) FPL, (3) minimally acceptable FPL, (4) good FPL, and (5) definitely excellent FPL. The global rating was based on the instructor's observations of the student operating the radar problems.

Analyses included the following: To determine the feasibility of using computer-derived measures, those measures were used in a regression equation to predict the (i) problem average (PA), (ii) instructor assessment (IA), and (iii) total score on

TABLE 2. A Listing of the Computer-Derived Measures and Their Corresponding Reference Numbers Employed in the NAFEC Study

1. Conflicts (5-mile separation)
2. Conflicts (3-mile separation)
3. No. Start Point Delays
4. Start Point Delay Time
5. No. Turn and Hold Delays (turns longer than 100 seconds)
6. Turn and Hold Delay Time
7. Aircraft Time-in-System
8. No. Aircraft Handled
9. No. Completed Flights (transfers to 130.5 must be from ghost position)
10. No. En Route Departures (Code 2)
11. No. Terminal Arrivals (Code 3)
12. No. Terminal Departures (Code 4)
13. No. Air-to-Ground Communications Time
14. Air-to-Ground Communications Time
15. No. Altitude Changes (pilot keyboard messages)
16. No. Heading Changes (pilot keyboard messages)
17. No. Speed Changes (pilot keyboard messages)
18. No. of Handoffs From Feeder Position to Subject
19. Handoff Delay Time
20. No. Beacon Re-Idents

the over-the-shoulder evaluation. Further, the individual problem scores from the over-the-shoulder evaluation were used in a regression equation to predict the global rating score for each student. A regression analysis was performed using the CDM and IA regressed on the global rating to compare with the PA and IA on the global rating. The indices of agreement reliability for the simultaneous over-the-shoulder evaluations were also computed and listed by problem and option. A reliability index (intraclass correlation) was also performed on the global rating data. Profiles across students and across instructors were computed by stratifying the errors on the lab forms according to error categories identified by a group of controllers who reviewed the worksheets. The frequencies of the errors were then summarized by category (Table 3 contains a listing of the over-the-shoulder measures). An orthogonal, varimax factor analysis was also calculated to group the measures in multidimensional space and to compare the underlying dimensions of the error categories in the over-the-shoulder and computer-derived measures.

The reliability coefficients for the over-the-shoulder problem averages were computed as previously described. The global rating and instructor assessment reliabilities are intraclass correlations across all instructors for each student. The reliability coefficients are important for several reasons. In the case of the over-the-shoulder evaluation, it indicates the proportion of times that two instructors agreed on a particular error marked against the student's grade. Disagreements occurred in two ways: The instructors recorded the same event as an error but differed in the type of error they called it, or one instructor recorded an error for an event while the other instructor either failed to see or did not judge it to be an error. It can be readily noted that the instructor assessment is more reliable than the problem average. The reliability of the problem average is important since the validity of a measure cannot exceed its reliability. Consequently, it is very important to standardize any portion of the grading procedures that requires instructor judgments.

(13)

TABLE 3. A Listing of the Over-the-Shoulder Measures and Their Corresponding Reference Numbers

I. SYSTEM ERRORS (S)

1 - Vertical
2 - Lateral, long
3 - Terrain
4 - Airspace outside radar coverage

II. SYSTEM DEVIATIONS (D)

1 - Airspace (lateral)
2 - Altitude (facility)
3 - Altitude (aircraft data block) (min. separation used, no alt. verification)

III. PROCEDURE (P)

1 - Keep them high
2 - Speed control
3 - Bad vector
4 - Delay
5 - L.O.A. (letter of agreement)
6 - Holding-EAC/EFC
7 - WAFDF (wrong altitude for direction of flight)
8 - Needless altitude change
9 - Radar contact not given to ACFT
10 - No reason for vector
11 - Traffic
12 - Position of ACFT. Incorrect or not given.
13 - SID (change in route)
14 - Missed approach instructions
15 - Remarks
16 - Improper coordination
17 - Beacon code
18 - Point out
19 - Route
20 - Altitude
21 - Transfer control
22 - Change of destination
23 - Change of ACFT. Status (VFR/IFR)
24 - Altitude verification
25 - Clearance

IV. OTHERS (O)

1 - Phraseology
2 - Strip marking
3 - Altimeter not issued
4 - Overrestriction
5 - Improper feedback of wrong information
6 - Data block update within sector
7 - Board management

TABLE 4. Reliability Coefficients for the Over-the-Shoulder Evaluation and Q-Sort by Option

	Problem Average	Instructor Assessment	Total Score	Global Rating
Terminal	.326	.582	.433	.234
En route	.294	.561	.427	.266

Model 1 (Table 5) demonstrates the ability of the computer-derived measures to duplicate the problem average in the over-the-shoulder evaluation. The Beta weights indicate the relative importance of each of the computer measures in the duplication process. The "R," multiple correlation, ranges from -1.0 to +1.0 and is a measure of the overall fit of the model. A .5212 is a moderate to good value; however, the

value could increase considerably if the unreliability in the problem average were minimized.

TABLE 5. Regression of Computer-Derived Measures (CDM) on the Over-the-Shoulder Problem Average (PA)

Model 1
Predictors = 1-20
R = 0.5212

V	BETA
1	0.1147
2	0.0365
3	0.2637
4	0.0128
5	0.1704
6	0.0126
7	0.0298
8	0.1649
9	0.1791
10	0.8536
11	0.0586
12	0.6821
13	0.2704
14	0.3552
15	0.2906
16	0.1167
17	0.0542
18	0.2582
19	0.1593
20	0.0507

Models 2 and 3 (Tables 6 and 7, respectively) demonstrate how well the computer-derived measures duplicate the instructor assessment and the total score. The increase in "R" for instructor assessment is probably due to a better reliability in the instructor assessment.

Model 4 (Table 8) demonstrates in the Beta weights a tentative schema for weighting the lab problems to form a composite lab score. The information provided by the problems is highest in problems 4 and 5. The maximum amount of information peaks at problem 4. Thus, a five-problem lab grading procedure offers the most information, but a four-problem procedure would be an efficient manner of maximizing information in the shortest time frame. The relative weightings for five problems would be 10, 10, 15, 40, and 25, and for four problems would be 15, 15, 30, and 40.

Models 4 and 5 (Tables 8 and 9, respectively) demonstrate how well the computer-derived measures can be used in place of the problem average in predicting the global rating. The multiple "R" drops from .4493 to .4299, an insignificant decline. For practical purposes, the computer-derived measures can be used in place of the problem average in forming an overall grade. This approach would have at least one very strong advantage. The computer-derived measures are completely reliable whereas the problem average is considerably unreliable. Combining the highly reliable computer-derived measure with the moderately reliable instructor assessment creates

(126)

TABLE 6. Regression of Computer-Derived Measures on the Over-the-Shoulder Instructor Assessment

Model 2
Predictors = 1-20
R = 0.5302

V	BETA
1	0.1547
2	0.0390
3	0.3446
4	0.0157
5	0.1669
6	0.0100
7	0.2337
8	0.1343
9	0.2099
10	0.8121
11	0.0000
12	0.7387
13	0.1292
14	0.1970
15	0.3665
16	0.2153
17	0.0169
18	0.1143
19	0.1602
20	0.0041

TABLE 7. Regression of Computer-Derived Measures on the Over-the-Shoulder Total Score

Model 3
Predictors = 1-20
R = 0.5247

V	BETA
1	0.1942
2	0.0387
3	0.3533
4	0.0107
5	0.1370
6	0.0546
7	0.3451
8	0.0039
9	0.2177
10	0.5529
11	0.0790
12	0.5079
13	0.1628
14	0.1861
15	0.3735
16	0.3786
17	0.0335
18	0.0842
19	0.1308
20	0.0808

TABLE 8. Regression of 5 (PA+IA) *Probs on Global Rating

Model 4
Predictors = 1-5
R = 0.4493

V	BETA	B
Prob 1	0.0928	0.0062
Prob 2	0.0742	0.0043
Prob 3	0.1376	0.0096
Prob 4	0.3029	0.0147
Prob 5	0.1923	0.0090

*Problems

REG. CONST. = 1.8253

TABLE 9. Regression of CDM+IA on Global Rating

Model 5
Predictors = 1-5
R = 0.4299

V	BETA	B
Prob 1	0.1851	0.0007
Prob 2	0.3511	0.0012
Prob 3	0.8663	0.0017
Prob 4	0.0515	0.0001
Prob 5	0.6531	0.0012

a problem average reliability of approximately .750, which is a significant improvement over the previously reported .433 and .427 for terminal and en route, respectively.

The evidence from Models 1-5 suggests that the computer-derived measures are useful and valuable contributions to the assessment process. The validity of the measures is not established by this study; however, using the computer measures in place of the over-the-shoulder problem average increases the reliability significantly, and reliability is the upper bound for validity.

The factor analyses offer a means to (i) identify cluster areas where general measures are incurred by students and (ii) provide a comparative basis for the underlying structures of the two grading systems. The factor analyses point out at least two major differences in the two evaluation models: delays and system deviations. Delays are difficult to determine over-the-shoulder and no method was available to measure system deviations in the computer-derived measures. A next step would be to attempt an optimal combination of over-the-shoulder measures and computer-derived measures to be averaged with the instructor rating.

Conclusions.

It was concluded from the regression models that the computer-derived measures predict a global rating criterion of potential ATC on-the-job success at least as well as the

over-the-shoulder evaluations (Models 1, 2, and 3). Further, it was found that the over-the-shoulder evaluations are not as reliable as the computer-derived measures (Table 3). Since reliability is in general the upper bound for validity, using computer-derived measures would enhance the probability for higher validity. The computer-derived measures, it appears, can be substituted for the over-the-shoulder ratings and used to form a composite laboratory score. Model 4 demonstrates that four or five problems should be employed in forming the laboratory composite with unit weights of (i) 15, 15, 30, and 40, or (ii) 10, 10, 15, 40, and 25, respectively. These results have implications for other training programs where expert observations and ratings are [illegible] scoring. This study indicates that computer scoring can provide a more objective and reliable measurement, and this increased reliability provides potential for enhancing a program's validity. Further research in this area should include extended analysis of factor structures of the two measurement techniques (Table 10) in an attempt to reach an optimal scoring schema using both computer scoring and expert observation. Future research should also include a long range validity study to determine which measurement technique is more valid in predicting on-the-job success.

TABLE 10. Factor Analyses of the Over-the-Shoulder and Computer-Derived Measures

Computer-Derived Measures		Over-the-Shoulder	
Measure	*Loading	Measure	*Loading
FACTOR 1	(Conflicts)	FACTOR 1	(Conflicts)
1. Conflicts	.7843	1. System Error	.8188
		2. System Error	.6441
FACTOR 2 (A/C Workload)		3. System Error	.6088
		4. System Error	.6836
1. Start Point Delays	.4214		
2. No. A/C Handled	.8762	FACTOR 2 (System Deviations)	
3. No. Completed Flights	.7304		
4. No. Arrivals	.6209	1. System Deviation	.5144
5. No. of Departures	.5347	2. System Deviation	.4163
		3. System Deviation	.5883
FACTOR 3 (Delays)			
		FACTOR 3 (Departures)	
1. Turn and Hold Delays	.4574		
2. A/C Time in System	.6302	1. Keep Them High	.7902
3. Handoff Delay	.4039		
		FACTOR 4 (A/C-Vectoring)	
FACTOR 4 (Communications)			
		1. Bad Vector	.5514
1. Air-to-Ground Contacts	.8253	2. No Reason Vector	.5108
2. No. Beacon Re-Idents	.4928	3. Holding	.4232
3. No. Heading Changes	.4291		
		FACTOR 5 (Arrivals)	
FACTOR 5 (A/C Direction Vectoring)			
		1. Missed Approach	.6057
1. No. Speed Changes	.7126		
2. No. Altitude Changes	.5284		

*Only loadings of .400 or better were retained. (TABLE 10 continued)--

(128)

TABLE 10 continued→→→

Measure	Over-the-Shoulder *Loading
FACTOR 6 (A/C Direction)	
1. Improper Coord.	.5511
2. Routing Error	.4294
3. Position A/C Incorrect	.4129
4. Altitude Verification	.5423
FACTOR 7 (Communications)	
1. Traffic	.4374
2. Remarks	.4863

References.

1. Arad, B., B. T. Golden, J. E. Grambart, C. E. Mayfield, and H. R. vanSaun: Control Load, Control Capacity and Optimal Sector Design. System Design Team, Washington, D.C., & Research Division, National Aviation Facilities Experimental Center (NAFEC), Atlantic City, New Jersey, Report No. RD-64-16, 1963.

2. Alexander, L. T.: Terminal Air Traffic Control System. SYSTEM DEVELOPMENT CORPORATION MAGAZINE, Santa Monica, California, July 1962.

3. Alexander, L. T.: Terminal Air Traffic Control Follow-on Research. System Development Corp., Santa Monica, California, Report TM-639/007/00, 1963.

4. Alexander, L. T., and M. Ash: Terminal Air Traffic Control: A Laboratory Model for Man-Machine System Research. System Development Corp., Santa Monica, California, Report SP-1016, 1962.

5. Alexander, L. T., and A. S. Cooperband: A Laboratory Model for Systems Research: A Terminal Air Traffic Control System. System Development Corp., Santa Monica, California, Report TM-639, 1961.

6. Alexander, L. T., and A. S. Cooperband: Schematic Simulation: A Technique for the Design and Development of a Complex System. HUMAN FACTORS, 6:87-82, 1964a.

7. Alexander, L. T., and A. S. Cooperband: The Effect of Rule Flexibility on System Adaptation. HUMAN FACTORS, 6:209-220, 1964b.

8. Alexander, L. T., and E. H. Porter: Terminal Air Traffic Control and Problems of System Design. System Development Corp., Santa Monica, California, Report TM-639/008/00, 1963.

9. Alluisi, E. A.: Human Engineering Aspects of Air Traffic Control Systems. Ohio State Univ. Research Foundation, Columbus, Ohio, Final Quarterly Report, 1956.

10. Anderson, C. M., and C. E. Dowling: Evaluation by Simulation Techniques of a Proposed Traffic Control Procedure for the New York Metropolitan Area. Civil Aeronautics Administration Technical Development & Evaluation Center, Indianapoli Indiana, Report No. 245, 1954.

11. Anderson, C. M., T. E. Armour, et al.: Dynamic Simulation Tests of Baltimore Friendship Airport at Increased Traffic Densities. Civil Aeronautics Administration Technical Development & Evaluation Center, Indianapolis, Indiana, 1957.

12. Armour, T. E., A. N. Johnson, T. K. Vickers, and R. S. Miller: Simulation Tests of the Factors Affecting IFR Traffic Capacity at Chicago O'Hare Airport. Civil Aeronautics Administration Technical Development & Evaluation Center, Indianapolis Report No. 341, 1958.

13. Baker, R. E., A. L. Grant, and T. K. Vickers: Development of a Dynamic Air Traffi Control Simulator. Civil Aeronautics Administration Technical Development & Evaluation Center, Indianapolis, Report No. 191, 1953.

14. Berkowitz, S. M., and R. R. Doering: Analytical and Simulation Studies of Several Radar-Vectoring Procedures in the Washington, D.C. Terminal Area. Civil Aeronautics Administration Technical Development & Evaluation Center, Indianapolis, Franklin Institute Lab Report No. 222, 1954.

15. Berkowitz, S. M., and E. L. Fritz: Analytical and Simulation Studies of Terminal-Area Air Traffic Control, Summary of Joint FIL-TDEC Simulation Activities in Air Traffic Control. Civil Aeronautics Administration Technical Development & Evaluation Center, Indianapolis, Franklin Institute Lab Report No. F-2384, 1955.

16. Berkowitz, S. M., E. L. Fritz, and R. S. Miller: Summary of Joint FIL-TDC Simulation Activities in Air Traffic Control. Civil Aeronautics Administration Technical Development & Evaluation Center, Indianapolis, Report No. 297, 1957.

17. Bottomley, D., R. E. Hansen, T. R. Johnson, H. T. Rohland, S. B. Rossiter, and E. H. Wright: Dynamic Simulation Study and Evaluation of a Proposed Air Traffic Procedural Plan and Control Equipment for the Washington, D.C. Area. FAA NAFEC, Atlantic City, New Jersey, Bureau of Research & Development Report No. 101-112V, 1962.

18. Brinton, J. H., Jr., and R. S. Miller: Summary of Joint FIL-FAA Research in Air Traffic Control During Period From September 1958 to January 1961. Franklin Institute Lab, Philadelphia, Pennsylvania, Report No. F-A2221, 1961.

19. Buckley, E. P., W. F. O'Connor, and T. Beebe: A Comparative Analysis of Individual and System Performance Indices for the Air Traffic Control System. FAA NAFEC, Atlantic City, New Jersey, NA-69-40, 1969.

20. Buckley, E. P., and R. H. Rood: Tower Training Using Radar Simulators. FAA NAFEC, Atlantic City, New Jersey, NA-77-12-LR, 1977.

21. Buckley, E. P., and R. H. Rood: CPM PROBE Experiment on Performance Information Feedback. FAA NAFEC, Atlantic City, New Jersey, NA-77-18-LR, 1977.

22. Buckley, E. P., K. House, and R. Rood: Development of a Performance Criterion for Air Traffic Control Personnel Research Through Air Traffic Control Simulation. FAA NAFEC, Atlantic City, New Jersey, Report FAA-RD-78-71, 1978.

23. Chapman, R. L., W. C. Biel, J. L. Kennedy, and A. Newell: The Systems Research Laboratory and Its Program. Rand Corp., Santa Monica, California, Report RM-890, 1952.

24. Cooperband, A. S., and L. T. Alexander: The Detection of Compound Motion. System Development Corp., Santa Monica, California, Report SP-1964/001/00, 1965.

25. Cooperband, A. S., L. T. Alexander, and H. S. Schmitz: Test Results of the Terminal Air Traffic Control Laboratory System. System Development Corp., Santa Monica, California, Report TM-639/004/00, 1963.

26. Faison, W. E., and A. L. Sluka: Dynamic Simulation Studies of Pictorial Navigation Displays as Aids to Air Traffic Control in a High-Density Terminal Area and a Medium-Density Terminal Area. FAA NAFEC, Atlantic City, Bureau of Research & Development Report, 1961.

27. Fitts, P. M., and M. I. Prosner: Human Performance. Belmont, California: Brooks/Cole Pub. Co., 1967.

28. Franklin Institute: Pilot Experiments Concerning Air Traffic Control Decision Making (internal document prepared for the FAA Bureau of Research & Development). Philadelphia, Pennsylvania, 1960.

29. Grant, E. E., S. L. O'Connell, K. L. Stoker: The Human Factors Laboratory. System Development Corp., Santa Monica, California, Report TM-561, 1960.

30. Harman, H. H.: The Systems Simulation Research Laboratory. System Development Corp., Santa Monica, California, Report TM-498, 1960.

31. Hagay, J. A.: The Development of a Procedure for Evaluating the Proficiency of Air Route Traffic Controllers. Civil Aeronautics Administration Division of Research, Washington, D.C., Report No. 83, 1949.

32. Henry, J. H., M. E. Kamrass, J. Orlansky, T. C. Rowan, J. String, and R. E. Reichenbach: Training of U.S. Air Traffic Controllers. FAA Office of Personnel & Training, Washington, D.C., Report R-206, 1975.

33. Hett, W. D., W. D. Coulopolos, W. Wolff, and R. A. Kowalski: Package D Testing With Air Movements Data Only (DAMDOT). MITRE Corp., Bedford, Massachusetts, final report No. TM-3339, 1962.

34. Hixson, W. C., G. A. Harter, C. E. Warren, J. D. Cowan, Jr.: An Electronic Radar Target Simulator for Air Traffic Control Studies. Aero Medical Lab, Wright Air Development Center, Wright-Patterson AFB, Ohio, TR-54-569, 1954.

35. Howell, W. C., R. T. Christy, and R. G. Kinkade: System Performance Following Radar Failure in a Simulated Air Traffic Control Situation. Wright Air Development Center, Wright-Patterson AFB, TR-59-573, 1959.

36. Jacobs, J. F.: Practical Evaluation of Command and Control Systems. MITRE Corp Bedford, Massachusetts, Report MTP-7, 1965.

37. Johnson, B. E., A. C. Williams, Jr., and S. N. Roscoe: A Simulator for Studying Human Factors in Air Traffic Control Systems. Univ. of Illinois National Research Council Committee on Aviation Psychology, Urbana, Illinois, Report No. 11, 1951.

38. Jolitz, G. D.: Evaluation of a Mathematical Model for Use in Computing Control Load at ATC Facilities. FAA Systems Research & Development Service, Atlantic City, New Jersey, Report No. RD-65-69, 1965.

39. Kidd, J. S.: A Comparison of Two Methods of Controller Training in a Simulated Air Traffic Control Task: A Study in Human Engineering Aspects of Radar Air Traffic Control. Wright Air Development Center, Wright-Patterson AFB, Report TR 58-449, 1959, 1961.

40. Kidd, J. S.: A Comparison of One-, Two-, and Three-Man Control Units Under Various Conditions of Traffic Input Rate. Wright Air Development Center, Wright-Patterson AFB, Report No. TR-59-104, 1959, 1961.

41. Kidd, J. S.: A Summary of Research Methods, Operator Characteristics, and System Design Specifications Based on the Study of a Simulated Radar Air Traffic Control System. Wright Air Development Center, Wright-Patterson AFB, Report No. TR 59-236 1959.

42. Kidd, J. S.: Some Sources of Load and Constraints on Operator Performance in a Simulated Radar Air Traffic Control Task. Wright Air Development Center, Wright-Patterson AFB, Report No. TR 60-612, 1961.

43. Kidd, J. S., and J. J. Hooper: Division of Responsibility Between Two Controllers and Load Balancing Flexibility in a Radar Approach Control Team. A Study in Human Engineering Aspects of Radar Air Traffic Control. Wright Air Development Center, Wright-Patterson AFB, Report No. TR 58-473, 1959.

44. Kidd, J. S., and R. G. Kinkade: Air Traffic Control System Effectiveness as a Function of Division of Responsibility Between Pilots and Ground Controllers: A Study in Human Engineering Aspects of Radar Air Traffic Control. Wright Air Development Center, Wright-Patterson AFB, Report No. TR 58-113, 1958.

45. Kidd, J. S., M. W. Shelly, G. Jeantheau, and P. M. Fitts: The Effect of Enroute Flow Control on Terminal System Performance: A Study in Human Engineering Aspects of Radar Air Traffic Control. Wright Air Development Center, Wright-Patterson AFB, Report No. TR 57-663, 1958.

46. Kidd, J. S., R. G. Kinkade, F. C. Ichniowski, W. G. Widhelm, and S. Urback: Overview of Project No. 102-8X. Aircraft Armaments, Inc., Cockeysville, Maryland, Report ER-3238, Vol. I, 1963.

47. Kidd, J. S., W. N. Widhelm, F. C. Ichniowski, and R. G. Kinkade: Method Development for ATC System Simulation Research. Aircraft Armaments, Inc., Cockeysville, Maryland, Report ER-3238, Vol. II, 1963.

48. Kraft, C. L.: A Broad-Band Blue Lighting System for Radar Approach Control Centers: Evaluation and Refinements Based on Three Years of Operational Use. Wright Air Development Center, Wright-Patterson AFB, Report No. TR-56-71, 1956.

49. Kraft, C. L., and P. M. Fitts: A Broad-Band Lighting System for Radar Air Traffic Control Centers. Wright Air Development Center, Wright-Patterson AFB, Report No. TR 53-416, 1954.

50. McKenzie, R. E.: An Exploratory Study of Psychophysiological Measurements as Indicators of Air Traffic Control Sector Workload. FAA NAFEC, Atlantic City, New Jersey, Project 157-524-03R, 1966.

51. Miller, R. S.: Summary of Joint FIL-TDS Research in Air Traffic Control During Period From April 1957 to September 1958. Franklin Institute, Philadelphia, 1958.

52. Muller, P. F., Jr., R. C. Sedorsky, A. J. Slevinsky, E. A. Alluisi, and P. M. Fitts: The Symbolic Coding of Information on Cathode Ray Tubes and Similar Displays. Aero Medical Lab, Wright-Patterson AFB, Proj. No. 7192, 1955.

53. Paul, L. E., and E. P. Buckley: Human Factors Evaluation of a Large Screen Radar Display. FAA NAFEC, Atlantic City, New Jersey, Report No. RD-66-105, 1967.

54. Ratner, R. S., J. O. Williams, M. B. Glaser, and S. E. Stuntz: The Air Traffic Controller's Contribution to ATC System Capacity in Manual and Automated Environments, Vol. I (Summary reports). Stanford Research Institute, Menlo Park, California, Report No. FAA-RD-72-63, 1972.

55. Ratner, R. S., J. O. Williams, M. B. Glaser, and S. E. Stuntz: The Air Traffic Controller's Contribution to ATC System Capacity in Manual and Automated Environments, Vol. II (Appendices). Stanford Research Institute, Menlo Park, California, Report No. FAA-RD-72-63, 1972.

56. Ratner, R. S., J. O. Williams, M. B. Glaser, and S. E. Stuntz: The Air Traffic Controller's Contribution to ATC System Capacity in Manual and Automated Environments, Vol. III (Terminal operations). Standard Research Institute, Menlo Park, California, Report No. FAA-RD-72-63, 1972.

57. Schipper, L. M., J. Versace, C. L. Kraft, and J. C. McGuire: Human Engineering Aspects of Radar Air Traffic Control. Aero Medical Laboratory, Wright Air Development Center, Wright-Patterson AFB, Proj. No. 7192, 1956.

58. Schipper, L. M., and J. Versace: Human Engineering Aspects of Radar Air Traffic Control: I. Performance in Sequencing Aircraft for Landing as a Function of Control Time Availability. Wright Air Development Center, Wright-Patterson AFB, Report No. TR 56-57, 1956.

59. Schipper, L. M., J. Versace, C. L. Kraft, and J. C. McGuire: Human Engineering Aspects of Radar Air Traffic Control: II and III. Experimental Evaluations of Two Improved Identification Systems Under High Density Traffic Conditions. Wright Air Development Center, Wright-Patterson AFB, Report No. 56-58, 1956.

60. Schipper, L. M., J. Versace, C. L. Kraft, and J. C. McGuire: Human Engineering Aspects of Radar Air Traffic Control: IV. A Comparison of Sector and In-Line Control Procedures. Wright Air Development Center, Wright-Patterson AFB, Report No. TR 56-69, 1956.

61. Schipper, L. M., J. S. Kidd, M. W. Shelly, and A. F. Smode: Terminal System Effectiveness as a Function of the Method Used by Controllers to Obtain Altitude Information: A Study in Human Engineering Aspects of Radar Air Traffic Control. Wright Air Development Center, Wright-Patterson AFB, Report No. TR 57-278, 1957.

62. Schipper, L. M., C. L. Kraft, A. F. Smode, and P. M. Fitts: The Use of Displays Showing Identity Versus No-Identity: A Study in Human Engineering. 1957.

63. Slattery, H. F.: Air Traffic Control Simulation Program Conference, Report of the Chairman. FAA NAFEC, Atlantic City, New Jersey, 1965.

64. Sluka, A. L.: Dynamic Simulation Studies of Pictorial Navigation Displays as Aids to Air Traffic Control in a Low-Density Terminal Area and in an Enroute Area. FAA NAFEC, Atlantic City, New Jersey, final report, 1963.

65. Test and Experimentation Division: A Report on Dynamic Simulation Tests and Study of the Bureau of Air Traffic Management Plan for the Positive Control of Air Traffic on an Area Basis in the Chicago and Indianapolis Air Route Traffic Control Center Areas. FAA NAFEC, Atlantic City, New Jersey, 1960.

66. Versace, J.: The Effect of Emergencies and Communications Availability With Differing Entry Rates: A Study in Human Engineering Aspects of Radar Air Traffic Control. Wright Air Development Center, Wright-Patterson AFB, Report TR 56-70, 1956.

67. Vickers, T. K.: Development of Traffic Control Procedures for Tactical Airlift Operation. Civil Aeronautics Administration Technical Development & Evaluation Center, Indianapolis, Report No. 235, 1954.

68. Vickers, T. K.: Simulation Tests for Army Air Traffic Control. Civil Aeronautics Administration Technical Development & Evaluation Center, Indianapolis, Report No. 298, 1957.

69. Vickers, T. K.: The Use of Simulation in ATC Systems Engineering. Civil Aeronautics Administration Technical Development & Evaluation Center, Indianapolis, Report No. 410, 1959.

70. Vickers, T. K.: Air Traffic Control Simulation Program Conference, Report of the Chairman. FAA NAFEC, Atlantic City, New Jersey, 1965.

71. Vickers, T. K., and R. S. Miller: Recent Developments in the Simulation of Terminal Area and Enroute Area Air Traffic Control Problems. IRE TRANSACTIONS ON AERONAUTICAL AND NAVIGATIONAL ELECTRONICS ANE-3:51-55, 1956.

BOYCE, Lt. D.G. and BELEC, Capt. B.E., Canadian Forces Personnel, Applied Research Unit, Willowdale, Ontario.

ATTITUDES TOWARD WOMEN'S ROLES AND ORGANIZATIONAL COMMITMENT WITHIN THE CANADIAN FORCES. (Fri A.M.).

This paper presents preliminary analyses of data currently being collected by the Canadian Forces in connection with the Women-in-Near-Combat Environments (WINCE) evaluations. Using Spence and Helmreich's Attitudes Towards Women Scale (AWS) and Cotton's Military Ethic Scale (MES), the attitudes toward women's roles in society of selected samples of CF personnel are examined in relation to a number of organizational commitment variables. In addition, the relationships between a number of biographic and demographic factors, and traditional versus equalitarian/profeminist attitudes are investigated. The theoretical and practical implications of these analyses are addressed.

ATTITUDES TOWARD WOMEN'S ROLES: A PRELIMINARY ANALYSIS OF CANADIAN FORCES' PERSONNEL*

Lieutenant Diane G. Boyce and Captain Brian E. Belec

Canadian Forces Personnel Applied Research Unit,
Toronto, Canada

INTRODUCTION

The Canadian Forces (CF) are currently conducting a series of evaluations in order to assess the impact, on operational effectiveness, of integrating women into near-combat environments**. These evaluations are being carried out as a result of two essential issues. The first of these was the promulgation of the Canadian Human Rights Act (CHRA) (March, 1978) which prohibits exclusion from occupational areas on the basis of sex - unless the employer can establish that such exclusion is based on a bona fide occupational requirement. The second issue concerns difficulties in sustaining current and future (projected) manpower levels. Socio-demographic analyses indicate that the traditional recruit population (i.e., males 17 to 24 years of age) will not provide an adequate source of personnel for projected manpower requirements into the 1980s and 1990s (Tierney, 1979; Tierney and Pinch, 1980). Given these legal and demographic trends, the Women-in-Near-Combat Environments (WINCE) evaluations are designed to determine if bona fide occupational requirements exist that might restrict employment of women in nontraditional areas, and the extent to which women's participation may be expanded within the military.

Research conducted in the United States has shown that, from the mid-1940s to the early 1970s, women consitituted less than two percent of total military strength, and were restricted largely to health care and administrative/clerical occupations. It has only been in recent years that women have been employed as electrical equipment repairmen, aircraft maintenance personnel, communications experts, and telephone linemen - among the many other nontraditional career fields (Binkin & Bach, 1977; Landrum, 1978). Similar trends have been visible within the Canadian military. Until recently, Canadian women were recruited to fill primarily traditional women's occupations. However, as a result of the socio-demographic and legal issues noted above, women's participation within the military has been expanded to include employment in nontraditional environments. One area which may have important implications for the integration of women concerns the attitudes toward women's roles in society held by CF personnel, both male and female, who may/will be employed in near-combat environments.

Attitudes toward women in society have generally been examined in terms of traditional versus profeminist/egalitarian attitudes (Spence & Helmreich, 1972). The traditional orientation emphasizes the "stereotypic" role of women in society,

---

*The views and opinions in this paper are those of the authors and not necessarily those of the Department of National Defence. The authors wish to acknowledge the comments of Major F.C. Pinch on earlier drafts of this paper.

**The term "Near-Combat Environments" refers to employment in land combat support units, in noncombatant ships and aircraft, and in isolated settings.

which distinguishes the kinds of activities considered appropriate for men and women. Conversely, the profeminist/egalitarian perspective downplays sex-typed roles and emphasizes equality of men and women.

One instrument that has been frequently used to measure traditional versus profeminist/egalitarian orientations is the Attitudes Toward Women Scale (AWS) developed by Spence and Helmreich (1972; 1973). The AWS was recently employed in research conducted at the three Canadian Military Colleges (CMCs) and several Canadian civilian universities (Prociuk, 1980). In addition, it was used in research conducted at three United States Military Academies (Durning, 1978). Thus, the use of the AWS in this study permits comparative analysis with the findings of Prociuk (1980) and Durning (1978).

Research over the past decade has shown that such socio-demographic factors as age (Thornton & Freedman, 1979; Mason et al, 1976), education (Thornton & Freedman, 1979; Mason et al, 1976), language group affiliation (Schreiber, 1975; Pinch, 1978), and community size (Pinch, 1978) are associated with subjects' attitudes toward women in society. This body of research suggests that better educated women, from more urbanized and industralized centres, hold more egalitarian attitudes toward women's roles in society. Research on the relationships between age and sex-role attitudes (Thornton & Freedman, 1979; Mason et al, 1976) and between language group affiliation and sex-role attitudes (Lipset, 1968; Schreiber, 1975; Pinch, 1978; Prociuk, 1980) has proved contradictory.

Accordingly, the relationships between age, level of education, language group affiliation and community size, and orientations toward the roles of women in society, for CF personnel, were examined.

In essence, then, this paper looks at the attitudes toward women's roles in society, and the relationshps between these attitudes and selected demographic variables for two groups (one female, one male) of CF personnel.

METHOD

For purposes of this paper, data collected from 405 Canadian Forces personnel (277 servicewomen and 128 servicemen) were analyzed. The female sample consisted of all the servicewomen who were preselected to participate in the WINCE evaluations. The male sample included 38 members serving on board HMCS Cormorant and 90 members serving with 4 Service Battalion. Data on the males were collected prior to the arrival of females at their respective units.

Two survey instruments were administered: the Women's Trials Biographical Questionnaire (WTBQ) and the Women's Trials Attitudinal Questionnaire (WTAQ). The WTBQ is designed to determine the socio-demographic background of the respondents, as well as some basic attitudes concerning their association with the Canadian Forces. The WTAQ is designed to measure respondents' attitudes toward women in society [as measured by the 25-item version of the Attitudes Toward Women Scale (AWS) developed by Spence, Helmreich & Stapp (1973)] and within the CF. The AWS is a Likert-type scale that measures attitudes concerning the rights, roles, and privileges women have, or ought to have, in such areas as vocational and educational choice, dating and courtship, sexual behaviour, and marital roles (Spence & Helmreich, 1972). More specifically, it deals with such factors as equality of opportunity for women in educational and vocational spheres, and reflects attitudes concerning social/sexual relationships between men and women, and what constitutes "ladylike behaviour" (Spence and Helmreich, 1972).

The AWS is considered a fairly reliable instrument. Measures of internal consistency (coefficient alpha) (Nie & Hull, 1977), taken in previous research, have produced coefficients of .838 for Queen's University students and .841 for University of Manitoba students (Prociuk, 1980).

Surveys were administered at the respondents' current places of duty by Base Personnel Selection Officers, under guidelines established by the CF Directorate of Personnel Applied Research, Ottawa. Coding of responses was completed by the research staff of the CF Personnel Applied Research Unit, Toronto.

ANALYSIS AND DISCUSSION

Each item on the AWS was scored from 0 to 3, with 0 reflecting the most traditional attitude and 3 indicating the most liberal/profeminist attitude. The respondent's score on the scale was the sum of the scores on each of the 25 items. Thus, scores ranged from 0 to 75.

Measures of internal consistency (coefficient alpha) for the two groups of CF personnel produced coefficients of .798 for servicewomen and .829 for servicemen, both of which are comparable to those reported by Prociuk (1980). Thus it would appear that the AWS was a fairly reliable instrument for these two samples.

Prociuk (1980) found that the AWS mean scores of cadets at either Royal Military College (RMC), College Militaire Royale (CMR), or Royal Roads Military College (RRMC), reflect significantly more egalitarian attitudes toward women's roles in society than those expressed by male cadets at either the United States Military Academy (USMA), United States Air Force Academy (USAFA), or Unites States Naval Academy (USNA). Appendix 1 lists applicable means and standard deviations. Similar analyses indicated that there are no significant differences in attitudes toward women's roles in society between CF servicemen and CMC cadets (either RMC, CMR, or RRMC). AWS mean scores of CF servicemen, like those of CMC cadets, reflect significantly more egalitarian attitudes toward women's roles in society than those expressed by male cadets at the US Military Academies (either USMA, USAFA, or USNA); (e.g., CF servicemen (48.37) versus USMA males (42.34); $t(1386) = 6.09$, $p < .001$). As indicated by Prociuk (1980), however, any firm conclusions must be tempered by the fact that the US Military Academies' data were collected in 1976 and general societal changes in attitudes toward women may have contributed to the obtained findings.

In addition, the AWS mean scores for CF servicemen versus civilian college male students were compared. No significant differences between CF servicemen and civilian male students (for any of the comparisons) were found. This is in line with Prociuk's (1980) finding which indicated no significant AWS mean score differences between CMC cadets and civilian college male students.

Next, AWS mean scores of CF servicewomen were compared to those of female cadets at each of the US Military Academies. No significant differences in attitudes toward women's roles were found for these groups. However, the AWS mean scores of CF servicewomen reflect significantly less egalitarian attitudes toward women's roles in society than those expressed by female civilian college students (either Queen's, Manitoba, or CEGEPs) (e.g., CF servicewomen (55.66) versus Queen's females (60.42); $t(456) = -4.51$, $p < .001$). Differences in the educational

levels of these two samples may explain, in part, the more egalitarian orientations of the female Canadian college students as compared to the less educated CF servicewomen. Similar results were reported by Mason et al (1976).

Obvious differences in AWS mean scores are indicated in all of the male versus female comparisons. CF servicemen (48.37) reflect significantly more traditional attitudes than those expressed by CF servicewomen (55.66); $t(369) = -7.43$, $p < .001$. Similarly, CF servicemen express significantly less egalitarian attitudes than female cadets at US Military Academies and female civilian college students. Finally, CF servicewomen reflect significantly more egalitarian attitudes than those expressed by male cadets at both the CMCs and US Military Academies, and male civilian college students. Appendix 2 shows all relevant statistics.

As reported earlier, CF servicewomen expressed significantly more egalitarian attitudes toward the rights and roles of women than CF servicemen. While cognizant of the importance of gender differences in determining attitudes toward women's roles (Erskine, 1971), it is essential to consider other factors that might influence the obtained results.

As the average level of education for the females was significantly different (greater) from that of the males ($t(399) = 3.26$, $p = .001$), the relationship between level of education and attitudes toward women's roles was examined separately for each sex. The Pearson product-moment correlation between level of education and attitudes toward women's roles, for CF servicewomen, was .16, $p = .005$. This finding is consistent with the results of previous research which suggest that egalitarian sex-role attitudes are positively related to level of education among women (Mason et al, 1976). This relationship was not significant for CF servicemen.

For the entire sample of CF personnel, the Pearson product-moment correlation between the size of community during adolescence and attitudes toward women's roles was found to be .10, $p < .04$. Although the small absolute value of this correlation means that any conclusions as to trend must remain tentative, the relationship is in line with the pattern found in previous research, where levels of urbanization and industralization were shown to be positively related to egalitarian attitudes toward women.

While Thornton and Freedman (1979) reported that there was an inverse relationship between age and egalitarian attitudes toward women, other research has shown either that there is a positive relationship between age and egalitarian sex role attitudes or that age has no effect on attitudes (Mason et al, 1976). The results obtained here are consistent with the latter finding, i.e., no significant relationships were found between these two variables for either males or females.

Examination of AWS mean scores reveals no significant differences between Anglophone and Francophone servicemen. This finding is consistent with the results obtained on male cadets at CMCs by Prociuk (1980). However, Anglophone servicewomen (54.6) were more egalitarian than their Francophone counterparts (51.0); ($t(275) = 2.21$, $p < .03$).

While Anglophones have historically been considered to hold more contemporary views on a wide variety of issues than their Francophone peers (Lipset, 1968), recent studies have shown either that Francophones are more egalitarian than Anglophones (both Canadian and American) (Schreiber, 1975; Pinch, 1978), or that there are no differences between Anglophones and Francophones in their attitudes toward women's roles (Prociuk, 1980). Given these contradictory findings, the problem becomes one of determining what factors might contribute to the differences in attitudes between Anglophone and Francophone servicewomen.

Comparisons of mean scores indicate significant differences between Anglophone and Francophone servicewomen with respect to age (26 years versus 24 years; $t(273) = 3.06$, $p=.002$), level of education (grade 12-13 versus grade 10-11; $t(275) = 2.51$, $p < .02$), and size of community (15,000 population versus 7,500 population; $t(273) = 2.72$, $p < .01$). Thus, in this sample, the Anglophones were older, better-educated, and from larger communities than the Francophones. Future research will ascertain the relative importance of each of these variables in determining the differences in attitudes toward women's roles for CF servicewomen.

These differences between Anglophone and Francophone servicewomen might suggest that Francophones would be more reluctant to assume nontraditional employment or, indeed, to remain in the nontraditional settings to which they are assigned, than their Anglophone counterparts. This was not borne out, however, as no differences were found between their decisions regarding participation in the trials. In other words, servicewomen who expressed traditional attitudes toward women's roles were as willing to assume nontraditional employment within the Canadian military as those servicewomen who expressed egalitarian/profeminist attitudes.

This suggests that factors other than attitudes toward women's roles may be operating. One possible set of factors revolves around differences in "military ethos" (Cotton, 1979; Prociuk, 1980); that is, between an "occupational orientation" (i.e., the roles and obligations of military personnel are limited to those relevant to civilian employment settings, and involve formal, contractual arrangements and limited liability) and a "vocational orientation" (i.e., military personnel have unlimited liability and an "implied contract" and should put service and duty first, regardless of personal consequences) toward military service. It may be that vocationally oriented servicewomen are more likely to assume nontraditional military employment than those who adhere to an occupational orientation. This issue remains to be addressed in subsequent research.

SUMMARY AND CONCLUSIONS

This paper presents a preliminary analysis of the data currently being collected by the CF in connection with the Women-in-Near-Combat Environments evaluations. Using the Attitudes Toward Women Scale, the attitudes of a sample of Canadian Forces personnel, toward the roles of women in society, were compared to a number of other military and civilian groups. Also, the relationships between several biographic and demographic variables and these attitudes were investigated.

The results indicated that the attitudes toward women, expressed by CF servicemen, were comparable with those expressed by other militiary and civilian male samples. The attitudes of CF servicewomen were found to be similar to those expressed by other military female samples, but tended to be more traditional than those of the selected civilian female samples.

The obtained results indicated that CF servicemen were more traditional in their attitudes toward women in society than CF servicewomen. One practical implication of this finding, with respect to the integration of women into near-combat environments, is the possible reluctance of servicemen to readily accept servicewomen in nontraditional roles. Such a lack of acceptance may have a negative effect on unit morale and operational effectiveness. In addition, the analysis revealed a significant, but weak, positive relationship between size of community and attitudes toward women for CF personnel.

Finally, for servicewomen, it was found that Anglophone servicewomen tended to be older, better-educated, and from larger communities than their Francophone counterparts. Moreover, Anglophones were more egalitarian than Francophones in attitudes toward women's roles. Despite no differences in respondents' willingness to participate in the WINCE evaluations, one implication of this finding is that Francophone servicewomen may have more problems adapting to nontraditional roles than Anglophone servicewomen. Subsequent research will address this issue.

REFERENCES

Binkin, M., & Bach, S.J. Women and the military. Washington, D.C.: The Brookings Institution, 1977.

Boyce, D.G., Tierney, E.C., & Pinch, F.C. Attitudes toward women's roles: a preliminary analysis of Canadian Forces servicewomen. Paper presented at the Annual Meeting of the Human Factors Association of Canada, Point Ideal, September, 1980.

Cotton, C.A. Military attitudes and values of the army in Canada (Report 79-5). Toronto: Canadian Forces Personnel Applied Research Unit, 1979.

Durning, K.D. Women at the Naval Academy: an attitude survey. Armed Forces and Society, 1978, 4, 569-588.

Erskine, H. The polls: women's role. Public Opinion Quarterly, 1971, 35, 275-290.

Landrum, C.S. Role of women in today's military. In Keeley, J.B. (ed.) The all-volunteer force and American society. Charlottesville: The University Press of Virginia, 1978.

Lipset, S.M. Revolution and counter-revolution: change and persistence in social structures. New York: Basic Books, Inc., 1968.

Mason, K.O., Czajka, J.L., & Arber, S. Change in U.S. women's sex-role attitudes, 1964-1974. American Sociological Review, 1976, 41, 573-596.

Nie, N.H., & Hull, C.H. SPSS batch release 7.0 update manual. Toronto: McGraw-Hill Book Company, 1977.

Pinch, F.C. Economic, social and cultural influences of military participation in two Canadian provinces (Report 75-6). Toronto: Canadian Forces Personnel Applied Research Unit, 1975.

Pinch, F.C. The social bases for expansion of women's roles into military combat in Canada and the United States. College Park: University of Maryland, 1978.

Prociuk, T.J. Women at Canadian Military Colleges: a survey of attitudes (1980) (Dept. Manuscript 80-1). Kingston: Royal Military College, 1980.

Schreiber, E.M. The social bases of opinions on women's roles in Canada. Canadian Journal of Sociology, 1975, 1(1), 61-74.

Spence, J.T., & Helmreich, R. The attitudes toward women scale: an objective instrument to measure attitudes toward the rights and roles of women in contemporary society. JSAS Catalog of Selected Documents in Psychology, 1972, 2.

Spence, J.T., & Helmreich, R. Who likes competent women? competence, sex-role congruence of interests and subjects' attitudes toward women as determinants of interpersonal attraction. Journal of Applied Social Psychology, 1972, 2, 197-213.

Spence, J.T., & Helmreich, R. Ratings of self and peers on sex-role attributes and their relation to self-esteem and conceptions of masculinity and femininity. Journal of Personality and Social Psychology, 1975, 32, 29-39.

Spence, J.T., Helmreich, R., & Stapp, J. A short version of the Attitudes Toward Women Scale (AWS). Bulletin of the Psychonomic Society, 1973, 2, 219-220.

Tierney, E.C. An overview of the socio-demographic changes and recruiting trends in the forces: 1968 to 1978 (Report 79-1). Toronto: Canadian Forces Personnel Applied Research Unit, 1979.

Tierney, E.C., & Pinch, F.C. Military implications of socio-demographic and related changes in the 1980's and 1990's (Working Paper 80-4). Toronto: Canadian Forces Personnel Applied Research Unit, 1980.

Thornton, A., & Freedman, D. Changes in sex-role attitudes of women, 1962-1977: evidence from a panel study. American Sociological Review, 1979, 44, 831-842.

APPENDIX 1

ATTITUDES TOWARD WOMEN SCALE SCORES FOR

CANADIAN MILITARY COLLEGES,

UNITED STATES MILITARY ACADEMIES,

CANADIAN CIVILIAN COLLEGES,

AND CF PERSONNEL

COLLEGE	MALES			FEMALES		
	N	Mean	S.D.	N	Mean	S.D.
CANADIAN MILITARY COLLEGES						
RMC	380	48.95	9.99			
CMR	296	48.34	9.60			
RRMC	188	48.46	9.91			
UNITED STATES MILITARY ACADEMIES						
USMA[a]	1277	42.34	9.91	115	57.38	9.96
USAFA[b]	367	43.10	12.38	42	54.09	12.35
USNA[c]	825	41.67	8.90	62	54.47	9.82
CANADIAN CIVILIAN COLLEGES						
Queen's	132	51.74	10.72	194	60.42	8.56
Manitoba	88	47.81	11.53	131	60.30	8.99
CEGEPs	100	49.33	11.09	112	60.90	8.35
CANADIAN FORCES PERSONNEL						
	111*	48.37	9.81	264*	55.66	8.11

[a]United States Military Academy; [b]United States Air Force Academy; [c]United States Naval Academy

*Variations from total sample sizes accounted for by missing data.

Source: Data for U.S. Military Academy cadets from Durning (1978; p. 578). Data for CMCs and Canadian civilian colleges from Prociuk (1980, p.34). Data for CF personnel from CFPARU, RIS (WINCE) File (September, 1980).

APPENDIX 2

TABLE OF t-STATISTICS
FOR COMPARISON OF AWS SCORES

COMPARISON GROUPS		df	t-STATISTIC	
CF servicemen vs.	RMC	489	- .531	n.s.
	CMR	405	.027	n.s.
	RRMC	297	- .08	n.s.
CF servicemen vs.	USMA males	1386	6.09	***
	females	224	- 6.53	***
	USAFA males	476	4.47	***
	females	151	- 3.18	**
	USNA males	934	7.44	***
	females	171	- 3.65	***
CF servicemen vs.	Queen's males	241	- 2.34	n.s.
	females	303	-13.39	***
	Manitoba males	197	.38	n.s.
	females	240	- 9.11	***
	CEGEP males	209	- .66	n.s.
	females	221	- 9.87	***
CF servicemen vs.	CF servicewomen	369	- 7.43	***
CF servicewomen vs.	RMC	642	9.40	***
	CMR	568	10.25	***
	RRMC	450	8.09	***
CF servicewomen vs.	USMA females	377	- 1.814	n.s.
	males	1539	19.88	***
	USAFA females	304	1.29	n.s.
	males	629	14.54	***
	USNA females	324	1.005	n.s.
	males	1087	22.92	***
CF servicewomen vs.	Queen's females	456	- 4.51	***
	males	394	3.93	***
	Manitoba females	393	- 5.06	***
	males	350	7.214	***
	CEGEP females	374	- 5.824	***
	males	362	5.85	***

**p < .01; ***p < .001; n.s. = not significant

BURKE, William P., Army Research Institute, Fort Benning, Georgia.

EFFECTS OF RESPIRATION CONTROL ON STRESS AND PERFORMANCE OF JUMPMASTERS (Wed P.M.)

The Jumpmaster Course at Fort Benning, Georgia, trains airborne personnel to conduct airdrops of men and equipment and features relatively stressful training jumps during which instructors grade the performance of students acting as jumpmasters for actual airdrops.

One class of Jumpmaster students was divided into pairs matched by rank and the members of each pair were randomly distributed into either an experimental or a control group. The experimental group was taught a method of respiration control to be used immediately before and during training jumps. The groups were then compared on heart rate, perceived stress, and grades received for performance as jumpmasters during training jumps.

The results showed that the experimental groups had significantly lower heart rates during the two night jumps of the course - jumps which, because of limited visibility, are somewhat more dangerous and therefore more stressful than daylight jumps. There were no other statistically significant differences between the groups.

EFFECTS OF RESPIRATION CONTROL ON STRESS AND PERFORMANCE OF JUMPMASTERS

William P. Burke

US Army Research Institute Field Unit
P.O. Box 2086, Fort Benning, Georgia, USA 31905

INTRODUCTION

Several laboratory studies have indicated that regulating the rate of breathing can serve to reduce physiological and psychological arousal in stressful situations. Brief training to reduce the breathing rate below normal in a stressful situation (the threat of receiving "painful" electric shocks), where the natural tendency of the body is to hyperventilate, has been shown by McCaul and his colleagues (1979) to reduce arousal, as measured by skin resistance and finger pulse volume (but not heart rate), as well as self reports of anxiety. Other studies varying both the rate and the depth of respiration have demonstrated that large decreases in heart rate can be produced by deep, slow breathing (Laird and Fenz, 1971; Westcott and Huttenlocher, 1961).

Respiration control has also figured prominently in a stress-management training program for novice skydivers. For several years, Walter Fenz of the University of Waterloo, in Canada, has been studying sports parachutists and has found that individuals who are rated as good performers at free-fall parachuting show a different pattern of heart rate response to the various activities preparatory to a parachute free-fall than do those rated as poor performers (Fenz, 1973; Fenz & Jones, 1972, 1974). He and his associates have found that, while poor performers show a relatively rapid increase of rate of heart beat from the time at which they first arrive at the airport on jump day, through boarding the aircraft, to reaching final altitude for the jump run, the heart rate of good performers peaks when boarding the aircraft and declines, thereafter, during engine warmup and the climb to final altitude. By the time of the jump run over the drop zone their heart rates have declined to the levels at which they stood when the individuals arrived at the airport that morning.

With reference to stress-management training, Dr. Fenz's most important project is one recently completed in collaboration with G. Brian Jones (see Fenz, 1975) in which they developed a program of mental and physical techniques (including deep, slow breathing) aimed at the control of involuntary stress reactions. In this research, they monitored the heart rate of individuals in two groups of novice parachutists, one group of which received the stress-management training and one of which did not, during their first jumps and during their first free-falls. The two groups showed heart rates which were significantly different after boarding the aircraft for both jumps. The most important aspect of these differences is that the heart rate response of the untrained jumpers resembled that of the poor performers in Fenz and Jones' earlier studies, while the heart rate of the trained group resembled that of the good performers in the prior studies. The training procedures had apparently prepared novice jumpers to approach the most critical of their early jumps in similar physical, and, presumably, mental states, to those of jumpers of far greater experience and ability.

These studies, from both the laboratory and the field, indicate that control of the rate and depth of respiration can lead to the control of both

psychological and physiological indicators of stress. Furthermore, the Fenz research suggests that heart rate, which has been shown in the laboratory to be influenced by patterns of breathing, may be related to quality of performance in stressful situations.

To test the effects of respiration control as a stress-management technique for use by the U.S. Army, an evaluation experiment was done with students going through the Jumpmaster Course at Fort Benning, Georgia. The Jumpmaster Course provides some of the most favorable conditions in which to test the effectiveness of stress-management techniques. Performance in this course is graded by instructors in an aircraft in flight with the students under extreme time pressure to complete a series of critical actions and inspections which prepare other men, equipment, and themselves for an airdrop. Only a brief association with jumpmaster students prior to their boarding the aircraft to make those graded training jumps is necessary to convince the observer that performance in the Jumpmaster Course is a stressful experience for most men. The stress appears to be the result of a combination of both harm anxiety and failure anxiety (Basowitz et al, 1955). Although most of the men in this course are experienced parachutists they have had little experience at working around the open door of an aircraft in flight and an initial apprehension about that experience must be overcome. Failure anxiety appears to be the prime stressor in the course, however, since these men, in the main, are highly motivated to do well in whatever training they attempt, and critical, irretrievable errors, causing their failure from the course, can be committed in seconds. The in-flight grading procedures used in this course are extensive and highly detailed and can serve as suitable performance measures against which to test the efficacy of any stress-management techniques.

METHOD

Procedure

The men used in this study were the students of one class of the Jumpmaster Training Course at Fort Benning, Georgia, and they ranged in rank from Captain to Private First Class The initial selection of individuals to be included in the experimental and control groups for the experiment proceeded as follows:

On inprocessing day, the first day of the course, the class was arranged by the jumpmaster cadre in descending order, by rank, with Captains listed first, followed by Lieutenants, senior NCOs, and so on, down to the lone PFC in the class. The students were then broken down into 2-man teams starting at the head of the list moving down and these teams, usually composed of men of equal rank, worked together throughout the course practicing and performing their skills. It was decided that subjects for this experiment should be matched by rank thereby to control to some extent for differences in airborne and service-related experiences. Accordingly, one man from each of the above mentioned 2-man teams was selected for inclusion in the experimental group which was to be taught the respiration control technique. Selection within teams was made by the flip of the coin, thereby providing random selection within matched pairs.

The Jumpmaster Course Curriculum. The course is conducted over two full weeks of training. The first week is primarily one of classroom training and hands-on practice at rigging and inspecting various parachute harness arrangements. The second week begins with a day of a written general knowledge exam

(15th)

and two hands-on harness inspection exams and consists from the second day on of flights and jumps over the drop zone.

There are five training jumps in the second week. First, there are both a day and a night orientation jump during which the students are taught to recognize, under both daylight and night-time conditions, the checkpoints on the ground indicating time and distance away from the drop zone. Every student makes a parachute jump at the conclusion of each of those orientation flights which will be designated hereafter as the Day and the Night Orientation Jumps.

During the next flight, one member of each pair of students is graded while performing as a jumpmaster and, after going through a series of commands, inspections, and decisions about position of the aircraft relative to the drop zone, he puts out the door a heavy bundle of equipment (weighing approximately 200 lbs) as well as his partner who is used to represent a line or "stick" of jumpers. When both door bundle and jumper are gone, the jumpmaster himself follows. This is the Door Bundle Jumpmaster routine for what will be referred to, here, as the Day Graded Jump.

Later that day, after night falls, the class flies again and this time every man in the class wears combat equipment and is graded while performing as a jumpmaster. For this routine there are no actual jumpers other than the jumpmaster himself even though the student goes through his routine as though he were giving jump commands to other individuals. This is the Combat Equipment Jumpmaster routine for what will be called the Night Graded Jump.

The final flight takes place the following day (weather permitting) when the remaining member of each team, the one who served merely as a jumper on the Day Graded Jump of the preceding day, is graded while acting as the jumpmaster and he, then, puts out a door bundle, his partner who is now serving as a jumper, and himself.

Grading System for Performance in the Aircraft. Each student, as he goes through his routine serving in his turn as jumpmaster, is closely attended by two members of the jumpmaster cadre one of whom grades his performance and the other records the result. Each student is given a cushion of 30 points out of which he may be penalized for errors in his performance and still pass the course. If, on any jump, he loses more than 30 points, he fails and leaves the course immediately thereafter. Each of the important actions of the jumpmaster routine is assigned a specific number of points which are lost if the individual either forgets to perform them or performs them improperly. Points lost range from a single point assessed for a weak or late performance of non-critical actions, such as being one second late in getting the jumper away, to a maximum of minus 35 points for failure to perform actions of extreme importance such as hooking up the static line (which automatically deploys the main parachute) -- a life-threatening error.

Training Session. All students initially chosen for the experimental group were taught the following breathing sequence:

> Sharp intake of deep breath; press diaphragm down gently; hold for 8 seconds: release over 4 seconds; hold without breathing 4 seconds; take one regular breath; repeat from the beginning.

It was related to the group that this method (slightly modified by insertion of instruction about pressing down with the diaphragm) was used by Fenz in training his novice skydivers (Fenz, Note 1) and was also used by him (and a similar technique by others) in laboratory experiments in which heart rate was

lowered by as much as 30 beats per minute (Laird & Fenz, 1971; Westcott & Huttenlacher, 1961).

To underscore the potential of this method, they were also informed that breathing techniques closely resembling this one (including the instructions about the diaphragm) are used in various practices of Zen Buddhism including the Zen art of archery (Herrigel, 1953), a form of archery practiced for centuries by the Japanese and widely known to produce archers who perform prodigious feats with a bow and arrow.

The students were asked to practise the technique often prior to the beginning of the training flights during the following week and to try to develop an ability to breathe in that manner without conscious thought, admittedly a difficult goal to achieve.

They were instructed to actually use the technique whenever they were beginning to feel themselves becoming "too tight" for their best performance while they were waiting to go through their jumpmaster routines or to make a jump. Since, as mentioned earlier, only two students could be graded at any one time, the remaining students would have long periods of flight time to bridge while awaiting their turn. It was during these periods and a similar period waiting to board the aircraft prior to a flight that the students were advised to use the techniques. In addition, since it was doubtful that most students would practice the technique enough to be able to do it in the recommended manner without conscious thought, they were told not to deliberately try to breathe in that fashion while being graded on their routines. It was feared that this would have a disruptive effect on their performance and unfairly penalize them relative to the members of the control group. These instructions consequently insured that any effects of the experimental technique would be residual ones from having used it prior to their actual performance.

Data Collection

Two types of data were collected for this experiment: measures of performance and measures of stress. The performance measures were the scores from the graded jumpmaster routines. The stress measures were heart rate measurements made immediately prior to the jumpmaster performances and end-of-course ratings by each individual of the stress he perceived himself to be under during certain key events of the course such as when going through the Door Bundle Jumpmaster routine.

Performance Measures. The performance measures came from the grades, described earlier, given to each individual for his performance as a Door Bundle Jumpmaster and a Combat Equipment Jumpmaster.

Heart Rate Measure. The physiological measurement of stress for this study was heart rate per 15 second period taken from the pulse in the carotid artery under the jaw at the side of the neck. The heart rate measure, while certainly not the only or necessarily the most valid indicator of the stressful condition, was chosen because it was the prime indicator in the skydiving studies (Fenz, 1975) and was, in that program of research, shown to be a lable measure of stress and one that tracked changes in respiration, both up and down, and led the more refractory measure of basal conductance from point to point.

The method of measuring the heart rate by taking the pulse by hand with a stopwatch from the carotid artery was chosen because it was a quick and reasonably accurate method of taking measurements on large numbers of individuals without disrupting their activities or disturbing their equipment, a sensitive point during airborne operations.

The heart rate readings were taken by the experimenter on each individual at that point in each flight into the drop zone at which the men to perform on that pass had risen to hook up their static lines and moved back toward the open jump doors to begin their graded routine. Since the first pass over the drop zone for all training flights was always to put out jumpers who were not in the class, the cue to take the heart rate measure in the first two and all succeeding pairs of students was that, when the jumpers of that pass stood up to hook up, the next two students in line, those who were then the ones seated nearest the open doors, would be measured. The measurement was thus taken while the students being measured were still seated and were watching the performance of the pair that had preceded them.

No difficulties were encountered in finding and measuring the pulses of most students in the aircraft. Some could not be located at the carotid artery and were thus measured at the wrist. For one individual, on one single flight, the pulse could not be located for measurement.

In order to transform these raw heart rate measures into a scale of units by which individuals could be compared, one to another, the final form of the heart rate data was to be change-in-heart rate expressed as percent-of-change from baseline measure. Baseline heart rate for each individual was established by taking the average of two readings made early in the first week of the course under nonstressful conditions. The first reading was taken on the morning of inprocessing day, the first day of the course. In order to produce a more stable measure to serve as a baseline for comparisons, a second reading was taken two days later during a break between classroom sessions. This second reading was then averaged with the first. The two baseline measures were highly correlated ($r = .79$, $p = .001$). Subsequent heart rate readings, taken under conditions of stress in the second week of the course, were then transformed into measures of percent-of-change from this baseline, thus establishing a comparable zero point for all individuals in the study.

Perceived Stress Measure. At the end of the course, in the morning of graduation day, the surviving class members (some had failed during the graded jumps and had departed) were given a questionnaire asking them to rate, on the scale below, the amount of stress which they felt themselves to be under during events such as the two orientation and the two graded jumps:

Not Successful At All	Borderline	Slightly Stressful	Moderately Stressful	Considerably Stressful	Extremely Stressful
1	2	3	4	5	6

Those individuals who failed during the graded flights and left the course immediately upon returning from the drop zone were given stress questionnaires by the cadre to be filled out before they left or done at home and returned by mail.

Selection of the Final Groups. The initial assignments of students to the experimental group yielded a group of individuals who were at least passive volunteers for the experiment, who attended the training session, and knew what to do and when to do it. A question of vital importance, however, for the interpretation of the results of the experiment was did they actually utilize the technique when they came under stress and, if so, to what extent? To insure that for the final data analysis the experimental group contained only those individuals who actually used the technique, questionnaires were distributed after the end of the course asking the experimental group members to rate the extent to which they used the deep breathing technique during the second week of training. The following 4-point scale was used:

Never	A little	Some	A lot
1	2	3	4

In addition, the questionnaire also provided space for free comments about the experiment. From a joint examination of both the extent-of-usage questionnaire and the free comments, augmented by discussion with individuals in the experimental group, it was apparent that the selection of the response "A little" on the above question was a polite way to say that the technique really wasn't used at all.

For those reasons selection of the responses "Never" and "A little" were taken as disqualifiers for inclusion in the experimental group and only those individuals claiming to have used the technique "Some" (six respondents) and "A lot" (two respondents) were included in the final experimental group. One individual from the initial experimental group who failed the course and didn't fill out the post-course usage questionnaire was retained in the final experimental group. In a similar manner, the final control group was constituted according to how they responded to a questionnaire which asked them what techniques, if any, they may have used to "psych themselves up" or calm themselves down prior to performing as jumpers or jumpmasters. Three of the control group students, including one who wrote as though his personal technique would be hard to believe, reported that they used slow breathing to help them during the stressful parts of the course and they were dropped from the control group as a consequence.

The final matching of the reduced list of experimental subjects with that of the controls was accomplished by starting with the PFC at the bottom of the list and, proceeding upward, retaining all the original pairs matched by rank which had survived all deletions and by creating new pairs where an experimental student had lost his old pair by his having either failed the course prior to the aircraft performance phase or been disqualified for having used slow breathing on his own. New pairs were made by matching unpaired experimentals with unpaired controls higher up the list, therefore, with a man of equal or, quite often, slightly higher rank. Individuals who failed during the aircraft phase for which the stress-management technique was intended were retained in the experiment and their data were analyzed along with the rest.

After deletion of individuals from the experiment due to classroom failures, jump injuries, and questionnaire responses, the final list contained nine pairs of individuals with two eligible control group individuals unpaired and unused.

RESULTS

The data from the three criterion variables for the comparisons between the experimental and the control groups are summarized in Table 1.

Heart Rate Differences

From the top section of Table 1, it can be seen that heart rates, expressed as percent change from baseline, were, at the time of measurement, lower on average, for the experimental students than for their matched controls for every training jump in the course.

Uncertainty about the actual level of measurement represented by scores on this and the remaining criterion variables, led to the choice of nonparametric statistic, the Wilcoxon matched-pairs signed-ranks test for the analysis of all the data in Table 1 and all analyses were done using the SPSS computer program for that technique (Nie and Hull, 1977). Wilcoxon tests on the heart rate data from each jump established that the experimental group had heart rates significantly lower ($z = -1.78$ and -1.96, $p = .04$ and $.03$, 1-tailed test, respectively) than their matched controls for both of the two night jumps but the differences between the groups on the day jumps were not significant.

The differences in the number of individuals in the groups over the jumps reflects, for the Day Orientation Jump, the inability to find a pulse on one of the experimental students while, for the Day Graded Jump, the differences result from the fact that two of the experimental students failed the course on the Night Graded Jump in what was, for them, their first and only performance as a jumpmaster. Thus, they were never tested as Door Bundle Jumpmasters during the day flight.

Perceived Stress Differences

Though the experimental group as a whole rated itself as being under slightly less stress than did the control group during each jump the differences were not significant.

The number of individuals listed for these comparisons is eight because one of the experimental students failed the course on a night flight. He departed the post without filling out the perceived stress questionnaire and failed to return a questionnaire sent to him by mail.

Performance Differences

Only the last two jumps listed in the table were graded and for neither jump was there a significant difference between the two groups in performance. For the Day Graded Jump, the experimental group, now reduced by two individuals who failed earlier, averaged considerably fewer points lost than the control group. Although nonsignificant, ($z = -1.26$, $p = .10$, 1-tailed test) this comparison suffered from the low number of individuals involved and the difference obtained is suggestive of a treatment effect.

On the Night Graded Jump, two experimental group members did poorly enough to fail and the higher average for points lost of the group reflects those failures.

(15-2)

Table 1

MEANS AND STANDARD DEVIATIONS OF HEART RATE CHANGE, PERCEIVED STRESS, AND POINTS LOST IN EXPERIMENTAL AND CONTROL GROUPS AND NUMBER (N) IN EACH GROUP OVER FOUR TRAINING JUMPS

	DAY ORIENTATION JUMP			NIGHT ORIENTATION JUMP			DAY GRADED JUMP (Door Bundle Jumpmaster)			NIGHT GRADED JUMP (Combat Equipment Jumpmaster)		
	HEART RATE CHANGE (Percent change from baseline)											
CONDITION	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N
Experimental	14.6	15.7	8	29.4*	18.4	9	30.9	18.8	7	30.8**	13.2	9
Control	19.4	17.2	8	43.8	17.6	9	36.7	13.9	7	38.2	14.3	9
	PERCEIVED STRESS											
	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N
Experimental	2.4	1.3	8	2.6	1.4	8	3.9	1.7	7	3.6	1.8	8
Control	2.9	1.5	8	2.8	1.5	8	4.6	1.0	7	3.9	1.1	8
	POINTS LOST											
	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N
Experimental	-	-	-	-	-	-	3.6	4.0	7	15.6	11.0	9
	Not Graded			Not Graded								
Control	-	-	-	-	-	-	6.9	6.3	7	11.2	9.7	9

*Experimental group significantly lower than control group, p = .04, Wilcoxon matched-pairs signed-ranks test.
**Experimental group significantly lower than control group, p = .03, Wilcoxon matched-pairs signed-ranks test.

DISCUSSION

The clearest outcome of this research is that an evaluation experiment of this type conducted in an operational environment is a difficult undertaking. Respiration control was presumably effective in reducing the heart rates of the experimental students relative to the controls on two training jumps, both conducted at night, when, due to limited visibility, parachute jumps are somewhat more dangerous and therefore more stressful. The technique was not, apparently, a great aid to performance in this setting. Only three students performed so badly during the graded jumps that they failed the course as a result -- two of them were members of the final experimental group and one had been included in the initial experimental group but was dropped from the final group due to having reported that he used the technique only "a little."

Considering the data from this experiment, overall, it is noteworthy that, for every comparison between the experimental group and their matched controls save one -- points lost during the Night Graded Jump -- the experimental group, on average, did better than the controls. While there is no statistic appropriate for assessing the statistical significance of this outcome, these results do seem to indicate a possibility that respiration control did have some generally helpful effect.

The only judgment which can be given at this time on the effects of respiration control as a stress-management technique is that it apparently worked at times in this experiment to control heart rate and that, although it wasn't established by the evidence here, it is still possible that it was of marginal value in performance. It was, perhaps, the nature of the men who participated in this experiment that worked to obscure whatever power the technique may possess to enhance performance. The Jumpmaster Course was chosen for this research because of the relatively exact performance measures which it provides and because the students who go through the course are, in the main, dedicated individuals who could be depended upon to give the technique a good test. However, they are also, on average, relatively experienced with and competent at training of a demanding nature similar to the activities in this course. A technique which might have been of considerable benefit to individuals less adept at performance under such conditions would have a smaller effect for these men and, thus, for that effect to show through all the uncontrolled variability in the scores, would require sample size much larger than the one possible here.

In sum, considering all the factors involved, respiration control appeared to be a limited success at controlling one aspect of stress under the conditions of this experiment. As such, it requires further study to establish its usefulness to the military.

REFERENCE NOTES

1. Fenz, W. D. Personal Communication, November 2, 1979.

REFERENCES

Basowitz, H., Persky, H., Korchin, S. J., and Grinker, R. Anxiety and stress: An interdisciplinary study of a life situation. New York: McGraw-Hill Book Co., Inc. 1955.

Fenz, W. D. Stress and its mastery: Predicting from laboratory to real life. Canadian Journal of Behavioral Science, 1973, 5 (4), 332-346.

Fenz, W. D. Coping mechanisms and performance under stress. D. M. Landers, D. V. Harris & R. W. Christina (Eds.). In Psychology of sport and motor behavior. Proceedings from the North American Society for the Psychology of Sport, Pennsylvania State University, May 19-21, 1975, 3-24.

Fenz, W. D. & Jones, G. B. Individual differences in physiological arousal and performance in sports parachutists. Psychosomatic Medicine, 1972, 34 (1), 1-8.

Fenz, W. D. & Jones, G. B. Cardiac conditioning in a reaction time task and heart rate control during real life stress. Journal of Psychosomatic Research, 1974, 18, 199-203.

Herrigel, E. Zen in the art of archery. (R.D.C. Hull, Trans.) New York: Pantheon Books, 1953.

Laird, G. S. & Fenz, W. D. Effects of respiration on heart rate in an aversive classical conditioning situation. Canadian Journal of Psychology, 1971, 25, 395-411.

McCaul, K. D., Solomon, S. & Holmes, D. S. Effects of paced respiration and expectations on physiological responses to threat. Journal of Personality and Social Psychology , 1979, 37 (4), 564-571.

Westcott, M. R. & Huttenlocher, J. Cardiac conditioning: The effects and implications of controlled and uncontrolled respiration. Journal of Experimental Psychology, 1961, 61, 5; 353-359.

FEASIBILITY OF LOW-COST SIMULATION FOR SHORT RANGE AIR DEFENSE

Richard J. Carter
US Army Research Institute for the Behavioral and Social Sciences

Edward W. Frederickson
Applied Science Associates, Inc.

Background

Studies are currently being undertaken to establish air defense command and control system requirements essential to the accomplishment of the air defense mission and to evaluate equipment performance and doctrinal concepts. An important, critical part of the above system requirements is information concerning short-range (SHORAD) and man-portable (MANPAD) air defense weapon system personnel detection and recognition capabilities. This information is however difficult to obtain because of the high cost of live aircraft support and of the nonavailability of foreign aircraft.

During the interval 1964 through 1976, the Army supported an extensive program of research aimed at developing a data bank concerning the capabilities of operators of forward area air defense weapons to detect, identify, and estimate the distance of low-flying aircraft. However, these studies had to be conducted in a part-task evaluation environment. That is, each of these critical tasks was studied in isolation from the other components of the total operational sequence. Part-task research only was conducted because the sponsoring agencies had a need for data indicating maximum and average target detection and identification ranges. As a result, there is a lack of information concerning an operator's abilities in performing part-task components when they are embedded in whole-task performance requirements.

Feasibility Study

A paper-and-pencil study is on-going at Fort Bliss to determine the feasibility of using a low cost simulation facility for assessing the performance capabilities of operators of man-ascendant forward area Air Defense weapon systems. A man-ascendant system is one that relies on human input to the control, operation and decision making functions of a system. These inputs are based upon perceptual, psychomotor and cognitive processes in man's functioning as a systems operator. The processes occur simultaneously, thus resulting in a complex man-machine operation. It is the measurement of the behavioral results of the interaction of these processes with the system's environment that is of interest in answering the simulation feasibility question.

The weapon systems of concern are the man-portable REDEYE and its successor, the STINGER, missile systems, and the short range Vulcan gun system and Chaparral missile system and their follow-on systems, the DIVAD gun system and ROLAND missile system. In all cases, the human operator must visually search for and detect aerial targets that pose a hostile threat to defended sites. They must identify the target as

hostile and then decide whether they can engage selected targets. Some systems require that the operator, without aid, track the target while gathering information and then throughout the engagement. The degree to which the human operator cannot perform the entire engagement sequence from searching to weapon firing defines the degree of system degradation in its capability of completing its mission.

SIMFAC Simulation Purpose

The driving aspect for developing the simulation is the purpose which it will accomplish. In SIMFAC, the primary purpose is:

> To determine the feasibility of building a simulation facility which could be used to develop data bases of forward area Air Defense weapons system operator performances that would not be significantly different from performance data obtained under real world conditions.

The question of how well the range of human operators can perform the entire sequence of procedures has not been empirically determined. Most studies of forward area weapon operator performance have been part task studies. Behavioral science research has shown time and time again that there is a distribution of performance levels for specific behaviors, and that each individual performer does not usually perform at the same level on different but related behaviors. In fact, in some cases, an individual may perform one behavior very well, but may do very poorly on another. The implications here are that the assumption of population mean behavior on each subtask may lead to incorrect conclusions about system capabilities and effectiveness. Therefore, the intended use of the simulation facility is to focus on whole task measurement rather than on a part task level but yet have the capability of gathering data on part task performance on a non-interfering basis.

The ideal situation for determining performance capability would be to use the real world environment that would be expected to exist in combat. This would be very costly, tie up tremendous amounts of resources and logistically be very difficult to manage. The decision then was to determine if it would be cost-effectively feasible to use some representation of the real world and still obtain performance data that would match or come close to that which would be obtained in the ideal situation. The major issues then became those of determining what aspects of the real world need to be represented to produce valid estimates of operator performance and how this should be done. The two questions are interrelated in that with some forms of representation of the real world some analytical data would not be needed but with other forms much more detailed information would be required.

An analysis of the SIMFAC purpose led to additional requirements that the simulation model must meet, further constraining the simplification

analysis. First, the operator whose performance is to be assessed must be allowed to complete the entire engagement sequence without interruption by the measurement process. It is required that performance be measured at least at the task level, and preferably have a capability of measurement at the skill level of job description. Performance is to be measured over a range of exogeneous parameters and variable levels. It is desirable that performance be measured under conditions where the interactions between variables change during the engagement. And, finally, it would be desirable to be able to use the same facility with all current and future MANPAD/SHORAD weapon systems. The data bases to be generated by use of the facility will produce statements descriptive of the performance relationships of individual tasks with each other, and also between individual tasks and the outcome of the entire engagement sequences.

The SIMFAC purpose would imply that isomorphism would be required in the measurement environment--preservation of the relationships between elements which must be represented in a one-to-one correspondence from the real world. The accepted approach, however, usually is to use homorphism, a "like form" representation of elements. Such real world representation is commonly referred to as the process of simulation. Simulation, however, is both narrowly and broadly defined. Redgrave (1962) provides a useful definition by emphasizing transformation for the convenience of meeting special purposes. Simulation, he says, is a representation or technique which transforms, either iconically or by abstraction, selected aspects of the real world out of their resident framework into a form more convenient for the analyst's purpose.

The broad definition of simulation explicit in Redgrave's definition will be used in this paper. Implicit in most definitions of simulation is the concept of modeling the real world. A model is the basic representation of a real situation or environment. A model consists of a description of the important elements of that which is modeled and their functional relationships. The descriptive model then is transformed into a simulation model to meet specific functional needs--either prediction or comparison. Initially, the SIMFAC project is interested in the prediction of operator performance over a variety of conditions typical of the Air Defense mission. This would embody the establishment of standards of expected operator performance. At a later time, there may also be an interest in comparing performance of individuals or groups under various sets of conditions. The use of the simulation facility would provide for a logical way to forecast the outcomes of alternative operational procedures. Given sufficient empirical data, future computer simulations of the descriptive model could be run that would provide for a systematic, explicit and efficient way for decision makers and planners to determine deployment and engagement doctrine.

When planned for or at least anticipated, simulation could be used for possible controlled experimentation in situations where real world experimentation would be impractical or prohibitive. In those instances where the impact of situational complexity on operator performance is of

experimental interest it is critical that the descriptive model be complete in every detail. Complexity derives from the interactions among system elements and the physical aspects they must deal with. Changing one aspect of a system or physical environment may well produce unpredicted changes or create a need for changes in other parts of the system. So the approach to simulation must meet all requirements for accomplishing its purposes.

Model Characteristics

In the development of simulation facilities, Shannon (1975) stresses that several modeling questions must first be addressed. The dimensions of the model must first be selected to the degree possible. At least the first three of the following four characteristics must be determined before a model can be built:

1. Static vs dynamic
2. Deterministic vs stochastic
3. Discrete vs continuous
4. Iconic vs analog or symbolic

Since the focus of the SIMFAC project is on whole task performance, we could not meet the measurement objectives by reducing the engagement sequence to static cross section segments. So the model must be dynamic. The MANPAD/SHORAD operators employ real-time control skills where error correction is essential in order to cope with deviations and unexpected, rapid changes in a target's behavior. Therefore, changing (dynamic) conditions in the environment are critical in accurate measurement of performance. To use other than a dynamic simulation would be an unjustified simplification leading to erroneous extrapolations of performance. A second dynamic requirement is that whole task performance studies have not used objective measures to a sufficient degree, and whole task performance may not be inferred from part task measures.

The specific impact on performance of many variables is not known, and many of the interactions between input variables and performance may not be linear. Therefore, we cannot use, at least initially, a deterministic model. The model must be stochastic. The purpose for the human in most systems is his versatility and capability to handle previously unencountered events. Man does this through his cognitive system, reaching decisions for action not possible in unprogramed control devices (including computers). The cognitive decision making processes are difficult to model, so for most situations, it is not tried. The human operator will interact with the simulated presentation of information from his unique combination of individual differences and may react in an unexpected and not a deterministic way. For the results of decision making to be validly assessed, the dimensions of the information sources must be representable and controllable, and motion is one of the important sources of information, which dictates a dynamic, stochastic model.

The model also must be continuous because the separate engagement stages are not independent, discrete events. In many situations, the behavioral requirements for one stage may have been met during a previous stage. Most behaviors required for carrying out the weapon system mission are directly correlated to the changing position, aspect, size, angle, speed and maneuvers of the target. The model must therefore provide for the continuous changes in the target/system relationship.

The fourth aspect of models that must be determined actually is embodied in the SIMFAC project. The feasibility issue can only be answered by evaluating the cost and capability of the various kinds of simulation vehicles for representing the MANPAD/SHORAD environments.

Types of Simulation

Shannon (1975) represents the range of simulation types in the following manner:

Real World	Artificial World	Abstract World
Physical	Graphs	Computer
Scaled	Charts	Math Model
	Management Games	
---Iconic---	--Analog---	---Symbolic---

In many instances, a combination of kinds of simulation are used. Aircraft cockpit simulators are mixed iconic and symbolic models. Some management game simulation is incorporated by artificially fixing some input parameters.

A major difference in 3-dimensional iconic vs analog and symbolic models is the flexibility of changing the element characteristics and their interactions. Such iconic models are essentially Gestalt in nature, the whole situation is more than the sum of its parts. Whereas the analog and symbolic models are only the sum of their parts. That is, only those elements and interactions that are explicitly identified and defined in the model will occur in the simulation. This, of course, makes the analog and symbolic simulations more controllable, but less realistic. Much of the model control occurring in analog and symbolic simulation can be effected in iconic simulation by using statistical controls--analyses of variance or co-variance.

Simulation Model Components

As the selection of a model/simulation type moves from the real world end of the simulation dimension to the symbolic end, simulation becomes more imprecise and Shannon indicates that this imprecision cannot be measured. The more precise the model, the more sensitivity it has to the changing of parameter values. The sensitivity of the model is accurate to the degree that interacting elements have been identified and included in the model.

Shannon (1975) has provided a general simulation equation that reflects the basic problem in modeling.

$$E = f(X_i, Y_j)$$

Where:

- E = the effect of the system's performance
- X_i = variables and parameters that can be controlled
- Y_j = variables and parameters that cannot be controlled
- f = the relationship between all variables and parameters, controllable and non-controllable.

This equation leads to the identification of the major ingredients of a model. Components are system elements and subsystems that perform a specified function. Variables are those factors whose values vary with function changes. Variables fall into two categories; independent and exogeneous, which are input variables and dependent and endogenous which are system status and output variables. Parameters are those factors whose values remain constant whether functional changes occur or not. Functional relationships describe how elements, subsystem variables and parameters interact as the system carries out its mission. The relationships for the SIMFAC model will be stochastic rather than deterministic because of the uncertainty of the output for given inputs. In many Air Defense situations, the inputs may even be uncertain, for example the number and type of targets that may have to be deal with. Constraints are factors that place limits on how the system can allocate or expend resources. Constraints may specify the value level at which parameters must be set. In Air Defense systems, the designated DEFCON status is a parametric constraint, as is the number of batteries a REDEYE operator carries. Criterion functions are the end result of system functioning. These are the desired outcomes when the mission has been completed. Submission or task outcomes may also be included here, such as the target identification decision.

Representing the Real World

In the SIMFAC project the measurement of the outcomes of the perceptual, psychomotor and cognitive processes are of primary interest. Therefore, the operator and his behaviors are not a subject for simulation. It is the physical elements of the system and its operating environment that are to be considered for simulation. In modeling a physical system, there are several factors that must be considered in order to maintain the essential psychological elements. First are the fundamental laws that operate in the real world and account for certain impacts and interactions which influence criterion functions. These laws must be accounted for in a simulation if they operate in the real world. Second, all procedural elements must be described and represented as well as the systematic variables that impact these elements. All policy (doctrines, tactics, and SOPs) inputs must be considered. Random components that have significant influence on functions must be identified and included. And, finally, the human decision

making requirements have to be specified so that information sources required for the decision can be represented.

Since the SIMFAC simulation is to be a scaling and/or an abstraction of a physical system, the sets of factors mentioned above at a gross level must be dealt with systematically so as to maintain the psychological validity. The specific purpose for the simulation provides an initial organizing structure for addressing the analytical questions of which specific factors in the various sets of factors are significant and thus must be represented in the simulation. The success of establishing a successful simulation then is a function of how well the significant system elements are identified and defined.

SIMFAC Model - World View

To this point the simulation model has been characterized in terms of general requirements. It must be basically a dynamic, continuous and stochastic model. The purpose to which the simulation facility under study will be put has been discussed in terms of the requirements it places on the feasibility decision. The next section will deal directly with the process of building a model of the SHORAD/MANPAD system environment. First, the systems and their environment will be described in terms of general systems theory. Then a "world view" taken for the analyses will be discussed. And finally, the elements of the environment that must be represented in any simulation will be presented with some comparisons of the capability of the various types (levels) of simulation to incorporate the required elements.

A general systems model consists of four major elements: the input, transformation/operation, output and feedback. The use of this model allows for the systematic identification of all critical elements and relationships of the Air Defense environment. Input variables are independent and external to the man/machine system. They cause the system to take various actions in order to carry out its function. Some input factors set some aspects of system status and act as parameters. Five major categories of input factors were identified for the SIMFAC model:

1. System mission
2. Command and control doctrine
3. Logistical support
4. Physical characteristics and conditions of the environment
5. Target characteristics and dynamics

Logistics, mission, and command and control categories can be assumed to be fixed as given parameters. Seldom, if ever, would factors in these categories vary during specific engagement sequences. The target and the environment both have variables which have changing values during the engagement. They both also have characteristics which do not vary. The important point here is that those target and environmental variables which change and interact non-linearly, or whose interactions are difficult to specify must be identified in terms of their impact on the criterion functions of the system.

Since SIMFAC at least initially, has little interest in evaluating machine capabilities or operations, the analysis of the man/machine functions needed to focus only on the operator tasks. The "world view" taken for building the simulation model was established here. In the initial listing of the operator tasks, it became obvious that all system events were keyed to perceptual information and the dominant perceptual system was vision. The auditory system does become involved at two points. First, early warning information may be provided over the communication net or if a helicopter target is in the area it may be heard before it can be seen. Second, the REDEYE and Chaparral system use an auditory signal to indicate IR source lock on by the IR sensor. These are important signals for keying operator events but not as significant as the visual information that must be sensed without sensory aids.

The output factors are embodied in the mission requirements of deterring, delaying, altering the mission of or destroying a hostile aircraft. These criterion functions are the end result of the engagement process and must be assessed for input information to the decision for subsequent action.

The analysis of the SHORAD/MANPAD systems engagement problem and discussions with other researchers working with performance assessment under simulation conditions led to the adoption of a visually-keyed event-orientation world-view of the forward Air Defense environment. The rationale for adopting this view was that the entire engagement sequence followed in these systems' operations depends upon a continual input and processing of visual information. Once the target has been visually detected, visual contact must be maintained at least until the weapon is fired. Reliability of system functioning is tied to detecting and discriminating cues which consist of visual detail. Any condition that interacts significantly with the visual cues so as to degrade cue detection or discrimination must be present in the simulation environment in order that performance measurements have generalizability to the real world. This requirement then leads to the decisions about presentation and control instrumentation.

The focus of the visual orientation of the SHORAD/MANPAD systems operators is the aircraft target, which is the source of cues that trigger the starting and stopping of the engagement events. Environmental variables, especially atmospheric conditions interact with target characteristics to degrade or enhance the visual perception of cue information. It is these interactions upon which the validity and generalizations of operator performance measures rests.

The state of the MANPAD/SHORAD systems then must be defined with reference to the dynamic portrayal of the aircraft targets. Scenarios would then be driven by the specific characteristics of the target during an engagement. Iconic simulation would present fewer scenario problems than would analog or symbolic simulations. All of the significant detailed target characteristics and their changing values would already be present in scaled model or film strip targets used in three-dimensional and two-dimensional iconic simulation. However, the use of analog or symbolic models would require a rather detailed analysis of the physical dimensions of the target features and their changes almost on a second-by-second time

basis to accurately protray the operator's visual perception of the target cues during the entire engagement. Each potential target scenario would have to be so analyzed. The subsequent effort to represent the myriad of cue source changes would be significantly more costly for analog and symbolic simulations than for iconic simulation. In addition, some interactions between atmospheric conditions (water content, particle content and shimmer) and cue information probably can not be specified and thus could not be represented.

Visual System Variables

There are three keys to representing the real world in a simulation facility: cueing, controlling and task loading. In the SIMFAC problem task loading is primarily a function of cueing. The cueing problem as inferred above is the visual presentation problem. Where different operator responses are required for different cues in the real world, the operator must be able to discriminate between the various cues in order to make the correct response. The sensitivity of the simulation, then, must be adequate to ensure that the operator can discriminate among the cues that must be represented. The task of the simulation designer then is to establish the minimum level of fidelity of cues that will ensure discrimination (Cream, 1974). In representing the visual system inputs the basic problem to be dealt with is how to present the smallest object (cue source) that must be represented. Object size and maximum discrimination range under ideal conditions are limiting factors. These factors create the visual problem of cue resolution in a wide field of view.

In viewing a target in the atmosphere, any serious limitation of visual range is due to what Middleton (1952) calls the atmospheric aerosol (the aerial colloids). This condition is due primarily to liquid droplets, the most important class of particles in the atmospheric aerosol. Large variations in the photometric properties of the atmosphere may occur as the content and density of the aerosol changes. A second significant particle in the air is dust, with a third, smoke, increasing in significance with time especially near large urban areas. The liquid droplets may vary in size from 10^{-6} to 10^{-1} centimeters in radius. The larger and more varied the atmospheric particles the more that light is scattered.

In a particleless atmosphere, light is scattered by the molecules of the permanent gases in the proportion to the inverse fourth power of the wavelength of the light (Middleton, 1952). In an atmosphere of a pure dry mixture of natural gases, visual range would be more than 350 kilometers. As non-permanent particles are added to the atmosphere, the visual range, as well as the amount of illumination, is reduced. Four critical factors influence the visual system in terms of how far and what we can see:

1. The optical properties of the atmosphere.
2. The amount and distribution of natural and artificial light.
3. The characteristics of the target objects.
4. The properties of the eye.

The interactions of the factors are both linear and non-linear. Shimmer, a disturbance of the atmosphere near the earth that occurs as the surface temperature increases above the atmospheric temperature, further complicates the visual system that must be represented in any visually oriented simulation. It becomes extremely expensive to create the conditions and produce the amount of information and its distortion in highly detailed large areas--the wide field of view problem.

The degradation of atmospheric conditions is defined in terms of the meteorological range--the range at which objects at known distances can be seen. As meteorological range is reduced the significant perceptual phenomenon of apparent target-to-background contrast ratio is also changed. Meteorological range is not necessarily omni-directional, thus possibly resulting in varying levels of contrast ratio in a wide search area. Apparent contrast ratio is also a function of the inherent target/background contrast, which usually changes as the target moves across the visual field because of: (1) the varying background; and (2) the sky/ground luminance ratio. Contrast is a subtle variable, probably of considerable importance in target detection. It is also important in recognition to the degree that critical target features may be non-discriminable.

The visual threshold for a given target in terms of distance is a function of target size, the amount of light (luminance) and the amount of time the target remains projected on the retina. In other words, it takes time to see (detect) a given sized target at specific light levels. Under a given set of atmospheric conditions, the limiting factor for detecting a target with specific perceptual and physical characteristics in the real world is the visual acuity of the observer. As visual acuity varies from near perfect vision, the degrading atmospheric and target factors interact to produce increasingly poor target detection and identification performance. Duntley (1948) offered an equation that gave the probability of detecting a target at or near threshold as a function of all the above mentioned factors (except shimmer) plus several others, such as target range and altitude, and several constants. The point in this discussion is that the visual target detection environment is very complex, and as mentioned earlier, very difficult to represent in any type of simulation other than an iconic model.

Simulation Presentation of Visual Variables

But not all iconic models are appropriate, given the importance of the critical atmospheric factors that degrade detection, identification and tracking performance--particularly the effect of shimmer, varying contrast/ratios and amounts of illuminance. Two dimensional iconic simulations using motion pictures can present a fairly high fidelity representation of the real world, capturing much of the atmospheric and target conditions that must be represented. However, instead of the visual acuity of the observer being the limiting factor in the visual problem, the resolution of the projected images becomes the determining factor for detection and identification tasks. The available resolution in films varies with the quality and speed of the film. Some films have resolution capabilities better than the eye, but is relatively expensive. The less expensive films

have a resolution level below that of the eye. But regardless of the film quality, the primary problem with photographic images is that amount of information available about a target image is fixed. Additional information cannot be obtained by using magnification (neither by projection lens nor by binoculars). With reference to the visual acuity of observers, operators with higher levels of visual acuity can obtain no more information from a photographic image than can the operators with 20/20 visual acuity. This problem could reduce the generalization of detection and recognition results.

The resolution ability of the eye takes at least two forms. First, the minimum detectable acuity--the smallest target the eye can detect--is less than ten seconds of arc. Second, the minimum distinguishable acuity (the ability to detect irregularities in form, shape or contour) is 40 seconds of arc. The computer generated images in the USAF simulators at Williams AFB, AZ, has a minimum resolution of six minutes of arc. A CRT with a 60^{o} diagonal field of view gives a minimum resolution of four minutes of arc. Although motion picture film can have a resolution that approaches the eye, two problems other than the one mentioned above exist. These problems are not with the film but rather with the projection lens and surface upon which the image is projected. The lens cause various kinds of distortion, as does the projection surface.

Understanding the limitations of the visual systems has always been a central problem of display research. As inferred in the previous paragraph, strong technical and thus economic constraints must be overcome to significantly improve the quality of the projected image in two dimensional representations of the real world. Improvements must take the direction of matching the performance of the display to the visual requirements of the observer. This direction follows from an analysis of predicting what an observer can see when he views a display. This is what Duntley (1948) attempted theoretically with his detection equation. Autonetics has recently conducted some detection research in which they emphasize that the design and specification of displays are a function of knowing what perceptual effects result from a specified set of visual system conditions, exactly the simulation design problem in our research.

Summary

The design of facilities for the purpose of assessing operator performance has received little attention. Where simulation has been used, the focus has been primarily on the design of training facilities and devices. The training orientation has been to represent critical context and task elements with as high a fidelity as possible within economic constraints. Physical/engineering fidelity has been emphasized although psychological impact has received some theoretical attention. From an assessment-of-performance point of view, the human factor considerations discussed above should drive the design of the measurement system in order to obtain valid, generalizable results. Engineering factors are important only in terms of their role in generating accurate performance assessment results under critical job conditions. Therefore, the position of this paper is that the degree of fidelity in the SIMFAC simulation should be based upon man's perceptual requirements. There should be perceptual

equivalence to the operational environment. All other things equal reduced engineering and physical fidelity with attention to cost considerations are desirable when it can be demonstrated that measure accuracy is not compromised.

This orientation, the perceptual focus in the design of a measurement system, is concerned with what is represented at the display and control interfaces so that measurement accuracy is maximized. This requires an analytical input about what perceptual details of the operational environment must be represented in a simulation to provide for performance measurement fidelity. The failure to represent a single critical element, feature or input may destroy the credibility of the results of measurement.

References

Cream, B.W. & Lambertson, D.C. Functional integrated systems trainer. Technical design and operation. Wright-Patterson Air Force Base, USAFHRL Tech. Rep. No. 76-6, 1975.

Duntley, S.Q. The reduction of apparent contrast by the atmosphere. Journal of the Optical Society of America, 38, 179-191.

Middleton, W.F.K. Vision through the atmosphere. Univ. of Toronto Press, 1952.

Redgrave, M.J. Some approaches to simulation, modeling and gaming at SDC. Santa Monica, CA: System Development Corporation, 1962. SP-721

Shannon, R.E. Systems simulation: The art and science. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1975.

CARTER, RICHARD J., Army Research Institute for the Behavioral and Social Sciences, Ft. Bliss, Texas.

PAPER-AND-PENCIL TESTING OF GEOMETRIC RADAR SYMBOLS (Wed P.M.)

Background

Every Army, radar-assisted air defense system has a unique set of geometric symbology. These sets not only have different symbols, but, when identical shapes are employed in different systems, they represent diametrically opposite and contradictory information. A need exists for a standard symbology because symbol confusion takes place when personnel must be retrained and reassigned to "new" systems.

During the first step in the research, shapes were identified which are stereotyped with the meanings friendly, hostile and unknown. Symbol sets of three and five symbols were put together based upon the results and tested in step 2.

Paper-and-Pencil Testing (Step Two)

The objective of this step, two experiments, was to find sets of symbols which can be easily discriminated and quickly reacted to in a paper-and-pencil mode. Radar console operators from Ft. Bliss searched for and marked through with a china marking pencil designated symbols in a field.

The field approximated a console screen at a moderate saturation level. A north-oriented 90^0 sector was plotted on 8x 10 1/2-inch white paper with X,Y coordinates located for 25 symbol positions. Ten sheets were arranged, using the sector as a blueprint, for each symbol set.

Reaction times and the number of erors of omission and commission were recorded for each subject. Repeated measures analyses of variance and Newman-Keuls tests were performed on the reaction time data. Chilsquare statistics were used on the error data.

PAPER-AND-PENCIL TESTING OF GEOMETRIC RADAR SYMBOLS

Richard J. Carter

US Army Research Institute For The Behavioral and Social Sciences
Fort Bliss Field Unit
PO Box 6057
Fort Bliss, Texas 79916

INTRODUCTION

Each of the following US Army air defense systems has a unique set of geometric symbology for display use:

a. AN/TSQ-51
b. AN/TSQ-73
c. Nike-Hercules
d. Hawk
e. PATRIOT
f. Roland
g. DIVAD Gun

These sets not only have different symbols, but, when identical shapes are employed in different systems, they represent diametrically opposite and contradictory information. For example, a circle may represent a friend in one system, and a foe in another.

As long as a radar console operator continues working with only one system the different symbologies cause no problems. However, when personnel are well trained in a system and then retrained and reassigned to a new system, as ancestorial systems are replaced, symbol confusion takes place.

Bergum and Burrel (1964) and Davis (1968) recommended that a standard symbology be adopted for employment across all Army radar air defense systems. The Army Research Institute's Fort Bliss Field Unit has initiated research aimed at fulfilling this requirement.

The first step in the research (Carter, 1979) was aimed at determining whether or not particular shapes exist in our population as stereotypes for the meanings friendly, hostile, and unknown. One hundred military fire control crewmen, 50 Hercules and 50 Hawk, sorted 60 cards into four categories. Each card had drawn on it a shape which was chosen either from symbology currently in the Army inventory, or symbol sets associated with systems in the procurement cycle, or simple shapes which have been used in discriminability studies and which can be generated by current hardware for presentation on cathode-ray tubes and/or plasma displays. The four categories were friend, hostile, unknown and other.

Stereotyping was found as follows:

a. The heart, 5-pointed star, flag, and circle are associated with the friendly meaning.

b. The swatstika, collapsed square, and X are associated with the hostile meaning.

c. The question mark is associated with the unknown meaning.

Symbol sets of five (two friends, two hostiles, and one unknown) and three (one friend, one hostile, and one unknown) members were put together based upon the results of the stereotyping survey. They were tested, via two separate experiments, in a mixed display which approximated a console screen at a moderate saturation level.

METHOD

Subjects

Sixty military subjects (Ss) from the US Army Air Defense Center at Fort Bliss, Texas participated in this phase of the research. They possessed either the 16E, Hawk Fire Control Crewman, or 16J, Defense Acquisition Radar Crewman, MOS. Half of the Ss were used in Experiment 1; the others in Experiment 2.

Apparatus

The stimuli were a heart (1), 5-pointed star (2), flag (3), swatstika (4), collapsed square (5), X (6), question mark (7), and 6-sided U (8). In Experiment 1, symbols were arranged into six sets as follows:

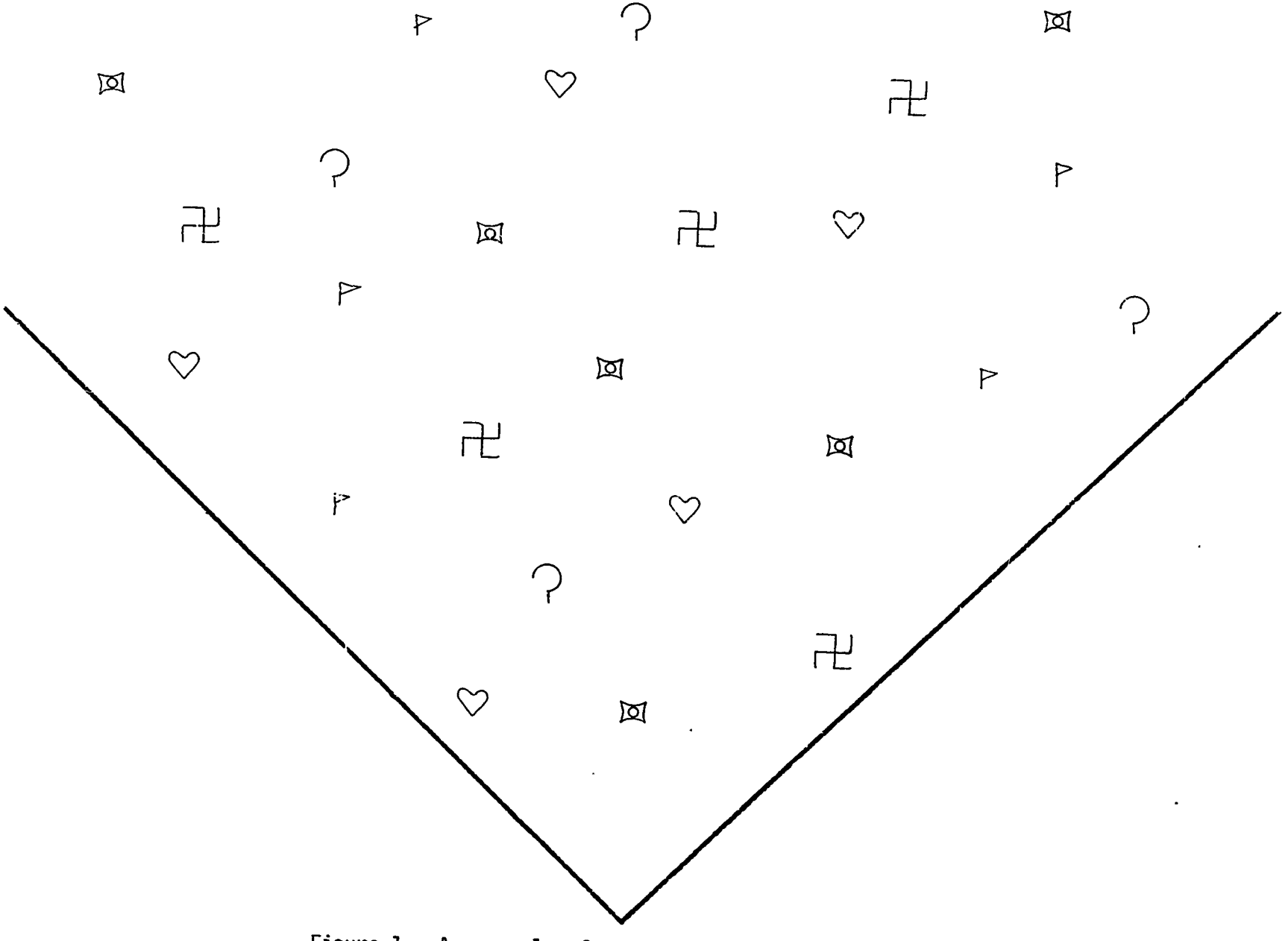
a. 1, 2, 4, 5, 7
b. 1, 2, 4, 5, 8
c. 1, 2, 4, 6, 7
d. 1, 2, 5, 6, 7
e. 1, 3, 4, 5, 7
f. 2, 3, 4, 5, 7

Eight sets were put together for Experiment 2. They were:

a. 1, 4, 7
b. 1, 4, 8
c. 1, 5, 7
d. 2, 4, 7
e. 1, 5, 8
f. 2, 4, 8
g. 2, 5, 7
h. 2, 5, 8

A set of specific practice symbols was assembled for Experiment 2. This set was comprised of a circle, square, and rectangle.

A north oriented, 90° sector was plotted on 8 X 10 1/2-inch white paper with X, Y coordinates located for 25 symbol positions - only 24 were used in Experiment 2. Fifty pages (27 for Experiment 2) were randomly arranged, using the sector as a blueprint, for each symbol set. Each symbol appeared either 4, 5, or 6 times on a page (7, 8, or 9 in Experiment 2) and twice in each position within ten pages (3 times within nine pages in Experiment 2). The symbols were drawn as large as possible, but dimensions were varied so that each symbol could be encompassed by a 1/4-inch circle. The pages were inserted into clear 9 X 11-inch document protectors and placed in looseleaf binders by set. Each binder also had blank sheets to divide the instrument into blocks of ten (nine in Experiment 2) pages. The looseleaf binder for practice in Experiment 1 was made up of ten pages from each one of the six sets. Examples of pages for five and three member symbol sets are pictured in Figures 1 and 2, respectively.



CAR-2-3

Figure 1. An example of a page used in Experiment 1

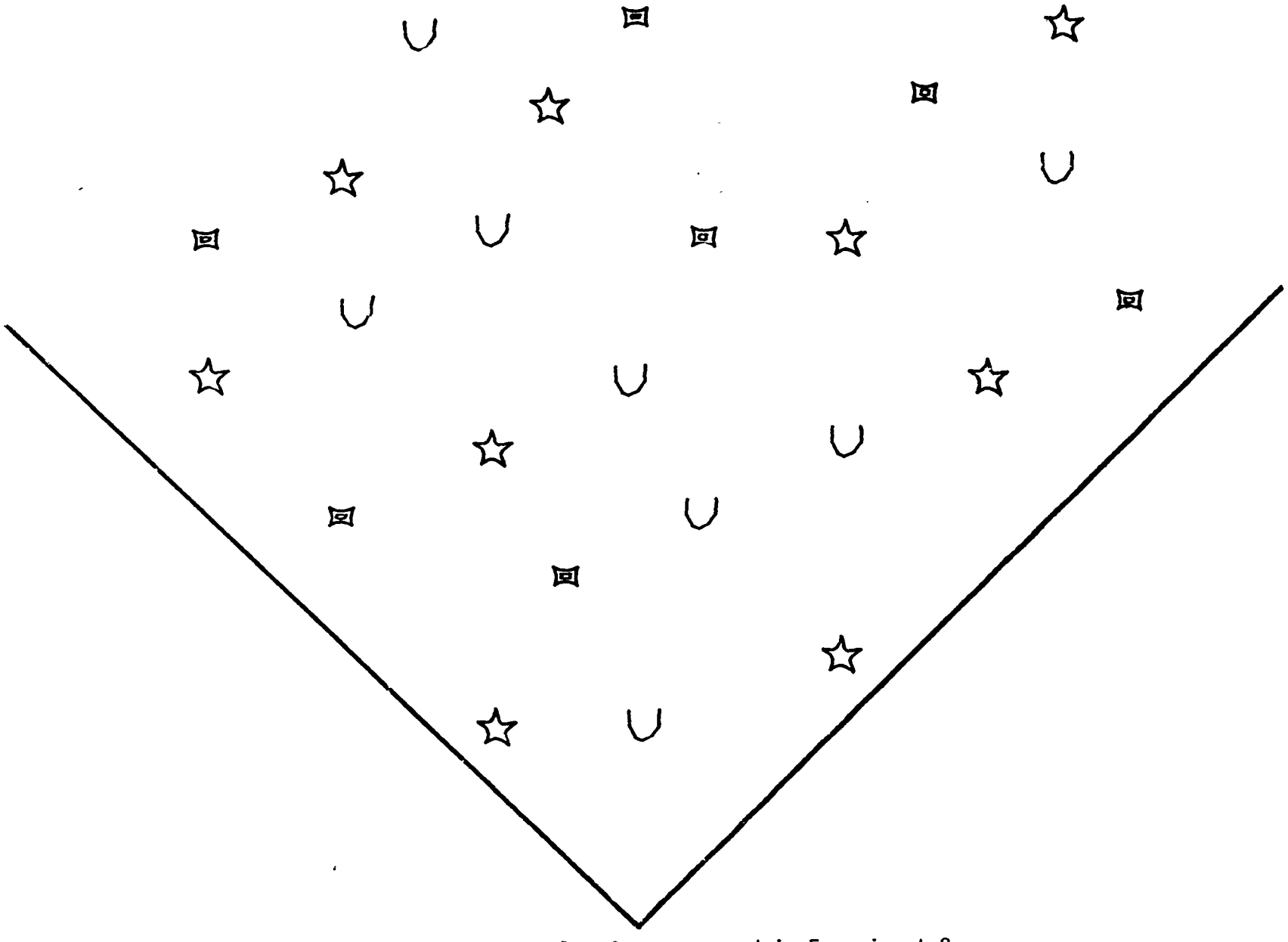


Figure 2. An example of a page used in Experiment 2

Other materials included china-marking pencils and stop-watches.

Procedure

Each subject was handed the looseleaf binder, containing the practice pages, and a china-marking pencil. The experimenter assigned a symbol as the "target" to be searched for and instructed the S to turn the front cover at the word "Go", and to find and put a line through all the representations of the designated symbol on the first page with his marking pencil. Each subject was told to pursue this routine for each succeeding page until he reached a blank one. This procedure was repeated for a symbol in each of the six sets (all three symbols in Experiment 2).

After practice was complete, each S was tested on the six sets (eight sets in Experiment 2) in a random order, following the process that was used with the practice material. The experimenter recorded reaction time to the nearest 1/10 of a second and errors of omission and commission for each block of pages.

RESULTS AND CONCLUSIONS

A Box test for homogenity of variance and covariance was performed on the reaction time data from both experiments. The obtained values of B were not significant at the .05 level.

The mean reaction time in seconds for a block of pages (symbol reaction time) was computed for each symbol. These times are presented in Table 1. Time to complete 50 pages (27 in Experiment 2) - total reaction time - was averaged for each symbol set. They were:

Experiment 1

a. Set 1 -- 217.4
b. Set 2 -- 207.7
c. Set 3 -- 217.6
d. Set 4 -- 212.2
e. Set 5 -- 209.3
f. Set 6 -- 207.3

Experiment 2

a. Set 1 -- 123.2
b. Set 2 -- 116.5
c. Set 3 -- 124.6
d. Set 4 -- 126.0
e. Set 5 -- 114.8
f. Set 6 -- 119.8
g. Set 7 -- 116.6
h. Set 8 -- 110.3

The repeated measures analysis of variance was used on the symbol reaction times within sets and by symbol type across sets for all symbols except the 6-sided U in Experiment 1. It was also utilized on the total reaction data across sets. Newman-Keuls tests were run on all the data which showed a significant difference. Tables 2 and 3 detail where the difference existed for the symbol reaction times within sets and by symbol type across sets, respectively. Set 8 was found to be significantly different from Sets 1, 3, and 4 at the .05 level in Experiment 2 when the test was run on the total reaction times.

(178)

Table 1

Symbol Reaction Times By Set

Set	Symbol 1	2	3	4	5	6	7	8
Experiment 1								
1	46.8	50.6		38.2	42.8		38.9	
2	43.9	47.4		41.8	43.7			30.5
3	45.6	47.9		39.4		40.7	43.6	
4	47.6	46.4			39.6	36.9	41.8	
5	44.8		42.7	36.1	44.2		41.4	
6		45.6	42.6	37.6	43.9		37.7	
Experiment 2								
1	45.5			36.6			41.1	
2	43.2			36.8				36.5
3	44.9				38.6		41.0	
4		45.2		41.2			39.6	
5	41.9				37.1			35.9
6		44.1		40.6				35.1
7		41.7			37.4		37.5	
8		39.2			36.8			34.3

Table 2

Results of Newman-Keuls Tests

For Symbols Within Sets

Set	Symbols
Experiment 1	
1	4 7 5 1 2
2	8 4 5 1 2
3	4 6 7 1 2
4	6 5 7 2 1
5	4 7 3 5 1
6	4 7 3 5 2
Experiment 2	
1	7 4 1
2	8 4 1
3	5 7 1
4	7 4 2
5	8 5 1
6	8 4 2
7	5 7 2
8	8 5 2

Note. Symbols underlined by a common line do not differ from each other; symbols not underlined by a common line do differ.

Table 3

Results of Newman Keuls Tests

For Symbol Type Across Sets

Symbol	Sets
Experiment 1	
1	2 5 3 1 4
2	6 4 2 3 1
4	5 6 1 3 2
5	4 1 2 6 5
6	4 3
7	6 1 5 4 3
Experiment 2	
2	8 7 6 4
4	2 1 6 4
7	7 4 3 1

Note. Sets underlined by a common line do not differ from each other; sets not underlined by a common line do differ.

Errors of omission and commission were collapsed, since very few errors of commission were committed in either experiment, and error data was summed across subjects. Table 4 presents the cumulative errors by symbol and set.

Chi-square statistics were used on the total number of errors committed by set and by symbol. Each of the two chi-squares in both experiments were significant at the .01 level.

Spearman rank correlations were computed between total reaction times and total errors by set. The coefficient for Experiment 2 was significant at .05.

In this phase of the research, symbols and sets of symbology were identified which can be easily discriminated and quickly reacted to. When the reaction time and error results were evaluated, the following findings were arrived at:

	Experiment 1	Experiment 2
a. Best friend:	Flag	Heart
b. Best hostile:	Swatstika	Collapsed Square
c. Best unknown:	6-sided U	6-sided U
d. Best sets:	2,5	8,5

REFERENCES

1. Bergum, B.O. & Burrell, W.E. Symbol confusion in fire direction systems. Consulting Report, U.S. Army Air Defense Human Research Unit, Ft. Bliss, Texas, May 1964.

2. Carter, R.J. Standardization of geometric radar symbology: Stereotyped meanings and paper-and-pencil testing. Paper presented at the 23rd Annual Meeting of the Human Factors Society, Boston, Massachusetts, October 29-November 1, 1979.

3. Davis, C.J. Radar Symbology studies leading to standardization. Technical Memorandum 5-68, U.S. Army Human Engineering Laboratories, Aberdeen Proving Ground, Maryland, February 1968.

(52)

Table 4

Errors By Symbol and Set

Set	Symbol 1	2	3	4	5	6	7	8	Total
				Experiment 1					
1	25	64		11	9		13		122
2	14	42		22	10			4	92
3	32	55		8		10	16		121
4	26	44			13	8	32		123
5	16		4	12	19		21		72
6		55	14	21	24		8		122
Total	113	260	18	74	75	18	90	4	652
				Experiment 2					
1	33			5			23		61
2	26			1				13	40
3	32				2		13		47
4		88		10			9		107
5	19				3			3	25
6		70		12				3	85
7		33			2		5		40
8		43			2			7	52
Total	110	234		28	9		50	26	457

CASSIDY, SQNLDR, Michael J. RAAF, DATKO, Louis M., and RUCK, Hendrick W., Air Force Human Resources Laboratory, Manpower & Personnel Division, Brooks AFB Texas.

OCCUPATIONAL ANALYSIS FOR DETERMINING JOB PROFICIENCY REQUIREMENTS (Wed P.M.)

This paper addresses the application of occupational analysis to determine job proficiency requirements. Using data routinely collected by the Air Force Occupational Measurement Center, innovative procedures for manipulation of data files have been developed to identify specific tasks required for proficiency in individual jobs. The development from a simple conceptual model to a more complex statistical model suitable for operational support of Air Force on-the-job training programs will be described.

OCCUPATIONAL ANALYSIS FOR DETERMINING JOB PROFICIENCY REQUIREMENTS

Michael J. Cassidy, SQNLDR, RAAF
Louis M. Datko, AlC, USAF
Hendrick W. Ruck

Air Force Human Resources Laboratory
Manpower and Personnel Division
Brooks AFB, TX 78235

Introduction

The on-the-job training (OJT) system used in the Air Force relies on two different sources of instruction. Career-wide knowledge and background are provided by Career Development Courses (CDC). Completion of CDC instruction programs is mandatory for being upgraded from the apprentice to the journeyman skill level. In addition, hands-on training must be conducted and certified by supervisors. Specialty Training Standards (STS) are used as guidelines for this hands-on training. Many supervisors have reported problems in identifying the tasks on the STS that should be trained (Stephenson & Burkett, 1975). Supervisors generally agree that training all STS tasks is impractical due to equipment and time availability while training only one task is usually insufficient. However, no system for identifying the important job tasks for hands-on OJT is operationally available. The Standardized Position Oriented Training (SPOT) system has been developed to meet this need. The conceptual model for the SPOT system has been reported elsewhere (Cassidy, Ruck, & Offutt, 1979). This paper reports on the development of an empirical model ready for field testing.

Background

To put this paper in context, some background on the SPOT concept is needed. SPOT shifts the focus of OJT from the specialty oriented STS to the job by using the results of occupational task analysis. For each job identified in the occupational analysis, a list of tasks is presented as the proficiency requirements for the job. As a supervisor checks an airman's proficiency on each of the listed tasks, he also compiles the airman's OJT requirements: those tasks on which the airman is not yet proficient. To compile the task list for each job in a specialty a conceptual model was devised (Cassidy, Ruck, & Offutt, 1979). The conceptual task selection model is a springboard for a technology which uses job analysis, detailed task information from the occupational survey program, to identify job proficiency requirements. The model uses jobs, relative time spent on tasks in the job (job relevance), and task difficulty as the starting point for the selection of tasks to specify job proficiency requirements. Figure 1 depicts the three screening processes which, when applied sucessively, define a segment of the model space representing the task list for a specific job.

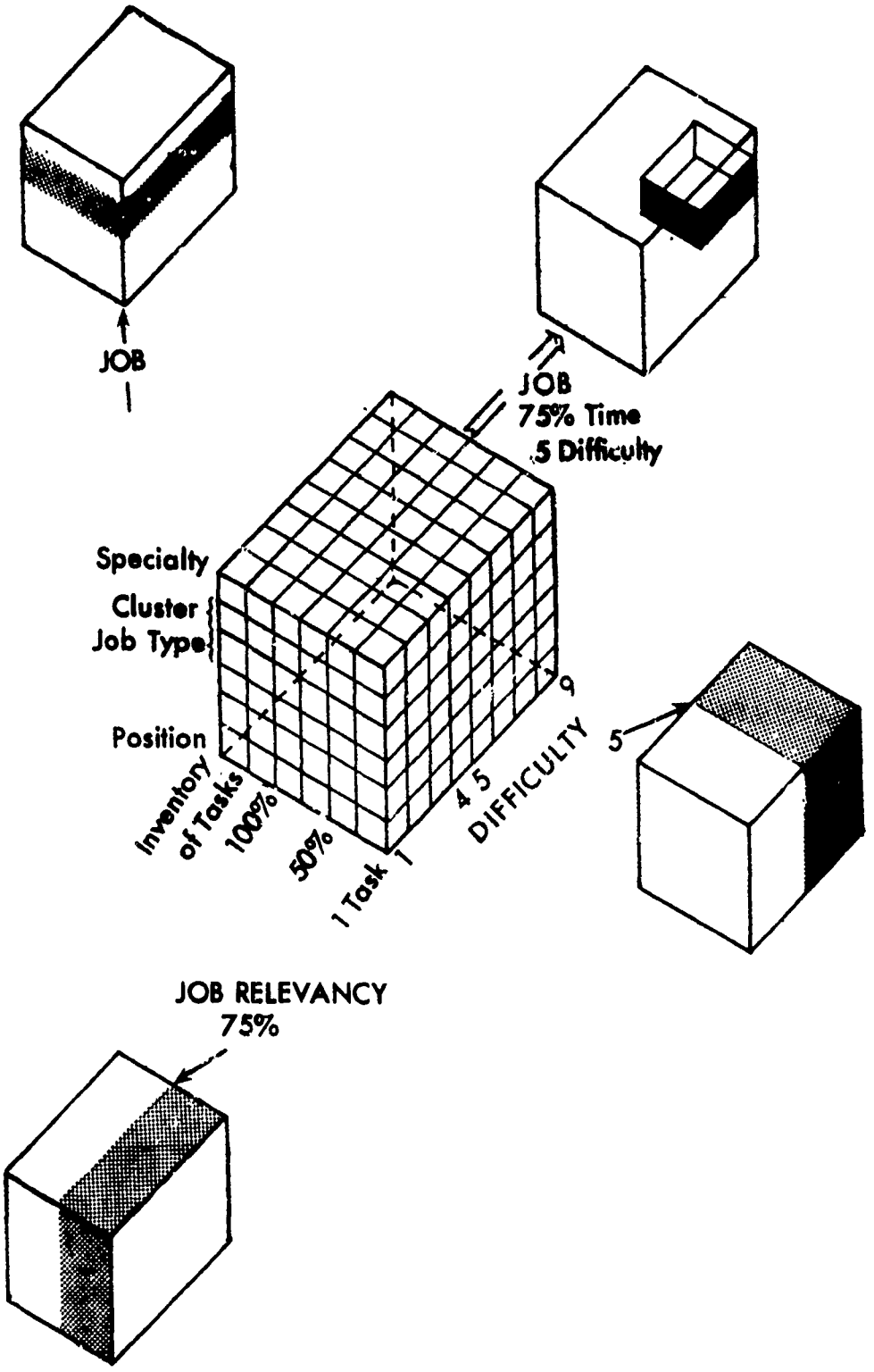


FIGURE 1. Conceptual Task Selection Model (Shaded areas represent tasks selected by each screening process, i.e., job, job relevancy, & difficulty & successive overlaying of these shaded areas intersect to define selected task list for a job.)

(186)

In the SPOT system, the task lists are compiled by automated data manipulation: initially by the three-way screening of the conceptual model; operationally by the multidimensional application of regression analysis. Thus, the SPOT goal of standardized training task lists for each job in a specialty for use in guiding hands-on OJT can be attained. Features of SPOT are: (a) tasks are listed by job; (b) proficiency (and consequently OJT) task lists are standardized throughout the Air Force; and (c) task lists are used instead of the more general STS for hands-on OJT. The SPOT system of guidance differs from the present system in three essential respects. First, the STS is replaced by a number of job-specific SPOT lists which are necessarily performance related listings. Second, the SPOT system allows for the inclusion of mandatory skills and knowledges (as defined in the Airman Classification Manual, but not covered in the STS) for each skill level of a specialty. Provision can also be made for contingency tasks, which are MAJCOM-specific or combat-related, to be included. Finally, when needed, SPOT lists can be provided for specific weapon systems.

Flexibility is inherent in the SPOT system which provides both guidance and direction. Local conditions, such as aircraft type, MAJCOM maintenance requirements, equipment, and manning, may result in particular jobs varying in task content from the SPOT specified job. Therefore, the supervisor will be able to modify the SPOT listing for a particular job within established policy guidance by deleting and adding tasks. This flexibility is the key to the utility of standardized guidance.

In brief, the STSs now in use are too general to meet OJT needs, requiring a great deal of interpretation and elaboration by working-level supervisors. The purpose of the research on the SPOT system is to develop a technology to identify specific job tasks associated with each significant portion of an Air Force specialty. In this way, job proficiency requirements and, by derivation, OJT requirements can be specified in terms of tasks associated with specific jobs.

Overview of Research Methodology

The first step in the research program was the selection of Air Force Specialties (AFSs) to be studied. Next, through automated data manipulation, the conceptual task selection model was used to provide SPOT lists (proficiency requirements) for each job in the selected group of specialties. The data collection phase involved briefing Major Command (MAJCOM) staff personnel, nomination and selection of bases, and briefing field-test base personnel. Job supervisors were then asked to review the proficiency requirements for those jobs under their supervision and to modify the automated listings according to their own perceptions of the job. Judgment analysis based on regression techniques was then applied to define empirical models that could replicate supervisors' specified requirements.

Selection of Sample of AFSs

To be able to demonstrate the potential for applying SPOT methodology to various types of Air Force specialties, considerable attention was given to

the selection of AFSs in the sample to be studied. Two attributes of Air Force specialties were deemed relevant: technicality and diversity. Using Air Force Regulation AFR 39-1 as the reference, all Air Force specialties were categorized in three broad divisions based on the degree of technical skills and experience required on the job. The category for each specialty was determined by the equipment worked on, the training required, and the experience requirements as presented in regulations. Highly technical specialties were defined as those requiring technical skills and/or experience for missile/aircraft, missile/aircraft systems, or equivalent skills; moderately technical specialties were those requiring technical skills and/or experience for non-missile/aircraft applications--for example, general purpose vehicle repair and plumbing; and non-technical specialties were those requiring little or no technical skills or experience. Diversity of an Air Force specialty was a measure of the number of jobs within the specialty. Two categories of diversity were defined: homogeneous specialties were defined operationally as those specialties containing 15 or fewer jobs for large (in population) specialties or 10 or fewer jobs for small specialties. All other specialties were classed as heterogeneous. So that data used in the SPOT research would reflect the actual work situations, the most recent OSR data available was obtained. Four specialties were selected: (a) 423X0 - Aircraft Electrical Systems (high technicality, homogenous); (b) 461X0 - Munitions Systems (high technicality, heterogeneous); (c) 645X0 - Inventory Management (low technicality, heterogeneous); and (d) 645X0A - Inventory Management Munitions shredout (low technicality, homogeneous).

SPOT List Construction

SPOT listings based on the conceptual model were generated for each of the four specialties. For each job in the four specialties, four different versions of SPOT listings were constructed. The SPOT lists from this conceptual model were not expected to be definitive. They were developed to expedite the OJT process by giving, for each job, preliminary task lists which would guide the knowledgeable supervisor in making logical and sensible decisions about job proficiency requirements. To allow for the inclusion of the more difficult tasks in the position description which did not get through the job relevancy screen, the field supervisor was given freedom to add tasks to a list as his knowledge of the job and task, and his experience, indicated.

A computer program was designed to format the task list in a manner similar to the currently used Job Proficiency Guide format with the exception that the level of proficiency required was deleted: the SPOT concept postulates that full proficiency for job performance was the only criterion required; partial proficiency on a task is not recorded in the SPOT system. The tasks were ordered from the least difficult to the most difficult within the listings. This presented the supervisor with a list of tasks based on the likely number of members performing those tasks so that he would probably delete tasks from the end of the list. (Generally, the higher the task difficulty rating is, the lower is the percentage of members performing the task).

Data Collection

The next phase of the research involved a detailed briefing to Headquarters USAF and functional managers at three MAJCOMs on the SPOT concept and the requirements for field evaluation of the conceptual model SPOT lists. Each MAJCOM then nominated four bases at which field evaluations could be obtained. On the basis of the distributions of seven- and nine-level supervisors, three bases from each MAJCOM were selected. At each base, management level staff were provided with a briefing on the SPOT concept and the research to be carried out. Supervisors were then briefed on the concept and the procedure to review and modify the SPOT lists. They were instructed to examine first the list of jobs within their specialty and to select the job or jobs which they were then supervising. Next, they evaluated the SPOT list for each of these jobs and deleted tasks for which they personally would not require proficiency. Supervisors then reexamined the SPOT list to see which tasks were not included in the list, but for which they would require proficiency. Such tasks were added to the list. To aid in their wording of these tasks, each supervisor was given a complete task inventory for that particular specialty and requested to use task statements from the inventory whenever possible. Data were gathered from 186 field supervisors.

Results

Analysis of Data

As a broad indication of responses received, Table 1 summarizes the average number of supervisors who participated in evaluating SPOT lists in each of the specialties, and the average number of recommended tasks in each job, the average number of tasks included by all supervisors, and the average number of tasks excluded by all. In one specialty, supervisors consistently reported jobs which had not been included in the job types reported from the Occupational Survey Report (OSR). Because this was so consistent across all bases, a reanalysis was done of the career field. Each of the jobs reported, with minor exceptions, was found to be a subgroup of a group reported as a job by the Occupational Measurement Center (OMC). Discussion with the original analyst at OMC indicated that the purpose for which the analysis had been carried out did not require as fine a job typing as would be necessary for SPOT purposes. From this it appeared that the analysis techniques used at OMC would be satisfactory in implementing SPOT as an operational system, provided that the occupational analyst was made aware of the purpose for which his analysis would be used. The response from these supervisors, coupled with the satisfaction of supervisors in the other three specialties with the job types identified in their specialties, indicates that specifying jobs at OMC job type levels is acceptable. Therefore, in developing the operational system, jobs will be specified in the routine job analyses done by OMC.

Judgmental analysis of supervisors' responses was performed in three steps using standard mathematical programs, Comprehensive Occupational Data Analysis Programs (CODAP) (a summary of the functions of the CODAP programs used is given in Appendix A), and special-purpose programs. Responses were scored as binary data: each task included in a supervisor's list was coded as 1; all

Table 1

Summary of Supervisors' Modification to SPOT Conceptual Model Lists

AFSC	# of Jobs	N_T [a]	N_s [b]	N_t [c]	Average # of Tasks Included by All	Average # of Tasks Excluded by All
423X0	6	559	1.3	25.6	4.0	469.8
461X0	27	435	4.0	32.8	13.8	374.9
645X0	18	691	7.4	23.4	8.1	629.4
645X0A	10	691	2.3	55.8	47.4	629.5

[a] N_T is the number of tasks in the total inventory

[b] N_s is average number of supervisors per job

[c] N_t is the average number of tasks included in modified lists

other tasks in the inventory were coded 0. Using a special purpose program, the dichotomous data were prepared for analysis together with job type and task factors (from Job Specials) data. (One specialty lacked data on one of the task factors but had data on two different factors; these data were included and dummy vectors constructed so that all four sets of specialty data remained comparable.) Using the dichotomous response data as the criterion, a standard correlation and regression program (TRICOR) was used to perform a regression analysis. Ten basic predictor variables, ten squared variables and nine interaction variables were considered in model development. Seven regression problems were solved for each rater. This process confirmed that job data was most important and specialty data, while making a contribution, was least important. The variability in the effect of square and interaction terms from rater to rater suggested further work should retain all twenty-nine terms of the full model (see Appendix A for a list of the variables in the empirical model). The goodness of fit [1] of the initial regression equations is indicated in Table 2.

Acceptable levels of goodness of fit were achieved with the least squares analysis. The intent of the follow-on analysis was to examine the deterioration in the goodness of fit caused by the grouping of individual models into a single specialty model. Thus, the second step in the analysis used another special program to aggregate raters by specialty giving one criterion vector for each specialty. By executing the TRICOR program on each group, a regression equation was generated for each specialty. The goodness of fit of these regression equations was still satisfactory; R^2 for

[1] Because the dependent variable (criterion) is dichotomous, the goodness of fit of the regression model is degraded compared to that for a continuously distributed dependent variable.

Table 2

Accuracy of Individual Rater Regression Equations by Specialty

AFSC	Category of Specialty[a]	N Rater	R^2 Range for Individual Models	Mean	SD
423X0	T/Ho	68	.1868* to .8735*	.5754	.2425
461X0	T/He	107	.0858* to .8901*	.5958	.1417
645X0	NT/He	134	.1971* to .9688*	.5836	.1591
645X0A	NT/Ho	23	.5379* to .7614*	.6600	.0702

[a] Categories are: T - technical, NT - non-technical; HO - homogenous, He - heterogeneous

*p .01

the specialties were: 423X0 - .3693; 461X0 - .4663; 645X0 - .3766; and 645X0A - .5767. See Appendix B for the form of the regression equation and the standard weights for each specialty equation. These regression equations form the empirical model are being used in the on-going research on the SPOT system.

In the third step of the analysis, a predicted SPOT score was generated for all tasks within jobs across each of the four specialties using each specialty's regression equation[2]. The specialty regression equations were found to correspond highly with actual judgments made by supervisors about specific jobs (see Table 3). The low values of R^2 for 645X0A are due primarily to the instability of the actual judgment policies which were based on a mean of 2.3 raters per job (refer to Table 1). The SPOT scores were then used in the CODAP program, FACPRT to generate the automated SPOT lists by including all tasks with a SPOT score greater than or equal to .10. This SPOT score cut-off was specified to produce lists which included those tasks commonly included by supervisors and which were slightly longer than supervisors' lists. Summary task factor data were also displayed to aid a final staffing of the SPOT lists before field use (see Appendix C for a sample SPOT list). As an aid to managers, an executive summary was also produced showing tasks common to SPOT lists for half or more of the jobs in an AFS (see Appendix D).

[2] CODAP programs used for this phase of the analysis were FACGEN, DECDEC, and FACPRE.

Table 3

Accuracy of Specialty Regression Equations Applied to Jobs

AFSC	N Jobs	R^2		
		Range	Mean	SD
423X0	6	.3173 to .8000	.5776	.1875
461X0	27	.3324 to .7916	.6011	.1415
645X0	18	.3651 to .7261	.5779	.1165
645X0A	10	.0182 to .3878	.1750	.1395

Summary

The objective of the SPOT system, to generate automated job proficiency lists, has been substantially realized. Through the application of judgment analysis, regression equations have been derived for each of the test AFSs. When these lists have been staffed, a further step in the analysis will be completed: investigating the feasibilty of generating a single regression equation which can be applied to all specialties. Concurrently, cross-validation of the regression equations will also be performed. The staffed lists will be field tested to aid in an independent evaluation of the SPOT system. The evaluation results and experience gained in the field shall provide a better definition of requirements for the operational SPOT system. This on-going research promises to make the task of the supervisor in OJT easier while improving the quality and tracking of hands-on training.

References

Cassidy, M.J., Ruck, H.W., & Offutt, S.V. Task Selection For Job Proficiency and Training. Paper presented to 21st Annual Conference of the Military Testing Association, US Navy, San Diego, CA, October 1979.

Stephenson, R.W., & Burkett, J.R. On-The-Job-Training in the Air Force: A System Analysis. AFHRL-TR-75-83. Lowry AFB, CO: Technical Training Division, Air Force Human Resources Laboratory (AFSC), December 1975.

APPENDIX A: Summary of CODAP Program Functions Used for SPOT

JOBSPC

Given the membership criteria in terms of computed or background variables, the Composite Job Description (JOBSPC) program identifies all the cases meeting those requirements and computes a composite job description for that group. All associated data may be stored on the job description file for further use.

FACGEN

The Factor generator (FACGEN) program may be used to modify and/or load task factors for future processing with the CODAP system. Input and/or output may be either standard task factor decks or request cards identifying vectors on the job description file. Optional operations for modifying the input vectors include (but are not limited to): (a) raise values to a specified power (2-9); (b) standardized values so that means = 5.0 and SD = 1.0; (c) substitute rank-orders for input values; (d) substitute values rescaled as percent of range; (e) expand or contract task list with a task category; (f) dichotomize or categorize the task inventory.

DECDEC

The Mathematically Interact Two Decks (DECDEC) program will accept two task factor decks as input and perform any simple arithmetic operation between them. This program may be used to add, subtract, multiply, divide, or exponentiate two task factors.

FACPRE

The Predicted Factors from Regression Equations (FACPRE) program will apply the regression equations developed elsewhere and produce a task factor deck representing this predicted factor. The following items are reported: titles for criterion including titles for the input vectors; the number of observations; product-moment correlation and product-moment correlations squared; and the standard error of the estimate. The following are reported in columnar format for easy comparison of the criterion versus the predictor factor: mean moment about the mean, standard deviation, coefficient of variation, minimum value, and maximum value.

FACPRT

The Task Factor Print (FACPRT) program allows the user to print any of the vectors on the Job Description file. In addition, the program is capable of calculating and reporting differences between vectors, maximums or minimums of sets of vectors, cummulative percentage, and categories of tasks by using any one of six arithmetic operators. The inclusion of sequence numbering and blank columns, sorting, as well as print suppression are under user control. If used, the suppression and category options automatically provide heading lines indicating the limitations being used. For SPOT Task Lists, a modification to the program was used to permit user formatting of the task factor vectors. Program identified task categories may be punched and added to the Job Description file for future references. Several types of reports may be produced; they include: (a) report by task, (b) report by task within duty, and (c) report by task within module.

APPENDIX B
Table B1 Variables in the Empirical Model

VAR NO	TITLE	ABBR	SYMBOL
1	Percent Members Performing Within Job Type	PMPJOB	M_J
2	Percent Time Spent By All within Job Type	PTSJOB	T_J
3	Percent Of All Performers With Job Type	PAL	A
4	Percent Of Total AFSC Members Performing	PMP	M
5	Percent Time Spent By Total AFSC	PTS	T_T
6	Average Percent Grade in Total AFSC	AVPGR	G
7	Average Task Difficulty Rating	ATDIFF	D
8	Average Training Emphasis Rating	ATREMP	E
9	Average Consequences Of Inadequate Performance	ACIP	C
10	Average Task Delay Tolerance Rating	ATOT	T
11	(Percent Members Performing/Job Type)**2	PMPJSQ	M_J^2
12	(Percent Time Spent By All/Job Type)**2	PMSJSQ	P_J^2
13	(Percent Of All Performers/Job Type)**2	PALSQ	A^2
14	(Percent Of Total Members Performing/AFSC)**2	PMPSQ	M_T^2
15	(Percent Time Spent By Total/AFSC)**2	PTSSQ	T_T^2
16	(Average Percent Grade In Total AFSC)**2	APGRSQ	G^2
17	(Average Task Difficulty Ratings)**2	DIFFSQ	D^2
18	(Average Training Emphasis Ratings)**2	EMPSQ	E^2
19	(Average Cons Of Inadequate Perform)**2	CIPSQ	C^2
20	(Average Task Delay Tolerance Rating)**2	TDTSQ	T^2
21	% Members Performing/Job By Average % Grade	V3 * V5	$M_J * G$
22	% Members Performing/Job by Task Difficulty	V3 * V9	$M_J * D$
23	% Members Performing/Job By Training Emphasis	V3 * V10	$M_J * E$
24	Member Performing/Job By CONS Of Perform	V3 * V11	$M_J * C$
25	% Member Performing/Job by Task Delay	V3 * V12	$M_J * T$
26	Average % Grade by Task Difficulty	V8 * V9	G * D
27	Average % Grade By Training Emphasis	V8 * V10	G * E
28	Average % Grade By CONS Of Inadequate Perform	V8 * V11	G * C
29	Average % Grade By Task Delay Tolerance	V8 * V12	G * T

(194)

Table B2. AFS Regression Equations - The Empirical Model

Form of the Empirical Model

$$Y=a_0U + a_1M_J + a_2T_J + a_3A + a_4M_T + a_5T_T + a_6G + a_7D + a_8E + a_9C + a_{10}T + a_{11}M_J^2 + a_{12}T_J^2 + a_{13}A^2 + a_{14}M_T^2 + a_{15}T_T^2 + a_{16}G^2 + a_{17}D^2 + a_{18}E^2 + a_{19}C^2 + a_{20}T^2 + a_{21}M_J*G + a_{22}M_J*D + a_{23}M_J*E + a_{24}M_J*C + a_{25}M_J*T + a_{26}G*D + a27G*E + a_{28}G*C + a_{29}G*T + ERROR$$

Regression Weights

	AFSC			
VAR	423X0	461X0	645X0	645X0A
1	-.011855	-.013437	-.014870	-.009565
2	.057829	.123549	.146133	.320033
3	-.000430	-.000173	-.000167	-.001078
4	.000572	-.000725	-.004620	-.009238
5	.116640	.053719	.133343	.898147
6	.040997	.013549	.037847	-.000131
7	.035533	.047187	.001349	.087667
8	.003053	NA	.000431	-.026684
9	NA	.041571	NA	NA
10	NA	-.004445	NA	NA
11	.000057	.000050	.000037	-.000066
12	-.005385	-.015950	-.013353	-.030363
13	.000006	.000032	.000005	.000007
14	-.000035	.000031	.000066	.000068
15	.081342	-.044894	-.076477	-.563883
16	-.002800	-.000825	-.001026	-.000662
17	-.002467	-.002542	.001355	-.008502
18	-.004571	NA	-.000135	-.000981
19	NA	-.005155	NA	NA
20	NA	.000001	NA	NA
21	.000387	.000476	.000140	.000577
22	.002253	.003149	.003261	.003407
23	.000417	NA	.000592	.000309
24	NA	-.000355	NA	NA
25	NA	-.000081	NA	NA
26	-.002821	-.006228	-.004939	.000007
27	.001829	NA	-.000871	.005356
28	NA	.002949	NA	NA
29	NA	.001858	NA	NA
Const	-.192808	-.243177	-.072275	-.206378

Note. NA appears in the table where a dummy vector was used in the analysis because data was not available for the AFS/task factor.

(196)

APPENDIX C: SAMPLE SPOT LISTING

TASK FACTOR INFORMATION ON SPOTS STUDY FOR 423X0 — 423X0 — PAGE 2

TASK FACTOR DECKS FOR MAJCOMS MAC, TAC & SAC ORDERED ON PREDICTED SCORE
423X0 JOBGRP 029 COMBAT LOGISTICS SUPPORT SPECIALIST
BLANK COLUMN = NO MBRS FOR JOBGRP 029 IN MAJCOM TAC

D	TSK	TITLES	SEQ NUM *	SPOTS SCORE 029	%MEM JOB 029	%MEM TOTL AFSC	%MAC %MEM JOB	MAC %MEM AFSC	SAC %MEM JOB	SAC %MEM AFSC	TAC %MEM JOB	TAC %MEM AFSC
J	450	PERFORM PROTO-TYPE TCTO'S	1	.766	73.9	16.8	50.0	16.9	.0	18.9		17.8
J	452	PERFORM TCTO MODIFICATIONS OF AIRCRAFT ELECTRICAL SYSTEMS	2	.763	87.0	59.7	50.0	56.9	50.0	67.2		63.7
J	524	REWIRE AIRCRAFT ELECTRICAL SYSTEMS	3	.708	65.2	62.9	50.0	69.8	.0	64.9		67.9
J	451	PERFORM SOLDERLESS CONNECTOR INSERTIONS OR EXTRACTIONS	4	.438	65.2	62.5	50.0	70.3	50.0	61.5		70.6
J	438	CRIMP WIRES TO CONNECTOR PLUGS, CONTROL BOXES ON CONTROL PANELS	5	.417	69.6	72.5	100.0	80.6	100.0	76.5		81.1
C	67	PREPARE OR INDORSE AIRMAN PERFORMANCE REPORTS (APR)	6	.406	39.1	32.4	.0	28.7	.0	24.2		30.1
J	445	FABRICATE WIRING HARNESS	7	.388	52.2	42.2	100.0	49.1	.0	43.3		39.8
J	405	ASSEMBLE OR DISASSEMBLE CONNECTOR PLUGS	8	.353	52.2	58.7	50.0	66.2	50.0	64.2		60.7
J	495	REMOVE OR INSTALL PINS ON CONNECTOR PLUGS	9	.310	47.8	68.3	.0	79.4	.0	69.5		73.2
J	440	FABRICATE COMPACT WIRE BUNDLES	10	.293	39.1	32.8	100.0	37.3	.0	35.6		33.4
J	465	REMOVE OR INSTALL CONNECTOR PLUGS	11	.256	43.5	66.6	50.0	77.3	100.0	71.6		70.8
J	453	POT CONNECTORS OR RELAYS	12	.255	52.2	54.1	50.0	35.0	50.0	62.2		61.3
J	519	REPAIR COMPACT WIRE BUNDLES	13	.237	30.4	45.6	.0	44.8	.0	43.5		56.5
J	528	SOLDER WIRES TO CONNECTOR PLUGS, CONTROL BOXES OR CONTROL PANELS	14	.184	43.5	70.4	50.0	81.6	50.0	73.9		74.3
F	121	OBSERVE IN-PROCESS MAINTENANCE OR MAKE ON-THE-SPOT CORRECTIVE ACTIONS	15	.180	21.7	34.0	.0	32.7	50.0	32.3		31.4
J	422	CLEAN CONNECTOR PLUGS	16	.177	47.8	66.2	.0	77.6	100.0	67.7		71.6
J	441	FABRICATE ELECTRICAL LEADS	17	.166	39.1	52.4	100.0	57.4	50.0	56.3		53.4
F	123	PERFORM SPECIAL INSPECTIONS OF AIRCRAFT ELECTRICAL SYSTEMS	18	.135	17.4	45.1	.0	45.1	.0	44.6		52.1
F	125	VISUALLY INSPECT AIRCRAFT ELECTRICAL SYSTEMS FOLLOWING MAINTENANCE	19	.132	17.4	45.1	.0	42.1	.0	44.6		49.7
G	173	VISUALLY INSPECT FIRE AND OVERHEAT DETECTION CIRCUIT COMPONENTS	20	.126	21.7	71.8	.0	78.1	.0	74.9		77.6
F	117	ADVISE MAINTENANCE PERSONNEL ON INTERPRETATION OF MAINTENANCE PROCEDURES	21	.113	17.4	21.7	.0	18.1	.0	20.5		15.2
B	41	SUPERVISE AES SPECIALISTS (AFSC 42350)	22	.107	13.0	28.6	.0	29.7	.0	25.7		30.8
J	475	REMOVE OR INSTALL FIRE OR OVERHEAT CABLES	23	.103	21.7	51.8	.0	72.0	.0	24.6		65.5

TASKS OMITTED FOR WHICH:
THE VALUE IN COLUMN SPOTS SCORE 029 IS LT .100

APPENDIX D: SAMPLE EXECUTIVE SUMMARY PRINTOUT

EXECUTIVE SUMMARY OF SPOT LISTS FOR 645X3 — EXCSUM PAGE 2

TASKS ARE ORDERED ON COMMONALITY INDEX – % OF JOBS IN WHICH TASK OCCURS AND ARE TRUNCATED WHEN COMIND < 50%; VALUES INDICATE THAT A TASK IS IN SPOT LIST FOR A JOB: VALUES = PERCENT OF MEMBERS IN JOB PERFORMING TASK

D	TSK	TITLES	SEQ NUM *	COM IND *D*	CMP 028 (F)	CPM 030 (F)	CPM 055 (F)	CPM 058 (F)	CPM 064 (F)	CPM 076 (F)	CPM 083 (F)	CPM 105 (F)	CPM 143 (F)
B	3	DEVELOP OR IMPROVE WORK METHODS OR PROCEDURES	1	83	85.	17.	21.	83.	47.	18.	40.	33.	
A	3	DETERMINE WORK PRIORITIES	2	78	76.	14.	27.	81.	42.		40.	26.	
E	25	RESEARCH PUBLICATIONS FOR GENERAL SUPPLY POLICIES OR PROCESURES	3	78	63.	21.	27.	59.	43.		25.	19.	
M	1	ACT AS CONTACT POINT FOR SUPPLY CUSTOMER PROBLEMS	4	78	17.	11.	16.	33.	23.	6.		9.	
D	5	CONDUCT ON-THE-JOB TRAINING (OJT)	5	72	39.		24.	42.	41.			26.	
E	2	ESTABLISH OR MAINTAIN SUSPENSE FILES	6	72		38.	37.	38.	40.			76.	
B	25	DRAFT CORRESPONDENCE	7	67	89.	15.	19.	89.	57.			17.	
D	8	DEMONSTRATE HOW TO LOCATE TECHNICAL INFORMATION	8	67	37.	13.		30.	22.			12.	
F	15	OPERATE REMOTE KEYBOARD UNITS	9	67			43.		65.		50.	33.	100.
G	34	PREPARE INQUIRY INPUTS	10	67	61.		40.		35.		30.	24.	100.
E	23	RESEARCH CATALOGS OR TECHNICAL PUBLICATIONS FOR SUPPLY TRANSACTION DATA	11	61	24.	24.		24.	36.			19.	100.
A	19	PLAN WORK PRIORITIES	12	56	63.		19.	65.	28.		20.		
B	1	CONDUCT OR PARTICIPATE IN STAFF MEETINGS	13	56	78.			86.	32.			29.	
B	2	COUNSEL SUBORDINATES ON PERSONAL OR MILITARY PROBLEMS	14	56	80.		33.	83.	37.		25.	26.	
B	33	INDUCTRINATE NEWLY ASSIGNED PERSONNEL	15	56	72.		33.	72.	32.			21.	
E	6	MAINTAIN CORRESPONDENCE FILES	16	56	35.		22.	36.	40.			21.	
E	22	RESEARCH CATALOGS OR TECHNICAL PUBLICATIONS FOR ITEM IDENTIFICATINO AND CLASSIFICATION	17	56		38.		26.	30.				100.
E	28	RESEARCH PUBLICATIONS FOR SUPPLY SYSTEMS PROCEDURES	18	56	54.		16.	33.	19.				
E	29	RESEARCH SUPPLY TRANSACTION DATA SUCH AS ITEM IDENTIFICATION	19	56				25.	24.			21.	100.
F	10	OPERATE KEYPUNCHES	20	56	41.		29.		38.		30.	21.	100.
B	36	INTERPRET POLICIES, DIRECTIVES, OR PROCEDURES FOR SUBORDINATES	21	50	76.		14.	72.	29.		15.		
B	40	SUPERVISE APPRENTICE INVENTORY MANAGEMENT SPECIALIST (AFSC 64530) PERSONNEL	22	50	26.		22.	26.	24.			19.	
C	25	WRITE ON INDORSE AIRMAN PERFORMANCE REPORTS (APR)	23	50	70.		25.	72.	33.		20.	17.	
N	10	ISOLATE CAUSES OF COMPUTER REJECTS	24	50	39.		14.		19.				
P	22	PREPARE TURN-IN DOCUMENTS	25	50			21.				45.		100.

TASKS OMITTED . OR WHICH:
THE VALUE IN COLUMN COMIND(D) IS LT 50.0

CAS-13

197

CICCHINELLI, L.F., Ph.D., Denver Research Institute, University of Denver, Colorado.

FACTORS LIMITING THE MEASUREMENT OF SIMULATOR TRAINING EFFECTIVENESS (Wed A.M.)

This paper focuses on some problems encountered in an evaluation of a simulated test station used for training intermediate level F-111 avionics personnel. The purpose of the research effort was to determine the relative training and cost-effectiveness of a three-dimensional simulator as compared to actual test station equipment.

In recent years it has become accepted, both in the military and civilian sectors, that simulators are viable and effective training tools. A review of the available literature, on the other hand, shows that information concerning the effectiveness of simulator training based on formal evaluation studies which rely on quantifiable data is quite limited. The available information on factors which influence simulator training effectiveness is usually based on subjective measures and often focuses only on the psysical characteristics of the simulator. Clearly, factors such as psychological and physical fidelity are important considerations. However, the observations made during this evaluation suggest that a number of other, less obvious factors inherent in the simulator training environment are probably directly responsible for the lack of valid quantitative assessments of simulator training effectiveness.

The specific factors discussed in this paper are training objectives, teaching styles, user acceptance, indicators of performance, and the intended role of the simulator in the overall training program. These issues are considered critical because they often limit the strength of the evaluation designs employed and reduce the reliability of the training effectiveness measures developed. It is suggested that if these factors are considered in the planning and design phases of future simulator development programs, the measurement of simulator training effectiveness would be much more reliable and valid.

FACTORS LIMITING THE MEASUREMENT OF MAINTENANCE SIMULATOR TRAINING EFFECTIVENESS

Louis F. Cicchinelli, Ph.D.
Denver Research Institute

Introduction

The purpose of this paper is to discuss some problems encountered in the instructional environment during an assessment of the training effectiveness of a maintenance simulator. It is suggested that more anticipation of these problems during the evaluation planning stage will not necessarily reduce their impact on the assessment or findings. Rather, it may be necessary to manipulate the training environment in order to determine the training benefits of simulation.

The Air Force currently trains its maintenance technicians in classroom settings which focus on both theory and practical experience. Upon successful completion of the training sequence, these airmen are assigned to field positions where their primary job is to isolate and repair faults in aircraft components. Associated with this task is the operation and maintenance of the test stations which are used in the diagnostic testing of an aircraft's Line Replaceable Units (LRUs).

Traditionally, the primary teaching aids used in the classroom have been operational test stations designed for field use rather than for training. In the past few years, however, extensive efforts have been made to utilize simulation techniques and equipment in providing maintenance training in the military. The decision to proceed with the development of maintenance training simulators has been motivated by many of the same factors that have stimulated the development of flight training simulators. Specifically, when compared to actual equipment trainers (AETs), simulators are expected to have the advantages of lower cost of purchase and operation, higher reliability, reduction of danger to unskilled trainees, reduced noise levels in the training area, and the increased availability of the actual equipment for operational use. An additional factor motivating the application of simulators to maintenance training is the potential for increased instructor control over the nature of the practical trouble-shooting training. In retrospect, this may be the single most important consideration underlying the decision to investigate the utility of alternative simulator designs and uses for maintenance training.

Differences in Flight and Maintenance Training Environments

Available literature on the use of simulators specifically for maintenance training is very limited. There is, on the other hand, a vast amount of information concerning the development and use of simulators for flight training which is clearly helpful in developing and testing maintenance training simulators. Despite the abundance of relevant information, quantitative assessments of simulator training effectiveness are conspicuously absent from the literature. After a review of ten Air Force simulator training programs, Caro (1977a) concluded, "that most . . . had not been subjected

to formal studies that would establish their training effectiveness in quantitative terms." Reported studies of maintenance training simulation often cite hands-on experience, reliability, safety, modifiability, and cost as benefits of simulators over actual equipment trainers. It is interesting to note, however, that simulator-trained student performance is usually reported as equivalent to that of actual equipment-trained students (Cicchinelli, et al., 1980; Daniels, et al., 1975; Hurlock & Slough, 1976; Wright & Campbell, 1975). It can be argued that these evaluations have not been sensitive enough to measure the improved performance of simulator-trained students.

Equivalent performance may be acceptable for flight training simulators because of the dramatic benefits realized in the areas of cost and safety. In the application of simulation to flight training, the cost savings has been reported at about 5:1 in favor of simulators by Orlansky and String (1977). The increase safety of training conducted on flight simulators is indisputable. Similarly, maintenance training simulators have been shown to result in cost savings and increased safety, but the magnitude of these benefits alone has not been sufficient to justify a major shift in Air Force training policy. Improvements in student performance and/or reduced training time must also be demonstrated as significant benefits.

Caro (1977b) points out that while flight simulators have become quite sophisticated, the effective use of these devices is limited by a number of factors in the training system. He suggests, for example, that the lack of communication between designers and users tends to result in simulators which are insensitive to the training process. In short, simulators are designed to simulate rather than train because they have not been developed in view of behavioral considerations related to how the simulator will be used. The lack of emphasis on simulator training technology pervades the design, test, and implementation phases of simulator development. To date the factors of cost and safety have precluded the need to demonstrate measurable performance increments for the basic application of flight training simulators. However, as more complex training options are added to basic flight simulators, it is likely that a demonstration of improved performance will also be needed to justify added costs. Thus, the issues discussed in this paper are likely to become relevant when measuring the training effectiveness of newer flight simulators.

In the flight training environment, the trainer simulates the aircraft in the true sense of the word. That is, it is possible to simulate all activities necessary for flight without actually flying an aircraft. To highlight the difference between the flight and maintenance training environments, it is useful to note two important characteristics of the flight training environment. First, in its most basic application, a flight simulator is a replacement for an actual equipment trainer and the actual equipment and simulator are clearly distinguished by their ability (or inability) to fly. Second, the objective of flying is defined by a set of observable tasks and proficiency at these tasks can be directly measured. In order to fly an aircraft, the operator must *modify* a basic learned skill in response to known circumstances.

By contrast, in the maintenance training environment a simulator is generally a supplement to an actual equipment trainer. In this setting, it is difficult to discriminate between simulators and actual equipment by such an obvious difference as "ability to fly." Further, with the exception of a few clearly defined procedural tasks, a major objective of maintenance training simulators is to teach the process of trouble-shooting--the identification of faults in malfunctioning aircraft components. This process is not easily described because the set of possible problems is infinite and always unknown at the outset. To trouble-shoot an aircraft component, an operator must *select* a series of responses from a set of learned skills in an effort to identify a fault. It is these relatively unspecified aspects of the maintenance training environment that offer a challenge to researchers who attempt to substantiate the utility of maintenance simulators as viable training devices.

Factors Limiting the Measurement of Training Effectiveness

The issues presented in this section are discussed in the specific context of the Air Force maintenance training and field environment in which the evaluation took place. Each of the problems cited is accompanied by a discussion of the solution employed. It is important to note that the assessment of the simulator's training and cost effectiveness were part of a larger ongoing Air Force simulation program. The simulator itself was designed in view of empirically determined specification. It should be emphasized that the test station simulated and the training environment do *not* constitute a unique situation. On the contrary, the test station selected for simulation was considered to be representative of the entire class of automatic test stations.

Planning a Maintenance Simulator Assessment

The information presented in the following section was obtained during a training effectiveness study of a maintenance simulator. The discussion focuses on factors found to limit measurement of the training effectiveness. At some level the issues are obvious and would be anticipated during the planning stage of an assessment by any competent researcher. This environment is constantly being changed to meet new training demands. However, it is also adjusted almost daily in an effort to maintain a consistent training level, despite fluctuations in the operational status of training equipment. It is now apparent, however, that no amount of planning, interviewing, and surveying would have accurately described the dynamic and relatively unspecified maintenance training environment as it existed during the evaluation period.

First, more often than not, an evaluation effort is likely to be designed in the absence of relevant background information. Further, once a specific approach is approved, it is not usually acceptable to immediately insert a lengthy observation and design refinement phase into the time schedule. The resulting extensive delays in data collection would reduce the timeliness of the findings with respect to the overall program effort and extensive changes in the approved approach would require justification and reapproval--a lengthy process also. Second, the introduction of a

simulator into the training environment, without removing the existing equipment, alters the environment. The simulator provides an alternate training device which can be substituted for malfunctioning actual equipment or used simultaneously to increase the training capacity as needed. In sum, the simulator is accepted as added equipment and maximizing its effectiveness through new teaching practices is not a primary concern. In some instances, the simulator may even be altered to more closely resemble the actual equipment so that established teaching methods and curricula can be maintained.

The following discussion is meant to provide specific examples of unanticipated problems encountered in a training effe iveness study of a maintenance simulator. It also challenges the basic assumption that simulator training effectiveness can be determined through one assessment of one instructional environment. The discussion supports the suggestion that simulator training effectiveness studies should be implemented in at least two distinct phases. Phase I should be used to examine the impact of incorporating a full-scale operational simulator into the training environment, to observe and document the dynamics of the environment itself, to specify the actual opportunity for simulation (if any) and needed equipment modifications to realize that approach, and to develop hypotheses concerning the proper use of simulation for training. Phase II can focus more directly then on a comparative analysis of those simulation techniques and strategies considered to have potential--based on the findings of Phase I. *A Phase II assessment presumes a significant modification in the actual training environment consistent with the hypothesis in question.*

Training Objectives

In order to appropriately assess the training effectiveness of a simulator, the training objectives used to assess effectiveness must be the same as those used to specify the simulator design. However, training must keep abreast of rapidly changing aircraft technology. Thus, if excessive delays occur between the design and implementation of a training simulator, it is likely that some capabilities will be outdated and new training needs will be identified.

Training on the AET occurred as part of a 23-week intermediate level avionics maintenance course. The objective of this course and the associated Specialty Training Standards (STS) has been in a continuous state of change over the past few years. A number of factors contributed to the need to modify course objectives and content. Perhaps the single most important factor is the evolution of the aircraft itself. As more sophisticated models are developed, course content must be modified to include instruction in the operation and maintenance of updated diagnostic equipment capable of testing the new aircraft systems. Prior to the time of the evaluation, students were trained as either test station operators or as test station maintenance personnel. When the evaluation began, however, these career options were integrated into a single career path, and a new STS was developed to reflect needed course modifications. In this combined course, theory and practical training on the test station was reduced from 14 days to a total of eight days.

Subsequently, still another plan of instruction was developed in accordance with an STS which was expected to become effective during the evaluation period. A comparative analysis of training objectives over the two years prior to the evaluation revealed that while most objectives remained the same, some objectives (e.g., training on specific LRUs) which were used to specify the simulator capabilities were eliminated due to the significant reduction in training time available. Since these training exercises were no longer employed, it was not appropriate to include the associated simulator capabilities in any tests of student performance.

While some course objectives did change over time, the relevant STS requirements did not substantially change. This finding led to the observation that the specialty standards are general enough to allow for interpretation, depending on one's perspective. Training and field personnel could easily assume that somewhat different skills are associated with specific requirements, such as "trouble-shooting." This lack of specific criteria of adequate performance made it difficult to measure training effectiveness even when objectives were identified. This difficulty was circumvented by considering only comparative training effectiveness and ignoring the more basic consideration of training adequacy. Thus, the study addressed only the question: "How do simulator-trained students perform as compared to actual equipment-trained students?"

Format of Training Program

The manner in which course material is delivered must be considered in the design of simulators. However, the format of the training course often changes to accommodate anticipated time, equipment and personnel constraints.

The comparative analysis of course content changes indicated that most of the objectives of the former maintenance and operations courses had been retained, although the training time allocated to each was greatly reduced. In fact, at the beginning of the performance data collection phase, only two days of practical training on the actual test station were included in the avionics maintenance course. This very limited contact with the equipment made it unlikely that performance differences as a function of training devices would be observed. Due to the short training period, it became necessary to control (experimentally and statistically) numerous variables which might obscure actual performance differences. Much effort was expended in controlling for individual differences, minor training deviations, instructor and supervisor differences, etc., in an attempt to reduce confounding of the performance measures by changing contextual factors.

Originally it was planned to rely heavily on existing test instruments to collect relevant data. Further, it was planned to focus on performance on the maintenance simulator only since previous test

scores from students trained on the actual test station equipment would be available to serve as baseline data. After observing classroom proceedings, it was apparent that changes in the training format necessitated changes in the performance test instruments used. The reorganization of the content of instruction blocks, together with numerous training "deviations" applied to nearly every class, made it inappropriate to use test scores from previous classes as baseline data. Performance measures were not based on consistent training experiences.

Other changes in format made it logistically impossible to equate training experience even during the evaluation period. In an effort to reduce the overall length of the course, it was decided to merge practical training on another test station into the same block of instruction for which the simulator was designed. To maintain the same amount of student-equipment contact in one-half the time, classes were divided into two groups. One group trained on each test station for two days. By reversing the groups for the remaining two days of practical training, all students completed their practical training on both test stations in four days rather than in eight days. The use of smaller groups and less time was expected to result in the same level of training as formerly given. However, since all of the performance testing took place on the day after the practical block of instruction was complete, it was necessary to control for the sequence of training received by each student. This procedure was essential to determine if the intervening training between the training and testing of interest had a detrimental effect on observed performance.

An additional change in the format of the maintenance training course involved the sequence of instructional blocks within the overall course. Training on the test station of interest was positioned earlier in the sequence of instruction blocks. This modification in the course resulted in less experienced students being trained and tested during the evaluation. The data collection plan was modified to carefully note any deviations in the sequence of training blocks. In theory, it was therefore possible to examine the performance of students in any instruction block in view of specific prior training experiences.

Reliability of the Training Equipment

A fundamental issue involved in the evaluation of simulated training devices is training equipment reliability. Clearly, equipment malfunctions have potential impact on the assessment of both performance and cost. Generally, it was expected that a review of equipment repair records together with direct classroom observation would be sufficient to predict the general nature of equipment malfunctions on training protocols. It was found, however, that recorded malfunctions did not correlate with training received and that the availability of the simulator provided a new alternative for dealing with faulty training equipment.

206

While malfunctions of both the actual test station and simulator were observed, no actual training time was lost due to the availability of the two training devices. The frequency of equipment failures, however, was sufficient to cause disruption of routine training and data collection. The major impact of malfunctions on the evaluation design was on the random assignment of trainees to experimental groups. Flexibility in evaluation design was essential to allow for unexpected changes in the training schedule and to minimize the potential loss of performance data.

It should be noted that impacts on training and performance due to equipment failures in prior training blocks were anticipated, but uncontrolled in this study. For example, in some instances, the first experience students had with an operational test station c during the evaluation period.

In large part, equipment reliability defined the format of the training approach used on the actual test station equipment. Unlike the approach with preprogrammed malfunctions on the simulator, the acquisition of practical experience was dependent on pre-existing or unexpected equipment failures. This was particularly true of the trouble-shooting aspects of training. In its extreme forms (which were observed during the experimental period), this dependency resulted in very limited training. At those times when all Test Replaceable Units (TRUs) and Line Replaceable Units (LRUs) operated without malfunction, it was not possible to demonstrate trouble-shooting techniques. At other times when a specific TRU failure caused the test station to be inoperative, no training was possible.

Finally, equipment malfunctions during training sometimes resulted in different training sequences for various classes. To maintain student flow, Air Training Command policy allowed instructors to move to another phase of training and to return to the block of instruction that was missed after the equipment became operational. In some instances, when specific AETs remained inoperable for long periods of time, a class did not receive any training in a particular block. A "training deviation" was filed for the class and graduation occurred as scheduled. With respect to the evaluation effort, these random variations in the training sequence made it impossible to assess the exact training experience of any given student at a specific point in time.

Assignment and OJT in the Field

From preliminary discussions with the training staff and field personnel, it became apparent that field assignments were made in view of the momentary demand for specific automatic test station operators. Thus, only a small number of airmen trained were expected to be assigned to the test station which had been simulated. Although the nature of the field assignments remained a design variable, it was clear from the outset that extremely unequal sample sizes for students assigned to the simulated and other test stations would be obtained.

Once trainees were assigned to the field, they entered the training phase called on-the-job training (OJT). Field training was relatively informal and was designed to develop the skills needed to operate and maintain the assigned test station.

Field site visits conducted in the course of the assessment revealed that the number of technicians available and their technical competency upon arrival determined the extent and type of OJT received. Due to the individualized approach to providing OJT in the field, it was not possible to use the extent and type of OJT required as an indicator of training effectiveness. An alternative method of assessing field performance considered was to record the number and cost of replacement parts requested and used by new technicians to perform test station and LRU repairs. Utilizing this indirect measure of training effectiveness was not feasible due to the complexities of obtaining such data in the field. In short, it was difficult to isolate clear measures of long-range impact of simulator training.

It was anticipated that the performance of simulator-trained personnel and actual equipment-trained personnel would be compared at specified intervals of time in the field. The effects of on-the-job training were expected to be present and constant in both groups and, therefore, any differences in performance could be attributed to the model of training. Clearly, with variable amounts of OJT, this assumption was not valid. In fact, if the originally proposed time series sampling framework was used, differences in performance due to training would be reduced as time in the field increased. Therefore, in order to obtain some measure of the long-range impact of simulator training on performance, it was necessary to devise a method of estimating job proficiency prior to field assignment, and to collect subjective ratings of field performance from supervisors shortly after the field placement (within about two weeks).

Conclusions

Again, it is important to point out that there is no reason to suspect that the simulated test station and the training environment studied are atypical in any significant way. These difficulties associated with designing and implementing a training effectiveness study of a maintenance training simulator suggest a number of critical factors which should be considered at the outset of efforts to investigate the role of simulators in maintenance training. Certainly, any researcher will consider these issues in planning a training effectiveness study. It appears that mere consideration of the factors does not ensure that adequate measurement of simulator training effectiveness will ensue. In fact, unless the interaction between the simulator and the maintenance training environment can be directly observed, it is unlikely that a strong research design can be implemented. It is for this reason that a two phase assessment plan should be considered a necessity. While Phase II can focus on measuring training effectiveness, perhaps outside the actual training environment, it is Phase I which serves

a planning and refinement function that will enable the researcher to adequately determine the issues relevant to the effective design and use of the maintenance simulator.

The following discussion presents some issues which should be considered in the design and use of maintenance simulators. This discussion illustrates the potential value of an exploratory Phase I. Additionally, alternative methodologies for implementing Phase II of a training effectiveness assessment are suggested.

Identification of Training Objectives

Any comparative analysis of simulators and AET must be based on a clear understanding of training objectives. The present study was conducted at a time when classroom objectives were rapidly changing to meet the changing requirements of the field assignment. Given the lead time necessary to develop the specifications of a simulator and to construct the trainer, it should be expected that the capabilities of the simulator were somewhat limited with respect to current training needs. Since it is unlikely and perhaps undesirable that the objectives of a training course can be stabilized for long periods of time, the useful lifespan of major simulators may be significantly increased by emphasizing training related to the general skills required to operate and maintain all test stations. The specialized functions of each test station could be simulated with less costly module simulators, which are either disposable or reprogrammable to meet changing needs. Such an approach is not unlike the "Test Station Replaceable Unit" (TRU) design of equipment already in use. While the main purpose of the TRUs is to reduce "downtime," the component system by making repairs easier, makes if easier to update (within limits) test stations to provide testing capability for new or modified aircraft systems. There is no reason to expect that a simulator designed to replace a test station will not be subject to the same limitations on utility due to improvement in aircraft design.

In view of changing course objectives and needs in the field, it seems that an explicit statement of minimal trouble-shooting standards should be made. These standards can, then, form the basis of a variety of training strategies and trainers.

Define the Anticipated Role of Simulation

From the outset of any investigation, the intent of introducing simulators into training should be clearly stated. That is, it should be determined if the simulator designed is expected to replace or supplement actual training equipment. Both are valid approaches to the use of simulation, of course, but require somewhat different equipment and research designs.

If the primary objective of incorporating a simulator into training is to replace more costly, less reliable, and more dangerous actual equipment, then it is clear that both psychological and physical fidelity must be considered. In this case, it is more likely that the simulator will include the capability of demonstrating basic operations procedures while duplicating, to a large extent, the physical appearance of the actual equipment. Physical fidelity is important since trainees will be assigned to actual equipment in the field. A lack of sufficient physical fidelity in the simulator and no exposure to actual equipment in training would almost certainly result in reduced initial performance on actual equipment in the field. However, further research may determine which tasks require full fidelity trainers.

If, on the other hand, the objective of incorporating a simulator into training is to supplement the use of actual test station equipment, then psychological fidelity should be emphasized. In either case, psychological fidelity of sensitivity to operator actions must be established. Interviews with students who had contact with the maintenance simulator, conducted directly after the practical instruction block, and responses from technicians on the field follow-up questionnaires, indicated disappointment with the simulator on this aspect of design. That is, while the two trainers looked similar, the simulator reacted differently (slower) to operator input. This flaw was also referred to by Becar (1978) in his evaluation of the maintenance trainer system. Supplemental simulators can more easily focus on more complex tasks, including training on equipment malfunctions which cannot be introduced or experienced on actual test station equipment.

The use of simulators for training maintenance skills offers an opportunity to provide consistent training since they are less subject to random malfunctions. Further, simulators designed to augment actual equipment trainers can be more easily used to train personnel in the operation and maintenance of test stations in general. More general skills (e.g., systematic problem solving) not unique to any specific test station can be provided as an introduction to maintenance training on specific test stations. In the present study it was found that students trained on the maintenance simulator performed as well as students trained on actual equipment. It is possible that the training benefits of the simulator were not realized either because the simulator was designed to replace the AET or because the assessment was insensitive to performance differences. Improved student performance was expected, however, because the simulator provided more consistent training experiences. In general, it seems that if the simulator is designed to replace the AET, the outcome of a cost comparison between trainers becomes the major factor in future procurement decisions, given approximately equivalent training capability. Given a "supplemental" objective, improved performance becomes the major factor considered.

Approach to Teaching

It is unlikely that a simulator of any quality will be accepted into existing training curriculum if it is not somewhat consistent with existing student-teacher interaction patterns. It seems that only a self-instruction mode in which the simulator guides the student through a series of problems might result in alteration of these patterns. To encourage teacher acceptance, the simulator should be effective as both a visual aid and demonstration tool. This allows the simulator to be effectively incorporated into training segments (e.g., theory familiarization) which do not include extensive practical trouble-shooting experience. Such a dual purpose simulator would be essential if replacement of existing equipment is planned. Also, the thorough introduction of the training simulator to the training staff, highlighting uses, real and potential, of the equipment in the overall training program, will improve instructor acceptance.

The environmental constraints on this study suggest that many instructional practices have evolved which are deemed necessary to maximize the effectiveness of actual equipment as trainers. Given this situation, it is not surprising that performance differences as a function of training equipment were not observed. The potential impact of simulator training on student performance may be realized only if a utilization strategy accompanies the placement of a simulator into an existing training environment. This plan for using the simulator would, in the likelihood that its unique training capabilities are tapped, and that benefits in terms of improved performance, consistent training, reduced training time, and cost savings are measured.

Generalization of Findings

Given the preceeding discussion of the impact of the environment on the evaluation effort, the generalizability of findings from this study are limited. While every effort was made to adapt experimental design principles for use in this natural experiment, it was not possible to rely on many of the premises of basic learning theory. Until parameters such as content, method, and duration of training, all known to affect learning, are subject to more careful control, an adequate analysis of simulation training effectiveness will not be possible. The point at which simulators provide the best training for a specified cost or the least cost for specific training can be determined only if control over relevant learning factors is possible. To answer the question, "Do simulators provide more cost-effective training than AET?" it must be able to maximize the use of simulator capabilities beyond those available on actual test stations. Simply stated, AETs are not designed for training purposes; simulators can be designed solely for that purpose.

Experimentation in the natural training environment is necessarily limited by the fact that simulators must provide training at lease equivalent

to that provided by existing actual equipment trainers (AETs). The adequacy of training on AETs in relation to the STS is usually assumed; however, the issue could also be investigated empirically. It is not feasible to risk the possibility of providing inferior training in the interest of research which is focused on defining the conditions of maximum cost-effective simulator training.

There are two viable solutions to avoiding the limitations imposed by the natural environment. First, the student could be deviated from normal training to participate in a well controlled research study and then subsequently re-enter the training sequence at the point of departure. The additional cost of deviating students (i.e., cost of additional day in ATC) would be part of the research costs. The disruption of student flow from the field's perspective would occur only at the start of such a project and should not cause any significant shortages of field personnel since student flow is normally somewhat erratic. The second alternative is to carefully structure a significant block of training time to allow the research project to be integrated into the existing training sequence. Students would complete training in the same time frame as usual (or sooner), and any adverse impacts of the research on performance could be corrected by OJT. The additional training required (if any) would become itself a measure of training effectiveness. While other alternatives may be possible, the main point is that true potential of simulation in training can be determined only by a focused research effort.

Clearly, some additional costs will be incurred by such research efforts--a small price, however, given the potential utility of the information in defining the future role of simulation in maintenance training. Such research would ensure that future investments in simulators would be based on factual information rather than assumptions. The information obtained should highlight the conditions under which the use of simulators in maintenance training is most effective. The overall cost savings would be extensive regardless of the findings.

REFERENCES

Becar, N.J. Software systems review of basic 6883 maintenance training system. Colorado Springs, CO: Kaman Sciences Corporation, 1978.

Caro, P.W. Some factors influencing Air Force simulator training effectiveness. AD-A043 239. Pensacola, FL: Seville Research Corporation, 1977a.

Caro, P.W. Some current problems in simulator design, testing and use. AD-A043 240. Pensacola, FL: Seville Research Corporation, 1977b.

Cicchinelli, L.F., Harmon, K.R., Keller, R.A. & Kottenstette, J.P. Relative cost and training effectiveness of the 6883 three-dimensional simulator and actual equipment. AFHRL-TR-80-24. Lowry AFB, CO: Logistics and Technical Training Division, Logistics Research Branch, in press.

Daniels, R.W., Datta, J.R., Gardner, J.A. & Modrick, J.A. Feasibility of automatic electronic maintenance training (AEMT). Volume I--Design development and evaluation of AEMT-ALQ-100 demonstration facility. Warminster, PA: Naval Air Development Center, May 1975.

Hurlock, P.E. & Slough, D.A. Experimental evaluation of Plato IV technology: Final report. San Diego, CA: Navy Personnel Research and Development Center, August 1976.

Orlansky, J. & String, J. Cost-effectiveness of flight simulators for military training. IDA Paper P. 1275. Institute for Defense Analyses, 1977.

Wright, J. & Campbell, J. Evaluation of the EC-11 programmable maintenance simulator in T-2C organizational maintenance training (NADC 75083-40). Warminster, PA: Naval Air Development Center, May 1975.

DAVIS, Brian C., SQT Management Directorate, US Army Training Support Center, Ft Eustis, Virginia, BRITTAIN, Dr. Clay, also of SMD, USATSC, and BERK, Dr. Ronald, The Johns Hopkins University.

SETTING SQT CUTSCORES - WAYS & MEANINGS (Wed P.M.)

US Army Skill Qualification Tests (SQT) measure how well soldiers can perform selected tasks in their jobs. Soldiers are scored competent (GO) or incompetent (NO-GO) on each tested task. A soldier's overall SQT score is simply the percentage of task on which he scored GO.

Cutscores for tasks are now set empirically. A cutscore for the overall SQT score is being developed to serve training and personnel management needs. The paper examines the presently identifed needs for an overall SQT cutscore, and relates these to several proposed methods of establishing that cutscore. Implications of each method for SQT content and administrative procedures are discussed.

SETTING SQT CUTSCORES - WAYS & MEANINGS

Perhaps no problem has so perplexed criterion referenced test (CRT) developers as the determination of appropriate competency-based cutscore(s). This may be because as Glass (197) contends, the issue of cutscores for CRT was born out of a semantic confusion between the original meaning of "criterion" as a "criterion variable" (i.e., "a behavioral scale articulated to a test") and the subsequent meaning of criterion as "standard", (i.e., a cutscore which divides competence from incompetence"). Referencing test scores to a behavioral variable is difficult enough, but divining which behavioral quantum (and associated test score) represents a qualitative transition from "unworthy to worthy", "incompetent to competent" poses metaphysical problems with which most psychometrists are ill equipped to deal. In fact, that divination is not a psychometric problem at all and consequently cannot be left to test developers. Setting competency based cutscores is actually a "policy capturing" exercise. Management sets policy. Psychometrists capture it. Metaphysical policy-making, however, is difficult even for managers and they are want to foist that responsibility onto the psychometrists. Nevertheless, then division of responsibility must be maintained: Management decides the policy on the distinction between competence and incompetence; then psychometrists operationalize this policy in setting CRT cutscores. In terms of the players, then, there are three general stages in setting a competency based cutscore:

1. MANAGEMENT: Defining competence with respect to the content domain

2. TEST DEVELOPER: Operationalizing the definition with respect to the test by setting a cutscore.

3. MANAGEMENT & TEST DEVELOPER IN CONCERT: Refining the cutscore to accommodate decision rule preferences.

STAGE 1" DEFINING COMPETENCE - THE MANAGER'S DILEMMA

Competence in CRT is defined with respect to the behavioral domain to be assessed. Therefore, a domain of "critical skills/behaviors/knowledge" must be identified. This is generally done through job analysis, mission analysis or content analysis depending on the subject matter. What results is a list of behavioral or learning objectives which represent the elements of the job, mission or training to be evaluated.

The manager's dilemma is then to divide the listed objectives into three types:

A. objectives which are essential for competence;

B. objectives which are relevant and desirable for competence but not individually essential;

C. objectives which are irrelevant to competence and which should therefore be purged from the domain to be tested.

The CRT will test those objectives which fall in type A or B.

Type A objectives

If all objectives are defined as "essential to competence", the manager has told you that examinees, to be called competent on the CRT, must be at least minimally competent on 100% of the objectives. This means that, if the portion of the CRT devoted to each objective were scored "pass" if the response was minimally competent or better, (NOTE: the minimally competent response may be less than perfectly correct) and "fail" otherwise, the nominal competency based cutscore for the whole CRT would be 100%. Having so operationalized competency as a 100% score on the CRT, stage 2 is accomplished. All that remains is stage 3, to refine the cutscore based on management's willingness to give the examinee some benefit of the doubt. (NOTE: "Doubt" is defined by some index of measurement error). Glass (1977) calls this, "counting backwards from 100".

Type B objectives

The psychometrist is more often faced with some or all objectives described as Type B. If objectives are descred as Type B, the manager has equivocated and so owes the psychometrist something more by way of definition. Fortunately at least two courses of action are open to the manager:

- Plead "competence is what competents do" and then either identify people who are representative of competency determined groups or identify judges (subject matter experts) who can estimate what competents would do on the test.

- Boldly assert that competence means mastering a specified proportion of the Type B objectives. (NOTE: It may be necessary to first weight the objectives in terms of contribution, to competence if all objectives are not equally important.) This weighting can be most directly accomplished by manipulating the relative proportion of test content devoted to each objective. More important objectives are accorded a larger part of the test content than less important ones.

If the former course of action is chosen, the psychometrist must operationalize the definition as described in stage 2. If the latter course of action is chosen the manager has operationalized competence (e.g., "at least minimally competent on 80% of the objectives) in the same way as it was operationalized at 100% for Type A objectives. This pre-empts the need for stage 2, leaving only cutscore refinement to be accomplished.

A mix of Type A and B objectives

If given a mix of Type A and Type B objectives, the most straightforward approach would be to create two tests - one made up of the Type A objectives with its own cutscore and one made up of the Type B objectives with its own

---

Footnote: Setting "an objective-level, competency-based passing score" for the portion (e.g., a question, or a sub test) of the CRT devoted to each objective, proceeds in the same way as for the whole CRT except that subject matter experts, instead of managers, define competence on the objective.

cutscore. Good examples of such tests are driver's licence exams, with a "basic signs" section (Type A objectives) requiring 100% mastery and a general operator knowledge/performance section (Type B objectives) requiring 80% mastery.

STAGE 2: OPERATIONALIZING THE COMPETENCY DEFINITION - PSYCHOMETRIST'S WORK

(NOTE: This stage is needed only for Type B objectives for which competence was defined as "what competents do"). A number of distinct cutscore-setting techniques have been proposed. These can be categorized in many different ways, but, for simplicity, only three dichotomous dimensions are needed to show the major differences in approach[2]. These dimensions are:

- Atomistic vs Holistic - Will the cutscore be determined directly from total score analysis (i.e., holistically) or indirectly from score analysis and accumulation of scores on each tested objective (i.e., atomistically)?

- Empirical vs Rational - Will the cutscore determination be based on actual test scores (i.e., empirically) or logical conjecture about how certain test takers would or should respond (i.e., rationally)?

- Contrasting Groups vs Anchor Group - Will the cutscore be set so as to optinally differentiate competents from incompetents (i.e., contrasting groups approach) or so as to pinpoint performance of the borderline competent (i.e., Anchor group approach)?

These three dichotomous dimensions yield eight different procedures for setting competency-based cutscores. These procedures and a sample cutscore algorithum for each are shown in figure 1. (NOTE: A variety of specific techniques may fit into a single procedure cell as for example the ATOMISTIC-RATIONAL-ANCHOR GROUP cell. This cell include the Ebil technique. the Angoff technique and the Nedelsky technique).

Comparative studies have shown that different techniques both within (Andrew and Hecht (1976), Hambleton (1978), Shakun and Kling (1980)) can give very different cutscores and results for the same test. Unfortunately, there is as yet no evidence to establish which procedure gives truest results, in part because "truth" with respect to competency is itself debatable. However, there are reasons to suspect the accuracy of some methods more than others.

RATIONAL VS EMPIRICAL: All methods require one of two types of judgments by managers or subject matter experts:

- judgment about what competents would or should be able to do on the test (RATIONAL METHODS), or

- judgement about who is clearly competent/incompetent or marginally competent (EMPIRICAL METHODS).

Footnote 2: The procedures which are discussed all require the assumption that the CRT scores are monotonically related to the criterion variable, or behavioral scale, as is true for most CRT. Where monotonicity does not obtain, multiple cutscores defining regions or profiles would be needed--but that is another paper.

FIGURE 1

METHODS OF OPERATIONALIZING COMPETENCE IN CRT CUTSCORES

	ATOMISTIC		HOLISTIC	
	EMPIRICAL	RATIONAL	EMPIRICAL	RATIONAL
CONTRAST GROUPS	$C = \frac{\sum_j^E (mc_j + mi_j)(w_j)}{2}$	$C = \frac{\sum_j^E (pc_j + pi_j)(w_j)}{2}$	$C = \frac{MC + MI}{2}$	$C = \frac{PC + PI}{2}$
ANCHOR (BORDERLINE) GROUPS	$C = \sum_j^E (ma_j)(w_j)$	$C = \sum_j^E (pa_j)(w_j)$	$C = MA$	$C = PA$

DA-1-4

= STAGE 1 PROCEDURES

= STAGE 2 PROCEDURES

C = CUTSCORE

F = NUMBER OF TEST ELEMENTS

W_j = "IMPORTANCE" WEIGHT OF j ELEMENT USUALLY A CONSTANT 1.0.

mc_j, mi_j, ma_j = MEASURE OF CENTRAL TENDENCY ON j ELEMENT SCORE FOR COMPETENT, INCOMPETENT AND ANCHOR GROUPS RESPECTIVELY.

MC, MI, MA = MEASURE OF CENTRAL TENDENCY ON TOTAL TEST SCORE FOR COMPETENT, INCOMPETENT AND ANCHOR GROUPS RESPECTIVELY.

pc_j, pi_j, pa_j = JUDGED PROBABILITY OF A PASS (OR EXPECTED SCORE) ON j ELEMENT FOR COMPETENTS, INCOMPETENTS & BORDERLINE COMPETENTS RESPECTIVELY.

PC , PI , PA = JUDGED PROPORTION OF ELEMENTS PASSED BY COMPETENTS, INCOMPETENTS, AND ANCHOR GROUPS RESPECTIVELY.

However, there is good reason to believe that the latter (EMPIRICAL METHOD) judgements are more accurate. First, people are, in general, more accustomed to making judgements about the competency of others, whereas they are unfamiliar with the task of "imagining" how hypothetically competent examinees would specifically perform on tests. Secondly, judgements about real people are less abstract than similar judgement about hypothetical examinees. Thirdly, inter-judge reliability is know to be poor for the RATIONAL METHOD judgements.

Consequently, EMPIRICAL METHODS would seem to be preferable to RATIONAL METHODS, or at least no worse.

CONTRASTING GROUPS VS ANCHOR GROUPS -

Given a preference for EMPIRICAL METHODS, the question becomes: can judges more accurately identify members of contrasting groups (i.e., clearly competent or clearly incompetent) or members of anchor groups (i.e., borderline competents)? Again it seems that one of these determinations would be more accurate than the other. CONTRASTING GROUPS are composed, by definition, of people whom judges can clearly identify as competent or incompetent. Borderline ANCHOR GROUPS are composed, by definition, of people whom judges cannot clearly categorize as competent or incompetent. This judgemental indecision results from some combination of two uncertainties: uncertainty about where, precisely, the line between competence and incompetence lies, and/or uncertainty about the capabilities of the people being judges. This former uncertainty, if not too great, is tolerable since the precise value is to be estimated from some central tendency statistic on the anchor group's scores. The latter uncertainty is very problematic since it subverts the meaning of the anchor group. That is, the anchor group would represent unknown performers rather than borderline performers. (NOTE: If ANCHOR GROUP METHODS are to be used, special care must be exercised to insure anchor group membership is not based on uncertainty about the capabilities of the people being judged.) Consequently, a concern for judgement accuracy leads to a preference for EMPIRICAL-CONTRASTING GROUPS METHODS.

ATOMISTIC VS HOLISTIC METHODS -

Since patterns of performance across test elements are not important, it makes little sense to go through the rigamarole of empirically determining competent vs incompetent performance on each test element in order to arrive at an overall cutscore, when a direct determination, based on overall score is much easier.

Furthermore, since empirical estimates of minimally competent performance on each test element are subject to sampling error, then consolidation of these estimates into an overall cutscore results in a concomitant accumulation of the element-wise errors of estimate for the overall cutscore. It is likely that estimation of an overall cutscore directly from overall scores would involve less estimate error.

Consequently, HOLISTIC METHODS are preferable where feasible and they are feasible under empirical approaches.

Following the above line of reasoning, then, EMPIRICAL - CONTRASTING GROUPS - HOLISTIC METHODS seem to be the most preferable approach. A pilot study using this approach to setting cutscores for selected US Army Skill Qualification Tests (SQT) will be conducted in 1981.

STAGE 3: REFINING THE CUTSCORE - CONCERTED WORK

At the end of the second stage, a cutscore has been set which satisfactorily operationalizes management's definition of competence. However, cutscores are ordinarily used to support decisions. For example, people are hired/not hired, promoted/demoted, retained/terminated/remediated depending on whether they scored above or below the cutscore. Consequently, decision consequences affect the desirability of a particular cutscore.

Desirability is defined in terms of misclassification costs. Since most CRT are not perfectly valid and since most cutscores are not perfect operationalizations of the competency definition, some examinees will probably be misclassified as competents or incompetents. Truly competent examinees who may erroneously score below the cutscore would be called "false incompetents." Truly incompetent examinees who may erroneously score above the cutscore would be called "false competents." Such misclassifications may incure serious costs in a number of areas:

- Actual fiscal losses resulting from poor performance of "false competents" who were hired/promoted/retained.

- Loss of potential gains from good performance which would have resulted if false incompetents had been hired/promoted/retained.

- Morale degradation associated with apparently erroneous personnel actions.

- Hazards incurred from using false competents to perform dangerous jobs.

- Costs of wasted remedial training for false incompetents.

Taking all such misclassification losses into account the manager must decide, or balance, which type of misclassification (i.e., false competent or false incompetent) is most tolerable. This decision will determine the direction of the cutscore refinement.

- If false competents are more tolerable, the cutscore will be lowered, allowing for relatively more false competents and relatively fewer false incompetents.

- If false incompetents are more tolerable, the cutscore will be raised, allowing for relatively more false incompetents and relatively fewer false competents.

- If false competents and false incompetents are equally tolerable, then refinement of the cutscore which equalizes numbers of false competents and false incompetents is needed[3].

Footnote 3: The cutscore set during stage 2 has probably accomplished this. Where stage 2 was precluded, and competency is only defined in terms of the CRT, then the cutscore can be moved to the value which equalizes decision consistency for competents and incompetenst across two or more test administrations.

The magnitude of the cutscore refinement depends on:

- The manager's regret of false competents relative to false incompetents (NOTE: Often called the regret ratio, and expressed as, for example, "I can tolerate three times as many false competents as false incompetent(s)".

The cutscore is then refined to finally acceptable value either:

- by trying out various values until the one which most closely captures the desired regret ratio is found, or

- (appealing to classical true score theory) by determining the value, based on standard error of measurement, for which the true score may equal or exceed the initially determined (stage 2) cutscore with a desired probability (i.e., this probability would be based on the regret ratio)[4].

SQT CUTSCORE - AN EXAMPLE (TO BE PILOTED)

The Skill Qualification Test (SQT) is a CRT which tests a selected set of performance objectives, called tasks, from the examinee's job. There are nearly 1000 'jobs' in the US Army and, for most, the critical tasks which comprise the job have been listed. Generally there will be an SQT for each job. The selection of tasks tested in the SQT is drawn from, but not necessarily representative of, the domain of tasks which make up each job. Soldiers are advised of the tasks in their job to be tested, and encouraged to "train for the test".

Competency-based cutscores are determined for each tested task during validation. Soldiers are scored 'GO', meaning task competent, or 'NO-GO', meaning task incompetent, for each task tested in their SQT. The overall SQT score a soldier receives is simply the proportion of tested tasks on which he scored 'GO'.

The cutscore problem for SQT is to determine the overall SQT score which differentiates competent job incumbents from incompetent job incumbents.

The fact that the SQT does not necessarily represent the job domain and that soldiers train for the test precludes inference from SQT results to some domain score. In any case, the tasks which comprise a soldier job have generally been treated as Type B (i.e., desirable but not essential to competence) objectives. In the past, program managers, at stage 1, asserted that 60% 'GO' would define job competence. However, a fairly wide spread in pass rates from job to job has undermined satisfaction with 60. The next logical approach for Type B objectives then is to define competence as "what competents do". Since SQT scores are not generalizable to the domain of tasks for the job, job competence defined on the domain cannot be rationally related to an SQT score. Consequently a statistical (i.e., empirical) linkage

---

Footnote 4: For example, if:

regret ratio $= \left(\frac{20 \text{ false competents}}{1 \text{ false incompetent}}\right)$; the SEM = 10; and original cutscore = 100, then 84 = 100 - 1.64(10) would be the refined cutscore.

is sought. For reasons discussed earlier, a HOLISTIC-EMPIRICAL-CONTRASTING GROUPS METHOD will be used to effect this linkage. Supervisors will be asked to identify soldiers from among their subordinates who they judge to be clearly competent or clearly incompetent with respect to the job domain of tasks. These soldiers will then be tested on the SQT during validation. When sufficient and equal samples of judged competent and incompetent soldiers are obtained, their SQT scores will be analyzed to determine the SQT cutscore which would best differentiates competents from incompetents. This stage 2 process can be accomplished by way of a contingency table analysis using competent/incompetent group membership as one factor and above/below SQT cutscore as the other factor. Then different SQT score values would be tried out to determine which optimizes some index of relationship (e.g., Cohen's Kappa; Phi; % of correct classifications or a discrimination index) between the two factors. Such a procedure will likely result in a different cutscore for each SQT.

Cutscore refinement will also be SQT peculiar in deference to the very different regret ratios likely for the very different jobs. For example false competents are a much more serious problem in nuclear weapons specialties than in bandsmen jobs, and this fact will be reflected in different regret ratios.

Since false competent and false incompetent rates will be known for various SQT scores from stage 2 analysis, a refined cutscore which provides misclassification errors approximating the regret ratio may be directly determined.

Scoring above/below such an SQT cutscore would then mean that a soldier performed <u>like</u> job competents/incompetents on the set of tasks tested in the SQT. Attaching such meaning to the cutscore is the sign of a job well done.

BIBLIOGRAPHY

1. Andrew, B. J. and Hecht, J. T. (1976) "A Preliminary Investigation of Two Procedures for Setting Examination Standards," Educational and Psychological Measurement, 36, pp. 45-50.

2. Berk, Ronald A. (1976) "Determination of Optimal Cutting Scores in Criterion & Referenced Measurement" Journal of Experimental Education, 45, pp. 4-9.

3. Hambleton, R. K. (1978) "On the Use of Cut-Off Scores with Criterion-Referenced Tests in Instructional Settings," Journal of Educational Measurement, 15, pp. 277-290.

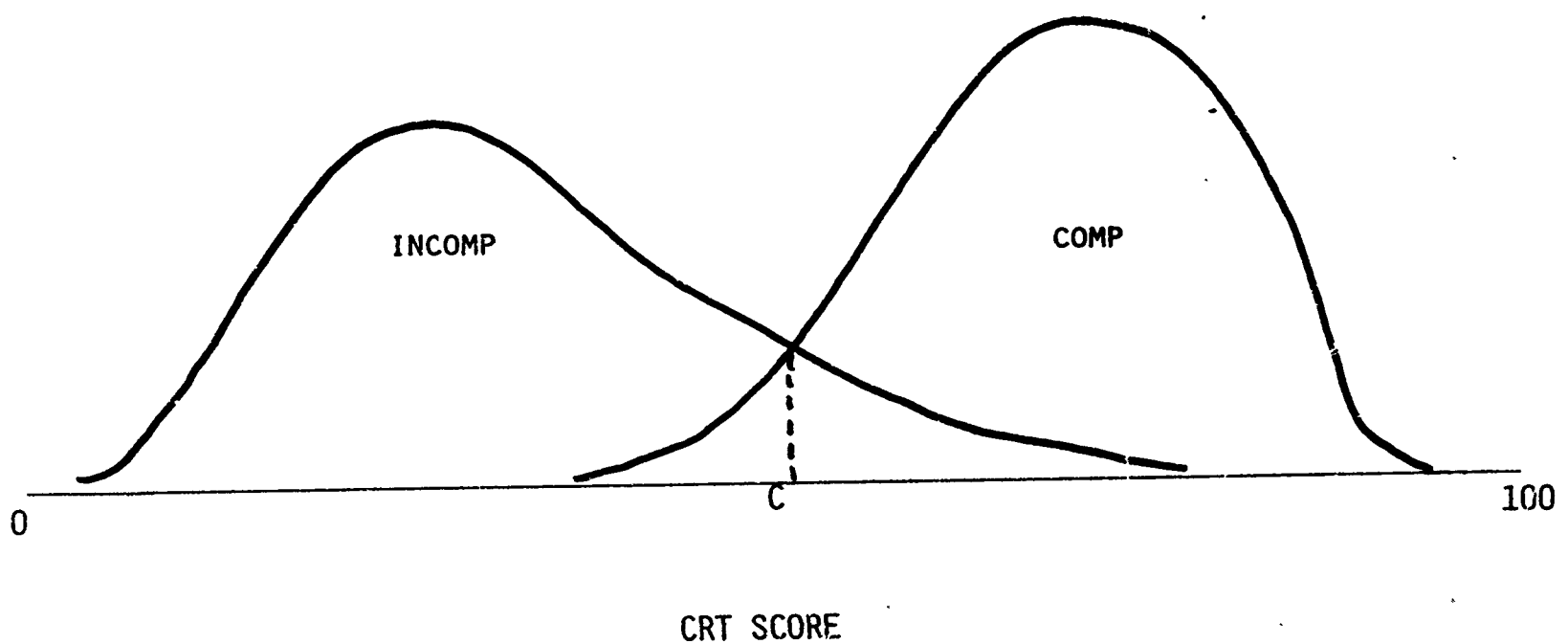
4. Jaeger, Richard M. (1979) "Measurement Consequences of Selected Standard Setting Models" in Practices & Problems in Competency-Based Measurement, Bunda, M. A., and Saunders, J. R., (eds)., National Council on Measurement in Education, pp. 48-58.

5. Koffler, Stephen L. (1980) "A Comparison of Approaches for Setting Proficiency Standards" Journal of Educational Measurement, 17, pp. 167-178.

6. Skakun, E. N. and Kling, S. (1980) "Comparability of Methods for Setting Standards" Journal of Educational Measurement, 17, pp. 229-235.

7. Zieky, M. L. & Livingston, S. A. (1977) A Manual for Setting Standards on the Basic Skills Assessment Tests. Princeton, New Jersey: Educational Testing Service.

HOLISTIC - EMPIRICAL - CONTRASTING GROUPS METHOD



SQT (C = CUTSCORE)

IDENTIFIED GROUPS		$< C$	$\geq C$	
	INCOMPS	A	B	A + B
	COMPS	C	D	C + D

INDICES OF RELATIONSHIP: ϕ

$$\phi_K = \frac{D}{C+D} - \frac{B}{A+B}$$

$$C = \frac{A+D}{A+B+C+D}$$

Determining Task Commonality in Navy Training

Two Methods Examined

by

Douglass Davis

and

Nancy N. Perry, Ed.D.

Staff, Chief of Naval Education and Training
Pensacola, Florida

ABSTRACT

Instructional technology in naval training is more than the sum of its parts. It is a systematic way of designing, carrying out, and evaluating the total process of learning and teaching in terms of specific objectives. Inherent in this process is task analysis, which results in the development of task inventories. From these task inventories, tasks are selected for training. There are presently two methods of comparing tasks for commonality in Navy training: (1) subjective judgment of Navy technicians, or subject matter experts, and (2) use of a computer in an experimental model which employs subject-matter-expert input in the recording of job data and the computer in storing and retrieving the data. This paper explores the relationships between subject-matter-expert task commonality rankings of given sets of tasks (from most of least common) and computer rankings of the identical (same) tasks within the framework of an experimental model.

Instructional technology in the United States Navy is known as "Instructional Systems Development," and is often referred to as "ISD." Instructional Systems Development is defined as

> a systematic way of designing, carrying out, and evaluating the total process of learning and teaching in terms of specific objectives, based upon research in human learning and communication and employing a combination of human and nonhuman resources to bring about more effective instruction (Commission on Instructional Technology, 1970, p. 19).

Scanland (1977) compared the ISD phases to the functions of management. Figure 1 (p. 7) below lists the functions of management down the left side. The next column, under the caption "ISD PHASES" lists the major components of nearly all ISD models. The first item under the heading "ISD STEPS," task analysis, is the most essential component of ISD. Task analysis produces task inventories, or lists of actual job tasks, which trainees will be able to perform upon completion of Navy training programs, after they have been assigned to a billet in the Navy work force.

MANAGEMENT	ISD PHASES	ISD STEPS
PLANNING	ANALYSIS & DESIGN	TASK ANALYSIS SELECT TASKS ALTERNATIVE STRATEGIES OBJECTIVES CRITERIA
ORGANIZING	DEVELOPMENT	MATERIALS METHODS EQUIPMENTS FACILITIES PERSONNEL
CONTROLLING	IMPLEMENTATION	VALIDATE IMPLEMENT PLAN CONDUCT
EVALUATION	EVALUATION	PROCESS PRODUCT FEEDBACK CORRECT

Figure 1. Management and ISD functions compared.

230

This paper addresses only one dimension of task analysis--that of determining commonality among related tasks. For purposes of establishing a common understanding, we may say that two tasks with identical performance requirements (skills required, tools required, conditions, actions, etc.) would be 100 percent common, or identical. Tasks having fewer identical requirements would be common to a lesser degree. At this point, it would be worthwhile to begin with a historical perspective of task analysis and the attention given thus far to the determination of task commonality.

Job/task analysis as a basis of Navy training program development emerged in the late 1960s (Bureau of Naval Personnel, 1968) and was institutionalized by Rundquist (1970) who employed a "systematic procedure" of analyzing jobs by subdividing them into component "duties" and further subdividing duties into component "tasks." This analysis procedure actually originated during the last decade of the nineteenth century, as a component of Frederick W. Taylor's "scientific management" (1947, p. 271). Taylor stated that tasks could be simplified after their content was known by subdividing them into elements which were less comprehensive (Butterworth, 1973), and emphasized, "Perhaps the most prominent single element in modern scientific management is the task idea" (1911, p. 39). It is generally acknowledged that the most prominent single element in ISD is still the task.

Tracey, Flynn, and Legere incorporated commonality into Army ISD (1970) through "universality" or general objectives having wide application. The rationale was that a skill or knowledge required for use in many job situations was more likely to be a candidate for training than one which occurred in only a few job situations.

Rundquist attempted to incorporate commonality into design of training by grouping similar tasks that could be consolidated for training purposes. This attempt involved "sorting Job Task Cards until satisfied that job tasks can be trained together have been brought together" (1970, p. 43).

In 1973, the Individualized Learning Development Group at the Naval Training Center, San Diego, California, developed a Job Task Analysis Manual which placed task commonality considerations at the "front end"--in the selection of tasks for training--of ISD. Tasks were grouped by universality codes during job/task analysis. For example, CODE 1 tasks were tasks that appeared to be common across the board. No significant deviation existed between afloat and ashore activities. CODE 2 tasks appeared to be performed primarily afloat, with no significant deviation from one afloat activity to the next. Other codes designated tasks that appeared to be performed primarily afloat at a single activity, etc.

Even though each of the attempts to deal with commonality differed in its systematic approach, researchers were successful in communicating the essential concept. Navy policy makers had begun to entertain commonality considerations by 1974, at which time the Navy restudied its work structure, personnel utilization, and training, and developed a "new" Navy Enlisted Occupational Classification System (NEOCS) which called for a reduction in duplicative work classifications and instructional efforts through the maximum utilization of schools common to numerous ratings and occupational fields (Chief of Naval Personnel, 1974).

In response to the "new" NEOCS, the Chief of Naval Education and Training enunciated the need to "Design and develop an ADP (automated data processing) system to identify common tasks/learning objectives across ratings/occupational fields" (1974, enclosure 2). The plan for such a system was published as the Naval Enlisted Professional Development Information System Automated Data System Plan (Chief of Naval Education and Training, 1977). In support of this plan, an experimental model task inventory file was developed by the Career Development Group, Pensacola, Florida, under the sponsorship of the Chief of Naval Education and Training. An essential contribution of the experimental model has been use of the computer in comparing tasks for commonality through storage and retrieval of numerous data regarding individual task performance requirements, including cues, standards, reference materials, conditions of performance, etc., and a set of descriptive characteristics for each task. The descriptive characteristics result from subject-matter-expert use of official operation and repair manuals approved for use in job situations (Ansbro, 1978).

The experimental model is the nucleus around which this paper has been prepared. Aside from this model which employs both subject-matter-experts and a computer, Fink (1978) reported one other procedure for identifying identical and similar tasks in the Navy ISD process. That procedure is based strictly upon the subjective judgment of Navy technicians. This paper addresses an overview of the mechanics of the experimental model and some of the preliminary results of comparisons of the two procedures, in terms of results.

The experimental model task inventory file, which is a part of NEPDIS (Naval Enlisted Professional Development Information System), is designed to be a computerized system for storing, processing, managing, and retrieving information from job and task data collected from official Navy personnel and technical data sources such as those listed in Table 1. When these data have been converted from job requirements into training requirements, they can be used for designing a training system that can deliver effective, efficient, and timely training programs.

Table 1

Examples of Sources of Official Navy Personnel and Technical Data

Technical Data:

- Training Manuals
- Specifications
- Instructions
- COSAL (Coordinated Shipboard Allowance List)
- 3-M (Maintenance and Material Management)

Personnel Data:

- PQS (Personnel Qualification Standards)
- EOSS (Engineering Operational Sequencing System)
- PARS (Personnel Advancement Requirement System)
- NOTAP (Navy Occupational Task Analysis Program)
- SMD/SQMD (Ship Manning Document/Squadron Manning Document)

Determining commonality of tasks is only one small portion of the NEPDIS capability, but it is an important one, because comparing tasks for commonality and eliminating the training of redundant tasks makes training more efficient. It is the CNET goal to provide a proficient occupant for every billet at minimum cost (Cagle, 1973) that has led to the portion of the NEPDIS experimental model under discussion.

Before we can show how the computer is used in this model to determine commonality among tasks, we must first describe how the data are collected and entered into the task inventory file.

As in the other method for comparing tasks for commonality, NEPDIS employs subject-matter-experts (SMEs) for the initial recording of job data. The difference is in how the SMEs are employed. The other method uses the subjective judgment of the SME to assess the degree to which two tasks are common; in sum, NEPDIS provides the SME with the opportunity to record objective data about each task and then uses the computer to compare sets of objective data among tasks to determine the degree to which the tasks are common.

When an SME is assigned to record task data for a given set of tasks for a rating, he first fills out a job data worksheet which contains descriptive information--e.g., what the task is, where it is performed, what cues initiate task performance, and what requirements there are for successful task performance (necessary tools, equipment, materials, and references).

In addition, the SME records the major functional category, or primary division into which Navy tasks are classified. Examples of major functional categories are maintenance, operation, and administration. The SME also records the duty subcategory of the task. Each duty subcategory is a further breakdown of a major functional category. For example, in the major functional category of maintenance, a task may be further classified as checking/testing/inspecting, performing preventive maintenance, or performing corrective maintenance.

Once the job data work sheet is complete, the SME turns his attention to the task data worksheet. After recording the rating in which the task is performed and the task number, the SME deals with a series of descriptive characteristics that define each task.

Descriptive characteristics are unique to the NEPDIS model. Within several different categories, the SME is asked to specify which of three statements best describes performance requirements of each task. These categories include a general category that applies to all tasks, the specific duty subcategory into which the task has been classified and five skill areas related to the use of references, tools and equipment.

Figure 2 contains examples of some of the three statement choices for the duty subcategory, "performing corrective maintenance." A "1" statement is the least complex; a "3" statement is the most complex.

A. REMOVAL/REPLACEMENT

1. Simple change of location - requires no fastening/unfastening (lift, unplug, push aside, etc.).

2. Dual action - fastening/connecting/unfastening/disconnecting in addition to change of locations.

3. Multiple action - requires other supporting actions in addition to fastening/connecting/unfastening/disconnecting and change of location. Example: remove/replace aircraft engine.

B. CONNECTORS (CONNECTING/DISCONNECTING)

1. Single connector - single type. Example: Cannon plug.

2. Multiple connectors - single type. Example: 2 leads on resistor or cannon plugs.

3. Multiple connectors - multiple types. Example: bolts, cannon plugs, and resistor leads.

Figure 2. Examples of Descriptive Characteristics from the Duty Subcategory "Performing Corrective Maintenance."

Figure 3 shows how the computer displays these data. The descriptive data about the task from the job data worksheet is printed first. The "signature block" comes from the task data worksheet. Each successive number in a horizontal row reprecents the statement selected for the respective choice in each category or skill area.

Once in the computer, these data may be used to calculate two additional pieces of information that further describe the task, and may lead to a description of the task in terms of its commonality with other tasks. The first is the complexity index, or a number derived from weighted values assigned to the descriptive characteristics. The complexity index is the weighted sum of the values of all the descriptive characteristics divided by the weighted sum of what the values would be if the characteristics were as complex as possible, i.e., all 3's, and then converted to a five-point scale.

The second value that can be calculated from the task data is the degree to which two (or more) tasks are common. If two tasks have the same numerical value for one descriptive characteristic, they are identical for that characteristic. Of course, the more characteristics they have in common, the more alike the tasks are.

Using the two concepts of task complexity and task commonality, tasks may be divided into four groups of tasks with respect to their commonality:

(1) Unique tasks. Tasks which are not common with any others in the set of tasks being examined.

```
NO. 2          RATING=AE        PKG=0004      TASK=0021       DS=01

PLATFORM=P-3 A/B                          SYSTEM=WSHLD HTG SYS
EQUIPMENT=NC                              COMPONENT=NC
MODULE=NC

      ACTION=IFT, ISOLATE FAULT/TROUBLESHOOT WSHLD HTG SYS 41330

      CUE.....................MALFUNCTION
      REFERENCE...............NA-01-75-PAA-2-2.4
      SUPPORT EQUIP...........POWER UNIT NC-12
      TEST EQUIP..............TEST SET WNDSHLD ANTIICE FT247
      STANDARD................IAW REFERENCE PUBLICATION
      TOOL....................COMMON HAND TOOLS
      TEST EQUIP..............MULTIMETER PSM-4
      TEST EQUIP..............BLOWER ASSY R1730-218-88

      COMPLEXITY=2.20

      GENERAL     21231100000000000000000000000
      DUTY SUB01  33232210000000000000000000000
      SKILL 1     21133000000000000000000000000
      SKILL 2     12110000000000000000000000000
      SKILL 4     23311330000000000000000000000
      SKILL 5     32333333000000000000000000000
```

Figure 3. Computer Printout of Task Data in the NEPDIS Task Inventory File.

(2) Identical tasks. Tasks which are 100 percent common with other tasks. When one trains one tasks, he/she trained all other tasks identical to it.

(3) Omnibus tasks. Tasks which contain other tasks. For example, in the AE rating, the task "Perform an operational check on the SW Rate Gyroscope CN 495A/AJB-3" includes the task "Perform a bench check on the automatic pilot warning lights flasser 52118." Therefore, the operational check has a higher complexity index than the bench check. In every instance where their descriptive characteristics are not identical, the descriptive characteristics of the operational check have a higher numerical value. If one has trained for the omnibus task which is the operational check, then one has necessarily also trained for the bench check.

(4) Embodied tasks. Tasks included in other tasks. The bench check in the preceding example is embodied in the operational check.

Examination of the total tasks contained in any rating to identify these kinds of tasks may result in a great reduction in the need for training. For example, the experimental job task inventory file, through the identification of identical tasks, reduced the task inventory for the Aviation Electrician's Mate (AE) Rating from 3,772 tasks to 1,180 tasks. A further reduction resulted in all of the 1,180 tasks being contained in 469 omnibus tasks. This goes to

say that a training program (or programs) which trained the 469 tasks would, within the parameters of the data stored for each task, in effect, have trained the 3,722 tasks.

Present paper and pencil task analysis does not make possible the realization of a Navy training system. The stockpile of data precludes manual manipulation; a computer must be used to bring job data into a manageable perspective. A rough calculation of 84 ratings multiplied by thousands of tasks per rating, suggests that the Navy requires an ADP (automatic data processing) system for tracking the skills required to operate and maintain the Navy's equipment just as surely as it needs ADP systems to account for that equipment.

This apparent success, or the promise which such a reduction seems to hold, is indeed encouraging, but not to be taken lightly. There are many questions which are yet to be answered. Some of them have been answered, at least in part.

For example, whether the computer can rank tasks for commonality as well as technicians can rank the same tasks, is one question that deserves an answer. The Naval Education and Training Command has conducted some preliminary studies in search of an answer, and the results are encouraging. In one study, commonality rankings of two groups of personnel were correlated with computer rankings. The first group was comprised of personnel in paygrades E-3 through E-5; the second group was comprised of personnel in paygrades E-6 through E-9. The study concluded that responses of senior personnel correlated highly with the computer data and the data are significant (alpha = .05) in the majority of cases. Data from less experienced personnel did not correlate as highly with the computer responses as did the data from the senior paygrades in this study, nor did the junior personnel agree among themselves to the high extent that the senior personnel did.

Other studies are being planned. One will involve inclusion of the computer rankings with technician rankings to see whether the coefficient of concordance is higher with technician rankings alone or after the computer rankings have been included. Other tests are planned to determine whether learning analyses of identical tasks can be utilized interchangeably and whether one can, in fact, perform an embodied tasks if the associated omnibus tasks can be performed.

More than three-quarters of a century elapsed before Navy trainers, with the help of Rundquist, actually employed the job/task analysis procedures advocated by Frederick Taylor. Computer storage and manipulation of discrete tasks is another bold movement, perhaps too bold for unanimous acceptance at this time, but surely a move with promise enough to deserve, and hopefully withstand, the tests that are yet to be applied.

REFERENCES

Ansbro, T. M. Using the computer to build the task inventory. Proceedings, 20th Annual Conference of the Military Testing Association. Oklahoma City: 1978.

Bureau of Naval Personnel. Fundamentals of Navy curriculum planning. Washington, DC: NAVPERS 98510-1, 1968.

Butterworth, R. M. Task analysis of U.S. Navy enlisted radiomen with emphasis on technical controllers at the U.S. Communications Station, San Francisco, California. Unpublished thesis. Monterey, CA: Naval Postgraduate School, 1973.

Cagle, M. W. Dimensions. TRANAVY: magazine of naval training. Pensacola, FL, 1973, 9, 1.

Chief of Naval Education and Training. NEPDIS automated data systems implementation plan. Pensacola, FL: 1977.

Chief of Naval Personnel. Navy enlisted occupational classification system. Washington, DC: Bureau of Naval Personnel, 1974.

Commission on Instructional Technology. To improve learning: a report to the President and the Congress of the United States. Washington: U.S. Government Printing Office, 1970.

Fink, C. D. An analysis of CTAD instructional program development training task analysis procedures, NAVEDTRA 106A, supplement 2. Alexandria, VA: Kinton, Inc., 1978.

Individualized Learning Development Group. Job task analysis manual. San Diego: Naval Training Center, Service School Command, 1974.

Rundquist, E. A. Job training course design and improvement. San Diego: Naval Personnel and Training Research Laboratory, 1970.

Scanland, Worth. Instructional technology in naval training. Proceedings of the National Security Industrial Association Conference on the State-of-the-Art Application of Advanced Training Technology. NSIA, Washington, DC: 1977.

Taylor, Frederick W. Shop management. New York: Harper & Brothers, Publishers, 1974.

Taylor, Frederick W. The principles of scientific management. New York: Harper & Brothers, Publishers, 1911.

Tracey, William R., Flynn, Edward B., and Legere, C. L. The development of instructional systems. Fort Devens, MA: U.S. Army Security Training Agency and School, 1970.

DELONEY, J.R., & BURT, J.A., U.S. Coast Guard Institute, Oklahoma City, Oklahoma.

AN ASSESSMENT OF OOD COMPETENCY IN THE WAKE OF THE BLACKTHORN INCIDENT BY APPLYING A MODIFIED MAPL PROCEDURE (Tue P.M.)

One of the results of the collision and sinking of the U.S. Coast Guard cutter Blackthorn was the development of a new test on navigation rules to be administered to the OODs (Officer of the Deck) in the Coast Guard. The purpose of the test is to determine knowledge of the rules of the road. A modified application of the Minimally Acceptable Performance Level (MAPL) technique, a procedure developed by Meredith (1977) based on the earlier work of Nedelsky (1954), was used to determine what score represented a minimal level of competency. The technique employs subject matter specialists to rate each distractor of each item as to whether a minimally acceptable OOD would or would not know whether the distractor was a wrong choice. Major modifications were made in the MAPL technique to obtain consensus on whether a minimally acceptable OOD would judge the distractor as correct or incorrect. The procedure and modifications are described in the paper with example data. Application of the procedure to other types of military testing are also discussed.

Assessment of Competency Following the BLACKTHORN Incident by Applying a Modified MAPL Procedure

Julia R. Deloney John A. Burt

Coast Guard Institute
P.O. Substation 18
Oklahoma City, Oklahoma 73169

One aftermath of the sinking of the Coast Guard cutter BLACKTHORN was the modification of the method for qualifying Officers of the Deck (OOD). The major modification was the development of a paper and pencil examination on navigational rules of the road to be used in addition to the standard certification procedure which required knowledge of or experience with factors such as shipboard organization, search and seizure, international agreements, pollution control, navigation, and shiphandling. The purpose of the test was to determine knowledge of rules of the road. A modified application of the Minimally Acceptable Performance Level (MAPL) technique, a procedure developed by Meredith (1977), based on the earlier work of Nedelsky (1954) was used to determine what score represented a minimal level of knowledge.

Method

Two parallel forms, Form 51 and Form 52, of the test were developed by Coast Guard Institute personnel based on manual CG-169 Rules of the Road. This manual contains the rules to be followed by vessels traveling in international waters and in inland waters. Each test contained 100 items broken into three sections; International Rules, Inland Rules, and Pilot Rules for Inland Waters. All three sections were further subdivided into General Rules and Terms, Light Displays and Dayshapes, and Sound Signals.

The original procedure for applying the MAPL technique employed a panel of experts to consider and rate each distractor of each item of a test. A MAPL score was then computed from the results of the ratings.

Six subject matter specialists from the Coast Guard Institute were employed in assessing each item of the two forms of the test. Time and financial constraints limited the choice of panel members to personnel available at the Institute. However, those who participated as panel members were experts in the field. The specialists were chosen based on their previous experience as OOD, as commanding officers of ships, and with the navigation rules. As seen in Table 1, two specialists were item writers who were former Merchant Mariners, two had served as OOD for 1½ and 2 years, and two had served as both OOD and commanding officer. One had also served as Executive Officer. Three of the experts held licenses as Merchant Mariners. The six specialists represented approximately 53 years experience with the navigation rules.

	CO	XO	OOD	MVP ITEM WRITER	LICENSED MERCHANT MARINER	TOTAL YEARS EXPERIENCE
SMS 1	X	X	X			4½
SMS 2				X	X	24
SMS 3				X	X	6
SMS 4			X		X	1½
SMS 5			X			2
SMS 6	X		X			15
TOTAL YEARS EXPERIENCE						53

TABLE 1. - EXPERIENCE OF SUBJECT MATTER SPECIALISTS.

The MAPL procedure as applied by Meredith required each member of the panel of six to eight subject matter specialists (SMSs) to independently rate each distractor of each item according to the following criteria. A rating of 0 was assigned for yes, meaning the minimally acceptable person would know this was a wrong answer; 1 for neither yes or no; and 2 for no, meaning the minimally acceptable person would not know this was a wrong answer. (See Table 2).

Would the minimally acceptable person know this alternative is a wrong answer?	
0	YES, S/HE WOULD KNOW THIS IS A WRONG ANSWER
1	NEITHER YES OR NO
2	NO, S/HE WOULD NOT KNOW THIS IS A WRONG ANSWER

Table 2. - Question and ratings Subject Matter Specialists applied to each distractor.

The correct option was always assigned a value of 2. The distractors of the first six items were rated independently. The panel members then discussed the ratings they had assigned until all members agreed on the rating to be assigned to each distractor. The purpose of this discussion was to insure that each panel member fully understood the criteria for rating each distractor. The SMSs then independently rated the distractors for the rest of the test.

Table 3 gives the MAPL computation procedure. To compute the MAPL score, the points assigned to each item were totaled for each SMS. Then the SMSs totals were summed over each item and the average value was determined for each item. This was called the Average Total Points (ATP). Next, the Alternative Similarity Index (ASI) was computed by dividing the ATP into the correct option value, 2. All ASIs were added to find the MAPL score. To convert this to a percentage score, the MAPL score was divided by the total number of questions and multiplied by 100.

ITEM (* = CORRECT CHOICE)

#	CHOICE	SMS-1	SMS-2	SMS-3	SMS-4	SMS-5	ATP	ASI
1.	A	2	1	2	2	1		
	*B	(2)	(2)	(2)	(2)	(2)	31 ÷ 5 = 6.2	
	C	1	0	1	2	1	2 ÷ 6.2 =	
	D	2	0	2	2	2		0.32
		7	3	7	8	6		
2.	A	0	1	0	0	1		
	B	0	0	1	1	1	17 ÷ 5 = 3.4	
	C	1	0	0	1	0	2 ÷ 3.4 =	
	*D	(2)	(2)	(2)	(2)	(2)		0.59
		3	3	3	4	4		
3.	*A	(2)	(2)	(2)	(2)	(2)		
	B	0	2	2	1	1	30 ÷ 5 = 6.0	
	C	1	2	1	2	1	2 ÷ 6.0 =	
	D	1	1	2	1	2		0.33
		4	7	7	6	6		

$\sum$ASI = 0.32 + 0.59 + 0.33 = 1.24 = MAPL Score
1.24 ÷ 3 = 0.413 X 100% = 41.3% = PERCENT SCORE of MINIMALLY ACCEPTABLE PERSON

TABLE 3. - MAPL computation procedure.

(242)

The aim of the MAPL procedure is to establish an overall score that represents the level of achievement required for a person to be considered minimally acceptable. The way the MAPL procedure produces a minimally acceptable cut score can be explained as follows. Suppose all distractors for all items in a test were assigned a rating of 0, meaning a minimally acceptable person would know that all the distractors for the items were incorrect. Assuming that the examinee can identify the correct response by eliminating the distractors known to be incorrect, a value of 2 is assigned the correct response; which means, as defined in Table 2, at minimum the examinee would not eliminate the correct response. This condition would represent the minimal situation in which a correct response could be attained (discounting guessing). Consequently, in calculating the ASI, the baseline 2 is divided by the ATP. In the case where all distractors for each item on a test are assigned a 0, the ATP value for each item would be 2, and 2 divided by the baseline 2 equals 1, the ASI (See Table 3). Summing the ASI's and computing the percentage yields a percent score of 100 for the minimally acceptable person (See Table 3). Again, suppose all distractors were assigned a rating of 1. The calculated ATP value is larger and the ASI value is smaller. Summing the ASI's and computing a percent score yields a lower percent score the minimally acceptable person must score to be considered minimally acceptable. Once more, assigning a rating of 2 to each distractor raises the ATP even more, further reduces the ASI, and again lowers the percent score the minimally acceptable person must achieve to be considered minimally acceptable. In this manner as the distractors in test items become more difficult to discriminate from the correct response by the minimally acceptable person, the lower the MAPL technique establishes the cut-point for the test.

The method used was modified by the Institute from the usual application of the MAPL technique to obtain consensus on the ratings assigned by the SMSs for each distractor. Whereas independent ratings by SMSs had been stressed by Meredith, it was judged that if consensus was not reached in the first rating assignment, there was a problem in interpretation of the question by the SMSs, disagreement as to the performance level of the minimally acceptable person, or the item needed revising. This was discovered in the discussion process necessary in reaching consensus. The result of this method of obtaining consensus was that a more accurate score of the performance level of the minimally acceptable person could be determined. (See Figure 1.)

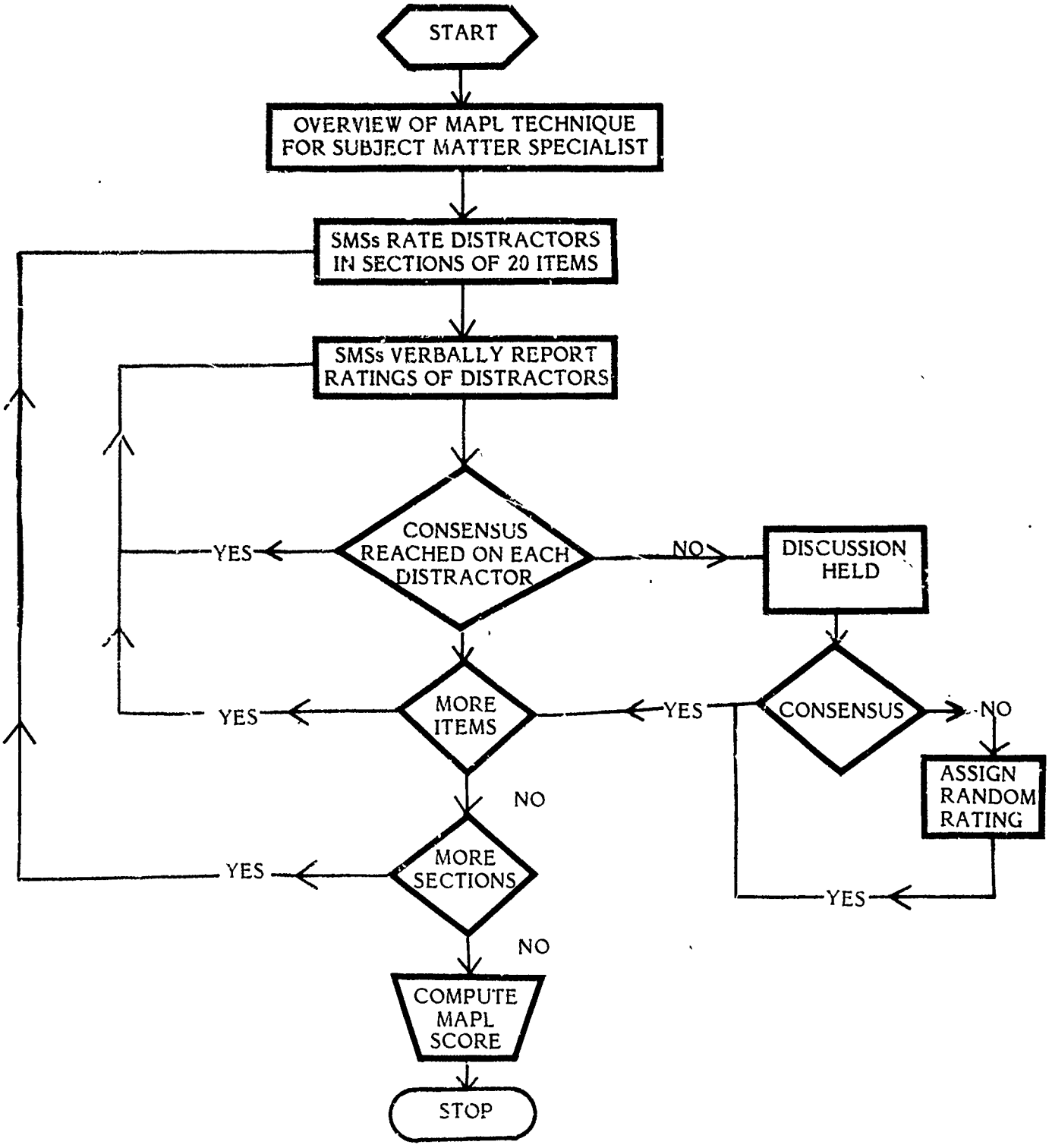


Figure 1. - Flowchart of modified MAPL procedure.

To implement this modification it was necessary to enlarge the role of the facilitator who in the original procedure participated only in the presentation of the MAPL technique to the panel and in the discussion following the rating of the distractors of the first six items. With the modification the panel rated the distractors in groups of 20 items at one time with a 10 minute time limit. Then each SMS verbally reported the rating he had assigned each distractor as the facilitator recorded each response. If four of the six SMSs had assigned the same value to a distractor, consensus was considered to have been obtained. If four had not agreed on the rating value, discussion was held in order to obtain agreement. Discussion was limited by the facilitator to two minutes to insure efficiency in reaching consensus and to keep the SMSs on task. The two minute limit was necessary because of strict time constraints. If consensus was not reached within two minutes, a rating was randomly assigned to that distractor by the facilitator. A random assignment was made based on the assumption that if agreement among the SMSs could not be reached it may be assumed that assignment of a value must be near random. This method encouraged consensus since the SMSs preferred to compromise rather than having a rating value randomly assigned. This procedure was followed until all items of the tests had been rated.

Obtaining consensus on each distractor of each item led to an adjustment in the computation of the MAPL score, as shown in Table 4. Since a consensus rating was used, it was necessary to sum over only the consensus SMS rating for each item rather than for all six SMSs to find the ATP. The ASI was then found by dividing 2 by the ATP. Summing the ASIs gave the MAPL score. To find the percent score of the minimally acceptable person, the MAPL score was divided by the total number of questions in the test and multiplied by 100.

ITEM #	CHOICE	SMS	ATP	ASI
1	A	1		
	*B	(2)		
	C	0		
	D	1		
		4	4	2 ÷ 4 = .5
2	A	1		
	B	0		
	*C	(2)		
	D	0		
		3	3	2 ÷ 3 = .67
3	A	1		
	B	1		
	C	1		
	*D	(2)		
		5	5	2 ÷ 5 = .4

(* = CORRECT CHOICE)

$\sum$ASI = .5 + .67 + .4 = 1.57 = MAPL Score
1.57 ÷ 3 = .523 X 100 = 52.3% = Percent Score of Minimally Acceptable Person

Table 4. Modified MAPL computation procedure.

The computation of the MAPL score resulted in a final cut-score of 66% for Form 51 and 73% for Form 52 of the test.

As seen from Table 5, of the 300 distractors for Form 51 of the test, consensus was reached immediately for 225 distractors, consensus was reached following discussion for 67 distractors, and rating values were randomly assigned to 8 distractors. For Form 52 of the test, consensus was reached immediately on 232 distractors, consensus was reached following discussion for 65 distractors, and rating values were randomly assigned to 3 distractors. Of a total of 600 distractors consensus was reached immediately on 457 leaving 143 to be discussed for a final rating value.

	CONSENSUS	CONSENSUS AFTER DISCUSSION	RANDOMLY ASSIGNED	TOTALS
FORM 51	225 (75%)	67 (22%)	8 (3%)	300 (100%)
FORM 52	232 (77%)	65 (22%)	3 (1%)	300 (100%)
Totals	457 (76%)	132 (22%)	11 (2%)	600 (100%)

Table 5. - Number of Distractors For Which Subject Matter Specialists Reached Rating Consensus, Consensus Following Discussion, and Number of Distractors having Ratings Randomly Assigned.

Conclusions

The results of the MAPL computation identified 66% on Form 51 and 73% on Form 52 as the scores representing the minimal acceptable knowledge level of navigational rules of the road for Coast Guard officers standing duty as underway OOD.

At least three aspects of the modified technique were identified as significant: (1) the role of the facilitator was expanded from that of the original procedure, (2) consensus versus independent ratings, and (3) implications for item revision.

The role of the facilitator is an important aspect of the procedure and can, in fact, "make or break" the process. With a panel of six to eight subject matter specialists, it was easy to disagree, to get off-task. Often the panel members miscued on insignificant or irrelevant issues in the discussion process. This seemed to occur more frequently with fatigue. The facilitator had to keep the panel members on task and to insure that the process operated correctly. As viewed in Table 6, it was necessary for the facilitator to intervene at various points throughout the procedure, and the intervention points were crucial to the procedure.

Table 6. - Facilitator's Role and Intervention Points

- Presented Overview of MAPL procedure to SMSs
- Initiated rating procedure for a group of items
- Timed rating sessions and called time
- Initiated and recorded verbal responses of ratings by SMSs
- Determined whether consensus was reached for each distractor
- If consensus was not reached for each distractor, called for discussion
- Limited discussion to two minutes
- During discussion, led panel in the attempt to reach consensus by informing them of what ratings had been given and possible directions to take in reaching consensus
- If consensus could not be reached, assigned a random rating value (may be done just prior to computation of MAPL score)
- Continued procedure until all items were rated
- Computed MAPL score

(246)

The issue of consensus is an important one. The purpose of the procedure is to determine a minimally acceptable performance level for a particular type of activity. When four of six SMSs agree on the assignment of a value to a distractor, this is assumed to be "prima facie" evidence for general agreement. If at least four do not agree, and particularly if there is a wide dispersion of values, this is an indication of a problem. It may be a disagreement as to whether the minimally acceptable person should or should not know the answer. However, other issues may be involved such as, "Is the question unclear, or is the question relevant". The process of attempting consensus can have one of three outcomes:

(1) Consensus is reached.

(2) The rating of the distractor is randomly assigned.

(3) The item or distractor is revised.

An unexpected dividend in using this modification of the MAPL technique was a system of item review. The items in the test had previously been submitted for review. However, during the MAPL procedure items were identified that had passed review but were questionable on a point previously overlooked. Therefore, it appears that this modification offers an improved test development procedure in the item revision step.

Since the purpose of the MAPL technique is to identify the cut-score on a test for the minimally acceptable person, or the minimally acceptable performance, this modified MAPL technique could be applied to any criterion-referenced testing situation. This would include lesson quizzes, end of course tests, promotion-selection tests, achievement tests, or any type of competency based test.

The MAPL procedure is also a policy-capturing procedure. The panel members represent the Coast Guard's policy for determining knowledge that the minimally performing person must have.

The Institute's experience with the MAPL technique has been a positive one. It has provided an objective and systematic process for determining a pass/fail cut-score with the added sidelight that the application of the MAPL technique served as a systematic approach to item review. It was objective in the sense that a criterion was set, which was the minimally acceptable person, and was used in assessing each distractor. This standard criterion was strictly adhered to across each item during the entire process. Using the same criterion systematically throughout the assessment procedure identified items that needed revision when the distractors were inappropriate for the criteria.

References

Meredith, J.B. The use of a criterion-referenced test methodology to assess training system effectiveness. Paper presented to the 19th annual meeting of the Military Testing Association, San Antonio, Texas, 1977.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

DEMPSEY, John R., Kentron International, Inc. and FAST, Jonathan C.,
Air Force Human Resources Laboratory, Brooks AFB, Texas.

DEVELOPMENT OF ENLISTMENT STANDARDS - AN APPLICATION OF THE LIFE MODEL (Tue A.M.)

The likelihood Function Estimation (LIFE) Model was developed by the Air Force to better model a dichotomous (binary) dependent variable. Under contract the prototype model was enhanced to improve its data handling procedures and the maximization algorithm. The LIFE model prototype had been used to develop an alternative Air Force enlistment standard designed to increase the number of qualified accessions with no increase in first term attrition. This enlistment standard was implemented in a test in October 1978 by the Air Force Recruiting Service and was called Project Image. The initial results from Project Image are discussed and a comparision with the same equation developed by the improved model is also presented. The two equations were then applied to a 1975 sample for cross validation purposes.

DEVELOPMENT OF ENLISTMENT STANDARDS: APPLICATION OF THE LIKELIHOOD FUNCTION ESTIMATION (LIFE) MODEL

J. R. Dempsey, Kentron, Inc. and J. C. Fast, AFHRL

I. INTRODUCTION

Enlistment standards are the mechanism through which the Air Force personnel planners control the quantity and quality of recruits. These mental and physical standards are designed to (1) qualify adequate numbers of applicants to meet Air Force manning requirements, and (2) maximize the aggregate quality of the first term airman force in terms of mental and physical attributes. Unfortunately long-range prediction of human behavior is difficult; consequently, enlistment standards generate two types of errors. First, they allow enlistment to some applicants who will be unsuccessful in the Air Force. Second, they deny enlistment to some who would have otherwise succeeded.

Because the recruiting environment became increasingly adverse with the adoption of the All Volunteer Force concept, the Air Force started a research effort to improve the methodologies by which post-enlistment behavior was predicted. The development of the Likelihood Function Estimation (LIFE) Model, designed to meet this need, was detailed in a previous technical report (Dempsey, Sellman, & Fast, 1979). This LIFE model was a prototype version intended to show feasibility. This feasibility is currently being tested in an official Air Force test, known as Project IMAGE. Concurrently, a development effort was undertaken at the Air Force Human Resources Laboratory (AFHRL) to enhance the prototype version, to eliminate some of the problems noted by Albert (1980). Section II describes the enhancement of the LIFE prototype. In Section III, Air Force's Project IMAGE is discussed and some preliminary analyses of the results are presented. Section IV describes the follow-on work done in enhancing the computer model and in developing a new enlistment equation using a 1975 data base.

II. PROJECT IMAGE TEST

With the advent of the All Volunteer Force, the Air Force experienced good recruiting years. This good market caused Air Force managers to raise enlistment standards, so that only the most qualified would be allowed to enlist. The enlistment standard used the four composites from the Armed Services Vocational Aptitude Battery (ASVAB), combined with the educational level of the applicant. As a minimum, the applicant must have achieved a 45 on the General aptitude composite and a total of 170 on all four composites. In addition, the Air Force attempted to limit the number of non-high school graduates by applying more stringent mental standards (as measured by the Air Force Qualification Test [AFQT] composite of the ASVAB). This resulted in a higher percentage of high school graduates among the recruit population. This combination of enlistment standards in general raised the quality of Air Force recruits, but at the expense of turning away many otherwise qualified applicants. When the more austere recruiting years arrived in the late seventies, the Air Force was faced with not being able to maintain the desired force level and their high enlistment standards.

The application of the prototype LIFE model to predicting attrition was brought to the attention of Air Force managers, and a plan to test it as an alternate enlistment standard was developed. The test was named Project IMAGE and, under the plan, the equation developed for the demonstration was to be used to waiver individuals into the Air Force. If an individual passed all

the enlistment standards, except the ASVAB General 45, composite 170 standard, he/she would be further processed through the IMAGE equation. If this equation predicted the individual would be successful in the first term, he/she would be allowed to enlist and would be assigned individually to a particular Air Force Specialty Code (AFSC). Each record of an individual who was allowed to enlist with an IMAGE waiver was flagged so that the IMAGE enlistees could be followed through training and into the Air Force. The Air Force Manpower and Personnel Center (AFMPC) required quarterly updates on the attrition rates of the IMAGE enlistees and comparisons with the other categories of recruits.

Test Results

The IMAGE test was started 1 October 1978 and was completed on 31 May 1979. During this period 3,911 people were waived into the Air Force under Project IMAGE. By comparison 62,704 were allowed to enlist in the Air Force during this same time period who passed the General 45, composite 170 standard, and educational qualification. These two groups will be tracked for 4 years (through the first term) and results reported at that time; however, it is worthwhile to discuss the preliminary results of the test through 30 September 1979. The analysis looked at the attrition experience of the two groups of recruits (pass G45, C170, and fail G45, C170) even though some of this group had just completed basic training and others had been in the field for 6 months or more. The analysis as a result shows only general trends and should not be interpreted as a comprehensive evaluation of the IMAGE equation as an enlistment standard. In addition, since IMAGE people were allowed to enter only selected, hard-to-fill AFSCs (with high attrition), the follow-on analysis will compare IMAGE vs. non-IMAGE people, after adjusting for the differences in AFSC attrition rates. The current analysis, however, was not broken out by AFSC, which may not reflect the actual utility of the IMAGE equation.

Table 1 shows general characteristics of the IMAGE enlistees. The important features are that only two were non-high school graduates and the vast majority were measured as mental category III-B by the AFQT composite. Table 2 shows the same characteristics for the other group of recruits who passed the G45, C170 standard. In this group, 18.1% failed to graduate from high school and they were fairly evenly divided between mental categories II, III-A and III-B. Table 3 contains the FY79 attrition analysis for the two groups of recruits. Overall, the IMAGE attrition rate of 8.9% is not significantly different from the current standard group rate of 8.8%. The male IMAGE group has attritted at a slightly higher rate through BMT and tech training. The female attrition rate for the IMAGE group for both BMT and TT was much higher than female current standard accessions; however, due to the small number of women, valid comparisons are difficult to make. The effect of these differences are absent in the total sample because the IMAGE group is predominantly male (97%), and females overall attrit at a higher rate. The limit on the number of females in the IMAGE sample was a result of allowing only IMAGE females to enlist in hard-to-fill, (non-traditional) female jobs. Not many IMAGE females who would have otherwise been qualified could qualify on the mechanical and electrical composites of ASVAB for these non-traditional jobs.

The IMAGE equation did fulfill its promise of increasing Air Force accessions by 6% without increasing attrition. Based on the promising results obtained through September 1979, the Air Force Deputy Chief of Staff for

Manpower and Personnel (AF/MP) declared the IMAGE test successful and ordered the IMAGE waiver installed permanently as part of the enlistment process. For operational use, IMAGE qualified personnel will be allowed to enlist in any Air Force job for which they qualify, and it will be done in the PROMIS system, rather than individually.

IV. LIFE MODEL APPLICATION

Model Enhancement

After enhancement to the LIFE model, the model was still able to handle only 3,000 observations and 12 independent variables. To solve this problem, AFHRL/MOMD computer personnel converted the model to reading the data into mass storage instead of into a matrix. This increased data holding capability to up to 10,000 observations with 20 independent variables, and this will allow task scientists to handle almost any binary prediction problem that will be likely to arise. This modification is not necessary for the research scientist who has access to a virtual memory machine. On this type of computer, the matrix can be expanded greatly to meet data requirements without exceeding core limits. The only limit becomes central processor time available, and the enhanced version of LIFE should make longer problems practical even on a busy machine.

Prediction Equation Development

The prediction equation used in Project IMAGE was developed 3 years ago from a 1972 data base. Although it has been successful, this equation needed to be updated by replacing it with an equation developed using the LIFE model on a more recent data base. As a result, work was initiated to develop two data base samples, taken from the population of 1975 Non-Prior Service (NPS) recruits into the Air Force. Two samples of 3,000 observations each were developed from this population, one for prediction equation development and one for cross-validation. After removing records with out-of-range ASVAB scores, the prediction development sample contained 2,541 valid cases and the cross-validation sample contained 2,526 cases. In the prediction sample, 744 were discharged from the Air Force within 36 months after enlisting and 839 were discharged within 42 months after enlisting. An attempt was made to develop a prediction equation for both criteria to determine the difference in predictive accuracy. These two equations developed using the LIFE model are shown in Table 4. The equations are very similar with only slight variations in the significant variables. The prediction accuracy of the two equations is compared in Tables 5 and 6. These hit tables show how well the two equations were able to identify actual successes and failures correctly. The equation developed was more accurate in predicting attritions for the 42-month criterion than the 36-month criterion (55.4% versus 52.0%). However, the equation was more accurate for predicting successes on the 36-month criterion than on the 42-month criterion. Because the specific purpose of IMAGE would be to waiver a predicted success into the Air Force and because the 36-month criterion is also the one used by the Office of the Secretary of Defense as the proper measure of attrition, 36 months attrition was used as the criterion of interest in the rest of this study.

Using the 36-month criterion, the next part of the study was to compare the ability of the original IMAGE equation for predicting success to the ability of a new equation using the LIFE model. In order to make this comparison, cases with incorrect AFQT scores were eliminated. This reduced the

prediction sample to 2,522 cases, and the cross-validation sample to 2,508 cases. It was conjectured that the current IMAGE equation would not predict success well on a new sample for several reasons. First, the IMAGE equation was developed on an all male sample of 1972 recruits, and the new sample included females in it. Second, the IMAGE equation predicted success using data from the 1972 sample (means and standard deviations) and these were very different from the data in 1975. In 1972, 86% of the sample were 18- to 26-years-old; in 1975, 98% of the sample were in this age group. In order to find a significant change in attrition behavior, this age bracket had to be closed to 18 to 23, which still included 92% of the sample. The English indicator changed in a similar fashion. In 1972, 6% of the sample had failed to complete a high school English course; in 1975, only 2% had failed to complete English.

Table 7 allows an inspection of the LIFE equation and the IMAGE equation. The AFQT score and the Trigonometry score were not significantly weighted in the LIFE equation and were left out since only 10 variables could be included in the original LIFE Model. The Physics variable did appear in the LIFE equation but not in the IMAGE equation. In addition, magnitudes and signs of the coefficients differed significantly between the two equations. These differences appear to be primarily due to the changes in samples from 1972 to 1975.

A third equation was developed using the expanded data handling version of the LIFE model. This equation included the AFQT variable and the Trigonometry variable, as well as the other 10 variables included in the LIFE prediction equation. This equation is also shown in Table 7. A total of four different prediction systems were generated for predictive ability comparisons. These were as follows: LIFE model with 10 variables (LIFE equation); LIFE model with 12 variables (LIFE equation with AFQT); IMAGE equation with 1972 means and standard deviations (old equation, old data); and IMAGE equation with 1975 means and standard deviations (old equation, new data). These four predictive systems were compared in two different ways--classification accuracy in a two-by-two contingency table and goodness of fit with a sum of squares statistic. Table 8 shows the contingency table accuracies for the prediction sample and Table 9 shows the contingency table accuracies for the cross-validation sample. The four prediction systems were very similar on the prediction sample, with success prediction accuracy ranging from 73.2% to 73.6% and failure prediction accuracy ranging from 49% to 51%. On the cross-validation sample (using the means and standard deviations from the prediction sample for a realistic application) the LIFE equation using AFQT performed better than the other three, but not by a large margin.

For the goodness-of-fit test, the actual occurrence (failure or success) was compared to the predictive probability of success, and the squared error was summed overall cases. Table 10 shows the comparison for the four systems and the two samples. There was no significant difference between the four systems on the prediction sample, but on the cross-validation sample there were significant differences. The old equation using the new data was significantly better at predicting the probability of success than the other three. This leads to the observation that the LIFE equation using AFQT will be best suited to the problem of predicting the occurrence of success among airmen, but will not be better than the current IMAGE equation for predicting the probability of this occurrence, using updated means and standard deviations.

(25-4)

IV. CONCLUSION

The LIFE model has been successfully transformed from a prototype to a useful analytical tool. The computer run time has been drastically reduced and the data handling capability enhanced.

The enlistment standard developed using LIFE has demonstrated the ability to increase the number of Air Force accessions without increasing first term attrition. In time of budgetary stringency, in which service attrition has an adverse impact on DOD manpower, resources and mission capability, enlistment standards must operate to allow enlistment to those applicants most likely to succeed. At the same time, enlistment standards should operate such that they do not turn away large numbers of potentially successful applicants. Project IMAGE has been successful in simultaneously demonstrating that these goals are achievable.

REFERENCES

Albert, W. G. Predicting involuntary separation of enlisted personnel. AFHRL-TR-79-58, AD-A082995. Brooks AFB, TX: Computational Sciences Division, Air Force Human Resources Laboratory, January 1980.

Berndt, E. K., Hall, B. H., Hall, R. E., & Hausman, J. A. Estimation and inference in nonlinear structural models. Annals of Economic and Social Measurement, 1974, 3(4), 653-665.

Dempsey, J.R., Sellman, W.S. & Fast, J.C. & Generalized approach for predicting a dichotomous criterion. AFHRL-TR-78-84, AD-A066661. Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, February 1979.

Table 1. Characteristics of IMAGE Enlistments
(Pass IMAGE - Fail Current Standard)

	Male		Female		Total	
	Number	%	Number	%	Number	%
TOTAL	3,780		131		3,911	
Ed Level						
HS Diploma	3,779	100	130	99.9	3,909	99.9
GED	....	...	...	....	.....	....
Other	1	...	1	.1	2	.1
Mental Cat						
I	.....	...	...	....	.....	....
II	4	.1	...	....	4	.1
III-A	285	7.5	8	5.9	293	7.5
III-B	3419	90.5	121	92.4	3540	90.5
IV	72	1.9	2	1.7	74	1.9
Mean ASVAB						
Mech	47.5		45.4		47.5	
Admin	46.6		53.8		46.9	
Gen	41.8		40.3		41.8	
Elect	50.7		46.1		50.6	
Comp	186.6		185.6		186.6	

Table 2. Characteristics of Current Standard Accessions

	Male		Female		Total	
	Number	%	Number	%	Number	%
Total	49,392		13,312		62,704	
Ed Level						
HS Diploma	40,548	82.1	10,784	81.0	51,332	81.9
GED	6,049	12.2	2,035	15.3	8,084	12.9
Other	2,795	5.7	493	3.7	3,288	5.2
Mental Cat						
I	3,171	6.4	625	4.7	3,796	6.1
II	18,552	37.6	4,434	33.3	22,986	36.7
III-A	18,437	37.3	5,835	43.9	19,092	30.4
III-B	9,219	18.7	2,415	18.1	16,814	26.8
IV	13	....	3	....	16	....
Avg ASVAB						
Mech	66.6		36.6		60.7	
Admin	64.9		74.3		66.7	
Gen	72.8		72.8		72.8	
Elect	72.6		60.4		70.2	
Comp	277.0		244.1		270.5	

(236)

Table 3. FY79 Attrition Analysis IMAGE Only
(Pass IMAGE - Fail Current Standard)

	Male		Female		Total	
	Number of Separations	% of Enlistments	Number of Separations	% of Enlistments	Number of Separations	% of Enlistments
BMT	214	5.7	19	14.5	233	5.9
Tech Trng	60	1.6	7	5.3	67	1.7
Post Trng	47	1.2	1	.8	48	1.2
Total	321	8.5	27	20.6	348	8.9

Current Standard Accessions (Pass Current Standard - Pass & Fail IMAGE)

	Male		Female		Total	
	Number of Separations	% of Enlistments	Number of Separations	% of Enlistments	Number of Separations	% of Enlistments
BMT	2,585	5.2	1,329	10.0	3,914	6.2
Tech Trng	733	1.5	193	1.4	926	1.5
Post Trng	516	1.0	148	1.1	664	1.1
Total	3,834	7.8	1,670	12.5	5,504	8.8

NOTE: Percentages may not add to 100% due to rounding error.

Table 4. Comparison of Coefficients of Prediction Equations

	Means	Coefficient For 36 Months	Coefficent For 42 Months
Intercept	.....	-.56	-.44
General	68.7	.004*	.005*
Composite	250.6	-.002*	-.001*
[1]Ed. Level	.11	.599*	.624*
[2]Algebra	.77	-.102	-.103
[2]Biology	.79	-.015	-.011
[2]Chemistry	.30	-.122*	-.083
[2]English	.98	-.107	-.064
[2]Geometry	.51	-.142*	-.146*
[2]Physics	.16	.088	-.053
[3]AGE	.92	.068	.110

3 - 1 if less than 18 or greater than 23, 0 otherwise
2 - 1 if taken in high school, 0 otherwise
1 - 0 if high school graduate or greater, 1 otherwise
*Asymptotic t value significant at .05 level

Table 5. Prediction Accuracy of Equations
42-Month Criterion

Category	Predicted Attritions	Predicted Successes	Total	Percent Correct
Actual Attritions	153	686	839	18.2
Actual Successes	123	1,579	1,702	92.8
Total	276	2,265	2,541	
Percent Correct	55.4	69.7	67.0	

Table 6. Prediction Accuracy of Equations
36-Month Criterion

Category	Predicted Attritions	Predicted Successes	Total	Percent Correct
Actual Attritions	138	606	744	18.5
Actual Successes	127	1,670	1,797	92.9
Total	265	2,276	2,541	
Percent Correct	52.0	73.4		

Table 7. Comparison of Equations

	IMAGE Coefficient	LIFE Prediction Equation 36-Month Criterion (10 Variables)	LIFE Prediction Equations Using AFQT
General	-.000006	.004	.005
Composite	-.001	-.002	.001
AFQT	-.0009	....	.001
Ed Level	.696	.599	.586
Algebra	-.120	-.102	.113
Biology	-.036	-.015	.008
Chemistry	-.027	-.122	.131
English	-.665	-.107	.131
Geometry	-.101	-.142	.147
Trigonometry	-.074	...	.017
Physics	....	-.088	.030
Age	-.198	.068	.092
Variance	1.065	1.012	1.0

(258)

Table 8. Contingency Table Accuracy for Prediction Sample (1975)

	Prediction Accuracy	
	Successes*	Failures**
Old equation, old data	73.3	50.8
Old equation, new data	73.6	49.0
Life equation	73.4	49.3
Life equation, with AFQT	73.2	51.0

*i.e., percent of predicted successes were actually successes.
**i.e., percent of predicted failures were actually failures.

Table 9. Contingency Table Accuracy for Cross-validation Sample

	Prediction Accuracy	
	Successes	Failures
Old equation, old data	73.5	48.1
Old equation, new data	73.7	47.2
LIFE equation	73.6	47.4
LIFE equation, using AFQT	74.9	49.6

Table 10. Comparison of Goodness of Fit

	Prediction Sample		Cross-validation Sample	
	SSQ	MSQ	SSQ	MSQ
Old equation, old data	253.50	.10	500.11	.20
Old equation, new data	299.50	.23	373.04	.15
LIFE equation	315.53	.13	514.41	.21
LIFE equation with AFQT	314.65	.12	510.54	.20

(?5?)

DICKINSON, Richard W., Occupational Research Division, Industrial Enginering Department, Texas, A&M University, College Station, Texas.

THE NEW CODAP SYSTEMS ENHANCED HIERARCHICAL CLUSTERING CAPABILITY (Wed A.M.)

In the task analysis arena, computer software to perform hierarchical clustering is generally restricted to a specific type of information (e.g., time spent on task indicies), allows few measures of similarity to be calculated and places constraining limitations on the type and number of objects that may be clustered. Such is not the case with the new CODAP systems enhanced heirarchical clustering capability. These enhanced capabilities are discussed with emphasis on ease of procedure use, lack of debilitating restrictions and simplicity of software modifications.

(26)

THE NEW CODAP SYSTEM'S ENHANCED HIERARCHICAL CLUSTERING CAPABILITY

Richard W. Dickinson

Occupational Research Division
Industrial Engineering Department
Texas A&M University
College Station, Texas 77843

INTRODUCTION

A new version of the job analysis computer software system, the Comprehensive Occupational Data Analysis Programs (CODAP), is being developed by the Occupational Research Division, Industrial Engineering Department, Texas A&M University. This paper discusses the incorporation of an enhanced hierarchical clustering procedure as part of the new CODAP system.

The hierarchical clustering procedure presently available to the majority of CODAP users restricts processing to tasks measured on time spent and allows only two measures of similarity to be calculated. Progress in the analysis of occupational information has mandated the development of procedures allowing greater flexibility in the selection of job analytic data for processing and in the type of processing such data is subject to. The new CODAP system is being developed to satisfy this need, with the enhanced hierarchical clustering procedure making use of such flexibility to enable the job analyst to apply clustering technology to a variety of occupational information.

The New CODAP System

At present, a new version of the job analytic software system (CODAP) is being developed that will greatly increase the job analysts' ability to answer questions of occupational information. In this new system, users access the occupational database through the use of an easy to learn, English-like language. There are no restrictions on data access and retrieval, allowing any piece of information residing on the database to be processed. The basic design philosophy of the new system is to conceptualize job analytic database information as a two-dimensional matrix in which incumbents represent the columns of the database, and the variables the incumbents are measured on representing the rows of the database (see Table I).

Using the new CODAP system, job analysts will have the ability to perform calculations on any variables (rows) in the database across any subset of incumbents (columns). The resultant calculations may then be added to the database for further processing. The flexibility of the

new system allows the added convenience of "symmetry," in which any calculations performed across database columns may also be performed across database rows. For a more in-depth discussion of the new system's operational capabilities and characteristics, the reader is referred to Dickinson (1979).

The Enhanced Clustering Procedure

The purpose of the enhanced clustering procedure is to perform a hierarchical clustering (Ward, 1963) on any set of columns of the database measured on any set of rows. Conversely, clustering may also be performed on any set of rows measured across any set of columns. In addition, the user will have the option of requesting any one of four techniques for calculating measures of similarity between columns or rows (see Appendix I for similarity formulae). Such flexibility will allow job analysts to perfrom such database operations as clustering incumbents on equipment usage or background characteristics. Further, the job analyst may not desire to cluster all incumbents, but a given subset of incumbents. Adhering to the concept of symmetry, the procedure will not be restricted to clustering just incumbents (columns), but will allow the clustering of tasks or any other set of rows in the database.

Output from the procedure will consist of a group membership report detailing the cluster breakdown, along with various vectors of information that will be stored on the database for future reference (these vectors will consist of hierarchical sequencing information that will be used by the enhanced DIAGRAM procedure. The enhanced DIAGRAM procedure produces a pictoral dendogram of the clustering operation, the discussion of which is beyond the scope of this paper).

The following system language source code will illustrate the use of the new procedure:

```
1)    BEGIN STUDYID EXECUTE.

2)    SELECT ROWS HISTINFO (H1-H2) 'HISTORY INFORMATION'.

3)    OVLGRP COLUMNS (INCUMBENTS) ON ROWS (HISTINFO) DSQUARE MAXIMIZE
            HSNID = HSN 'HIERARCHICAL SEQUENCE NUMBERS FOR COLUMNS'
         LOHSNID = LOHSN 'LOW HSN VALUES FROM CLUSTERING OPERATION'
         HIHSNID = HIHSN 'HIGH HSN VALUES FROM CLUSTERING OPERATION'
        SIMCOFID = SIMCOF 'SIMILARITY COEFICIENTS'
        WITHINID = WITHIN 'WITHIN COEFICIENTS'
         HEADING = 'EXAMPLE OF ENHANCED OVLGRP'.
4)    END.
```

The above example consists of a complete run sequence in the new system's language. Statement 1 is the first executable line in the new language. It alerts the system that a job is to be run, what study ID resides on the database of interest and that the following statements are

to be executed. Statement 2 is forming an aggregate of rows (a module), in this case H1 and H2, and assigning the name 'HISTINFO' to the module. Statement 2 is an example of the 'SELECT' procedure that is part of the new software system. This procedure's function is to identify aggregates of database rows or columns (based on user supplied instructions), attach user supplied names to them and them store them on the database for later reference. Any future occurance of the aggregate name will identify the associated rows or columns to the system for processing. Statement 3 is syntax for the enhanced OVLGRP procedure. The user is requesting that all incumbents in the study ('INCUMBENTS' is a system reserve keyword) be clustered based on their responses to the rows identified by the module name 'HISTINFO.' Distance squared (DSQUARE) is to be used to calculate overlap and 'MAXIMIZE' indicates that most similar columns are to be clustered first. The user supplied ID's HSNID, LOHSNID, HIHSNID, SIMCOFID and WITHINID are assigned to their respective vectors, stored on the database, and are later referenced by the enhanced DIAGRAM procedure. Any heading information will be printed at the top of the group membership output. Statement 4 ends the source language program.

CONCLUSION

The new job analysis computer software system being developed at Texas A&M University should open up new horizons in the field of occupational research. The incorporation of the enhanced hierarchical clustering procedure in the new system will allow the job analyst to apply clustering technology to a host of questions and, it is hoped, make the answers to these questions easier to come by.

TABLE I

CONCEPTUAL REPRESENTATION OF JOB ANALYTIC DATABASE

	I1	I2	I3	I4	I5	I6	I7	I8	I9
H1									
H2									
T1									
T2									
T3									
T4									
E1									
E2									
E3									
E4									

H = History or Background Information
T = Task Information
E = Equipment Usage
I = Incumbent

REFERENCES

Dickinson, R. W. The New Codap System -- Design Concepts and Capabilities. Paper presented at the Military Testing Association annual meeting, San Diego, California, 1979.

Ward, J. H., Jr. Hierarchical Grouping to Optimize an Objective Function, American Statistical Association Journal, 1963, 58, 236-244.

APPENDIX I

SIMILARITY MEASURE FORMULAE

SIMILARITY MEASURE	FORMULA
EUCLIDEAN DISTANCE	$D=\left[\sum_{i=1}^{i=n} (X_i - Y_i)^2\right]^{1/2}$
SQUARED EUCLIDEAN DISTANCE	$DSQUARE = \sum_{i=1}^{i=n} (X_i - Y_i)^2$
ABSOLUTE OVERLAP	$OVL = \sum_{i=1}^{i=n} \text{minimum } (X_i , Y_i)$
BINARY	$BINARY = \dfrac{\text{\# Tasks in Common Between X and Y}}{\text{\# Tasks Performed By X} + \text{\# Tasks Performed By Y} - @}$ @ = # Tasks in Common Between X and Y

NOTE: The symbols in the above formulas are defined as follows:

X_i and Y_i represent the ith elements in the data vectors of jobs X and Y, respectively.

X and Y represent the data vectors of jobs X and Y, respectively.

n = the number of elements in the data vectors of Job X or Job Y.

COMPUTER ASSISTED PERSONNEL SELECTION FOR THE ROYAL NAVY JUNE 1980

BERNARD T DODD
SENIOR PSYCHOLOGIST (NAVAL)
Ministry of Defence
Archway Block South IV
Old Admiralty Building
Whitehall, London SW1A 2BE

(27)

OVERVIEW

1. UK recruiters for the Royal Navy and Royal Marines are responsible for local schools liaison, manning the recruiting office, test administration, interviewing and allocation to a specific trade. Their accept/reject decisions are seldom reversed. With small entries they are faced with the choice of accepting an applicant or waiting in the hope that a better one will appear later or at another office. The penalties for errors are unfilled placed or entrants of lower quality than necessary. The Senior Psychologist (Navy) advises on selection procedures and attempts to set standards that optimise entry quality whilst minimising unfilled places. His division has developed a Computer Assisted Personnel Selection system (CAPS) which advises the recruiter on the eagerness of the Service to recruit the applicant whose personal characteristics and test scores he has just reported. This eagerness is expressed as a value attached to each entry date for each trade or branch for which the applicant is eligible. If recruiting is going well this value will be low. The recruiter is invited to accept his applicant if a high-valued opening can be found which satisfies the preferences of the applicant for type of work and joining date.

2. Numerical simulations are being used to prove the controlling programme and routines of the "electronic office" type are being developed to relieve the recruiter of his current heavy clerical load. Equipment procurement is in progress. What remains to be discovered is the extent to which the recruiter will take notice of the advice offered by the data link concerning the value of a particular applicant. Rigid adherence to the advice will tend to perpetrate the current system where places are filled largely on a first-come, first-served basis without reference to the need for even flows of talent throughout the year and a proper sharing out between trade groups.

The Basic Concept

3. The CAPS system in its present form appears to the Careers Adviser (CA) as two distinct sub-sustems; one tells him the state of recruiting, the other is an electronic booking clerk. The purpose of the first is to help him decide whether to accept the man before him or wait for a better: the second is to enable him to book his man and quickly get through the paper-work and statistical returns so that he can spend as much time as possible at his primary task of engaging the interest of potential recruits.

The booking function will require further programme development if the computer is going to write all the letters and compile all the statistical returns with the minimum of effort from the CA. To date, this project has done no more than illustrate the possibilities in this direction. What is not quite so straightforward us the design of the communication link from Director of Naval Recruiting (DNR) to the CA to advise him whether to offer a place to a specific enquirer. It is to this problem that the main effort of the CAPA R&D has been directed.

The simplest report on the state of recruiting is possibly the number of applicants forwarded for the current recruiting period and the number of places yet unfilled. These numbers are available to DNR and could be reported by telephone to the Careers Information Offices (CIOs). The figures will be slightly out of date due to various delays and the need to gather nation-wide bookings before an accurate picture can be developed. Under stable recruiting conditions, especially with large intakes, more timely feedback would not improve recruiting and the present regional quota system might well be satisfactory. However, if there are wide differences in applications between regions or at differerent times of the year,

it is possible for standards at one time and place to be less than optimum when the long term national result is considered. For example, a transient surfeit of high quality applicants for a particular branch ought to be accepted regardless of local quotas, even if extra recruiters have to be drafted in to process them. Similarly, a very small lowering of standards at all offices for a short time might make good a transient shortfall and avoid an irrevocable waste of new entry training places whereas a large long-term drop in standards tends to dilute branch quality levels and start a chain-reaction which may take many years to rectify.

Manipulation of entry standards should be purposeful and decisive, lasting only as long as is necessary to put the situation to rights.

4. Many SP(N) studies have shown the value of the Recruit Test (RT) score as the prime index of manpower quality. Higher scorers tend to stay and do well. Recruiting also takes into account other factors which are not deducible from test scores. These factors have been grouped into what are referred to as Personal Qualities and a summary of the CA's assessments of them is recorded as a Personal Qualities Assessment score. Research is still going on to refine the PQ assessment process and determine the value of PQ scores in predicting progress in the Service.

At the present state of research it seems reasonable to use the PQ total as an index of manpower quality in conjuction with the RT score although there are insufficient grounds for publishing minimum PQ scores. It has not been shown that low PQ scorers have a low probability of success.

5. Central to the argument for running CAPS on a computer with terminals at the recruiting sites is the possibility of advising the CA of the very latest recruiting achievement figures. A national data network could provide all users with the number of applicants in the pipeline for each branch for each entry into the foreseeable future. From this information the CA could work out that, say, an EM entry class 10 weeks away was fully booked and thereafter bookings tailed away to zero by week 20. Comparison with other branches would indicate whether this situation was a matter of general recruiting climate or something confined to that branch. If data were to hand, comparisons could be made with previous years, and so on.

The outcome of the assessment of the state of recruiting would be an impression of whether the number of persons booked for entry on a certain day to a certain branch was greater or less than should have been booked at that time. More people should have been booked for near entries than for distant: more for large intakes than for small.

Each CA would have his own ideas about how many should have been booked by a certain date and he would assess in his own way the urgency of particular shortfalls. But the comparison of how many had been booked with how many should have been booked is just the kind of thing a computer could do, provided it had a way of calculating how many should have been booked and an up-to-date picture of the real state of bookings. Its conclusions could be summarised as a number expressing the value of one more booking in a particular branch for a particular entry date. High values would indicate to the CA that he should relax his selection standards because time is getting short and there is a high probability that new entry training places will be wasted.

The value of an opening is correlated with the probability of wasting a branch opening. Its computation in CAPS is discussed after its practical application in personal selection.

It should be noted in passing that the value attached to a branch vacancy for a particular entry date is used by CAPS to advise the recruiter on the wisdom of relaxing or tightening his selection criteria for a particular opening, subject to minimum standards, but always with the aim of filling every vacancy. This is a significant change in selection policy.

The value of Booking at the Right time

6. The CAPS system is trying to put the best people into the new entry training openings but not at the expense of wasting any of the places. Regional monthly quotas run this risk as do unrealistically high recruiting standards. Low standards allow branch entries to be of diluted quality sometimes to the exclusion of good applicants. One simple adjustment is to publish the value of an entrant for each branch for each entry week computed by comparing the number who should have been booked by that time with the number actually in the pipeline.

Such a system would indicate the degree of shortage and the urgency of doing something about it. The CA could then choose an entry date which was convenient for the applicant and which had a high value. Once a booking had been made, the value for the same opening would decrease, to zero if all places had been booked.

Of course it might not be possible for the applicant to enter on the date suggested by the very highest value, but insofar as the CA's tended to book high valued dates, so the probabilities of wasting new entry places would be lowered. Early overbooking could be controlled by temporarily closing branch entries that had attracted more bookings than targetted that far in advance. As time passed the urgency to complete the class list would increase and the particular opening would allow an overbooking margin based on the numbers likely to fail to join after final approval.

The Value of Booking the Right Person

7. The application of the value of an opening to the counselling of an applicant, as thus far described, takes no account of the requirement, not only to fill places, but to fill them with the best people available. And, furthermore, as the discussion of branch recruiting policies emphasises, it is not in the interests of the Service to fill one branch with talented applicants regardless of the requirements of the other branches.

It is here that the talent matrix (figure 1) which is used in declaring branch recruiting policies finds its key application. Each enquirer who has taken the Recruit Test and been interviewed can be placed in 4 x 5 matrix according to his or her RT and PQ scores. Multiplying the percentages of the branch policy by class size shows the number of persons in each RT and PQ position that ought to be recruited to ensure that each entry class is of comparable RT and PQ make-up and that each branch gets its fair and agreed share of the talent expected to join in the year ahead. Figure 1 illustrates this computation for one seaman entry class of 26 men.

Figure 1 Make-up of a Seaman class (26 men)

MATRIX A (%)					MATRIX B (men)				
0.2	0.4	0.8	0.4	0.2	0.5	0.10	0.20	0.10	0.05
3.0	5.9	11.8	5.9	3.0	0.78	1.53	3.07	1.53	0.78
6.3	12.6	25.2	12.6	6.3	1.64	3.28	6.55	3.28	1.64

(275)

Matrix B of Figure 8 shows the ideal make-up for the Seaman Entry for a particular date. It is a target only and hence the decimal fractional men can be tolerated. The Jan 80 version of CAPS keeps the branch recruiting policies constant which assumes that the distribution of RT and PQ scores in the entry population is going to remain stable throughout the year. This is not likely to be valid but there is no obviously better assumption to make until the CAPS system has been running long enough to gather its own performance data. Its objective is to extract the best talent from the applicant population. If it is more successful than the present procedures the branch recruiting policies will be able to call for higher proportions of better quality entrants.

The Value of Booking the Right Person at the Right Time

8. Previous sections discussed attributing a value to an opening which depended on the probability of wasting a new entry training place for that branch on that date. Distant entries attracted lower values because there was more time left to recruit. Large entry numbers pushed up the probability of unfilled places. Bookings decremented the value of an opening until it reached zero when the number of recruits who had been booked by that time equalled the number who should have been.

Manpower quality considerations are brought in by computing the number who should have been booked, not simply in terms of places, but as places ideally intended for entrants of a particular position on the RT-PQ matrix. This would then be in accord with the agreed set of branch recruiting policies. Thus it could happen, for example, that a branch was well booked for middle quality entrants but low on Row R1 of the matrix. Any R1 applicant prepared to join on the date in question would then be seen as more highly valued than other applicants for this branch.

In summary, the value of an opening depends on the RT-PQ position of the enquirer, on the needs of a branch for persons of that quality, and on the time left in which to fill such places as are still vacant.

(224)

DRISKILL, Dr. Walter E., and MITCHELL, Lt. Col. J.L., USAF Occupational Analysis Program USAFOMC Randolph, AFB Texas.

THE USAF OCCUPATIONAL ANALYSIS PROGRAM: AN EVOLVING TECHNOLOGY (Thu P.M.)

The recent history of the USAF Occupational Analysis Program has been one of steady progress. We have been learning from management inspections that our continuing analysis of enlistment specialties can be improved by tailoring occupational data products to the users needs. The development of Utilization and Training Workshops for each specialty where major command, Air Force Manpower and Personnel Center, air staff functional managers, training staff, training development representatives, and occupational analysts meet to discuss the Air Force use of personnel and negotiate training requirements is a major innovation in the use of occupational data and in the Instructional Systems Development (ISD) model. The Air Force specialty training standard has taken on new meaning as a "contract" between Air Force users and the training community. Such workshops evolved out of the Hasty Grad program initiated by the ATC Commander. The result has been an improved training decisions system, and research is now underway to further develop the system. In certain areas, particularly where contingency operations are involved, our standard occupational survey technology does not let us directly assess the training needs; in the security police area, it was necessary to develop a scenario approach using an inventory of tactics knowledges, and equipment to determine what should be included in a new Air Base Ground Defense course. Over 900 security police officers and senior NCOs rated the items in the inventory and correlational analysis revealed very high agreement between groups on what should be trained. Thus, scenario-based analysis is a new technique which can help us deal with future contingency operations. We are currently also adapting the Training Emphasis technology for use in other specialized applications; it is presently in the field in one Electronics Principles Inventory and was also used in a fashion in the Officer Professional Military Education project which is now near completion. While originally designed to assess first term training requirements, we believe that the Training Emphasis methodology can be used in meeting a number of special needs. We anticipate that our occupational analysis technology will continue to evolve to help in meeting Air force management needs in the 1980s.

THE AIR FORCE OCCUPATIONAL ANALYSIS PROGRAM - A CHANGING TECHNOLOGY

Walter E Driskill, Ph.D., Jimmy L Mitchell, Ph.D.,
Joseph S. Tartell

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

As occupational analysts, one of the criticisms we hear most from managers and trainers is that job analysis "doesn't tell us what people should be doing." Fortunately in the US Air Force occupational analysis program we now have the techniques to counter such criticism, be they legitimate or not-so-legitimate.

Several arguments may prevail with the criticism levied by users of job data. After having analyzed the jobs of over 800,000 enlisted job incumbents in 300 plus occupations as well as over a third of the officer force, we are convinced that such criticisms primarily stem from three sources.

First, in many instances (a vast majority, in fact), the "what is" of the work environment, as revealed by job analysis, is the real world - and contentions about "what people should be doing" likely to be spurious. We are all aware of the differences that exist between stated policies and the implementation of those policies, and when the discrepancies are investigated, many times policy is at odds with the real world.

A case in point occurred several years ago. We were surveying the fire protection occupation. During our interviews with firefighters we identified a number of tasks relating to the maintenance of runway barriers. When the trainers and staff managers of the occupation saw these tasks, they wanted to purge them from the fire protection task inventory. They stated categorically that runway barrier maintenance was a function of the power production occupation, and that directives were in effect to prohibit fire protection personnel from such activity.

Even though the trainers and staff managers were adamant that no fire protection personnel performed any runway barrier maintenance, we left the runway barrier maintenance tasks in the fire protection inventory, following our usual policy of not eliminating tasks based on contentions like those that were made. Instead, we let the results of the job analysis give us the answer.

After administering the job inventory to the fire protection personnel, the analysis showed that over 35 percent of the journeymen level fire protection specialists were maintaining runway barriers. A follow-up inquiry of base commanders and their subordinate representatives for airfield operations revealed that, in fact, the commanders were using fire protection personnel for barrier maintenance. The reason they gave was that the power production specialists, who were by policy and directive responsible for those systems, did not perform night shift operations. When a runway barrier became inoperative in the middle of the night or there was a wind direction change which necessitated a relocation of the barrier, the quickest and simplest remedy was to turn to the fire protection specialists. They are always on the flightline, are mobile, and could make the necessary repairs or change locations long before a power production specialist could be routed from his bed to take care of the work. As a result of this finding, minor runway barrier maintenance responsibilities were added to the fire protection specialty. Thus, the policy guidance was brought into line with the real world.

The second source of confusion between "what is" and "what ought to be" sometimes lies in obscure management or training problems. In a survey of the Air Force environmental health occupation a few years ago, we found that none of the personnel from apprentice through supervisor were performing any of the new tasks associated with the protection of the environment that had been brought about by environmental protection legislation. Since our resident training schools were teaching these new tasks, we conducted a follow-on investigation to find out why these tasks were not being performed. The results showed that while the more recent graduates of the technical training knew how to perform the tasks, their supervisors and managers did not permit them to do so. The reason: the supervisors and managers and more experienced personnel had not received training and did not know how the new tasks were to be performed. Thus, the new tasks had not been implemented by operational units. Obviously, this problem was one of management and the training of the managers and supervisors.

The third source is the inconsistency of perceptions of what "ought to be," brought about by the differing levels of management involvement of the individuals concerned with an occupation, and brought about by the changing nature of the work environment. Different management levels have different perspectives about what personnel in an occupation should be doing. At the very top level of management, because of their broader scope of responsibility, we find personnel who want every worker to be able to perform every task in the occupation. At the very bottom level, that is, at the first line supervisor level who is responsible for getting the job done, the expectancy is that people will do only that which is required in their jobs. They usually are less concerned with breadth of ability and more concerned with specialized ability. Depending upon the number of levels between the top and bottom, you can have very different perceptions about what the incumbents of an occupation should do. Once again this is a management problem. The issue is that there is no average job in an occupation - that there really are many different jobs, and unfortunately, as Driskill and Mitchell (1979) pointed out in the paper to the Military Testing Association the separate jobs are frequently viewed as error variance. In reality, it is impossible to talk about what people should be doing from any particular perspective - level of management involvement. An occupation must be viewed from the perspective that there are many different jobs which emerge because of situational factors.

The changing nature of the work environment presents an unique problem and is brought about by the introduction of new equipment, new concepts, or new factors which cause or require the forecasting of what will be in the future. This is an extremely difficult problem, for there is no experience upon which to draw conclusions about what the jobs will be. Instead, it is a matter of forecasting. We have been able in the Air Force occupational analysis program to do some forecasting and believe that we have a handle on ways of tackling the problem.

We have taken two approaches to the problem. The first of these is in the use of utilization and training conferences which convene periodically after the completion of an occupational survey. In these conferences staff management personnel, operational personnel from the major using commands, personnel classification and assignment representatives, trainers, and occupational analysts meet in a workshop. There, with job analysis data from the real world, these representatives can work out the problems of management, of utilization, and of training. In more than 100 of these workshops over the last two years, we have found that the problems of differing concepts of what jobs should be, as well as differences between what people should be doing and what they are doing, are usually worked out satisfactorily. To some extent also, the question of forecasting changes in jobs can be addressed

on the basis of collective experience of those present at the workshop. With air staff representatives and major command representatives present, future equipment procurement and shifts in management policy can be discussed and the impact of such changes on training and utilization can be ascertained. These utilization and training workshops represent a new and more effective way of dealing with the management of specialties. In the Air Force, we feel that they have been tremendously successful, and we view such workshops as a major innovation in Instructional Systems Development and personnel management.

Another approach we are using to address the question of forecasting is an adaptation of the training emphasis technology used in the Air Force occupational analysis program which was reported by Ruck, Thompson, and Thomson (1978) at an earlier Military Testing Association conference. In this methodology we use the job inventory for an occupation as a basis for collecting opinions from senior supervisors in the occupation on two elements. The first element is an identification of the tasks within the job inventory for the occupation that require formal or structured training prior to an airman's entry into the work or occupation; and the second element is the ratings of these supervisors of the emphasis that should be provided in training on each of the tasks that they selected for training. They use a nine point scale to make these ratings.

Our first application of the adaptation came in a study to determine training requirements for security policemen in Air Base Ground Defense Tactics (Mitchell, 1980). During the process of conducting an occupational survey of the Security Police occupational field, questions arose about initial and follow-on training for personnel associated with Security Police Element for Contingencies (SPECS) concept for base defense. At the time of the survey this program was to be replaced with a new Air Base Ground Defense (ABGD) approach in which the Air Force assumed complete responsibility for the protection of Air Force bases.

The evolution of the Air Base Ground Defense concept caused immediate recognition of the need for specialized training in the area of defensive combat skills. The challenge facing the staff of the Security Police Academy was to identify, from the domain of all possible combat skills, those skills and knowledges most directly applicable to Air Force air base defense, and to develop an appropriate course of instruction which would meet the defined training requirement and still be within the constraints of an austere training budget. In agreeing to provide some assistance in defining the ABGD training requirement, the staff of the USAF Occupational Measurement Center recognized that the normal occupational analysis methodology, aimed at the quantified description of the tasks personnel perform in their jobs, would not be applicable. The ABGD concept requires training for contingency operations; that is, for what might happen rather than for what tasks Security personnel perform in their day-to-day peacetime jobs. Thus, the normal "Task List" approach would not meet the objective of this project.

One obvious approach was to define the possibilities which might occur and ask senior Security Police personnel what training should be given. This approach required a structured questionnaire, which defined the situation very precisely ("what might occur"), and structured the responses of the individuals surveyed so that the data could be easily summarized and compared. This approach would involve a "scenario" which had not previously been used in the USAF occupational analysis program. Thus, in this study, we were at the very edge of the "state of the art", and there was no previous history of research which would lead to the clear specification of decision criteria (such as exists for the regular occupational analysis program).

While it may have been possible to specify a series of scenarios and ask for ratings of what should be trained, such an approach would have generated more data than could have meaningfully been analyzed. In discussions with the Security Police Academy Commander, it was agreed to limit this project to a single scenario - the one which would be at the upper end of the spectrum of possible serious threats and yet would still be within the realm of a credible defensive response. In these discussions, it was further agreed that a list of the skills, knowledges, equipment, and weapons would be used to gather data rather than the usual task list. It was considered that such a list of skills, knowledges, and weapons would be more relevant in this case, and would be more directly interpretable for the design of ABGD training.

It was further agreed that the data gathered would consist of ratings by senior Security Police personnel of the relative emphasis which each item (skills, knowledges, weapons) should receive in an ABGD training course. The Air Force Human Resources Laboratory (Occupational and Manpower Research Division), Brooks AFB, Texas, developed a rating procedure to identify tasks (in a normal USAF Job Inventory) which should be given the greatest emphasis in initial training (defined as resident training or formal on-the-job training). This training emphasis research was conducted in 1978 and was released for operational use by the USAF Occupational Analysis Program in February 1979. This technology is currently being used in every occupational survey to gather data for technical training curriculum developers and training managers.

In the ABGD situation, where a hypothetical scenario was used, there had been no research to determine if Training Emphasis ratings could be gathered reliably and have some relevance or validity. In using the Training Emphasis technology in a new situation, the present project extended and modified the original purpose of the training emphasis research. For the ABGD project, the question asked of raters had to be more specific - they were asked what the relative emphasis should be on various skills, knowledges, equipment, weapons, etc. in a new Air Base Ground Defense Tactics course. This question was much more specific than is the case with usual training emphasis ratings; logically, if reliable and valid ratings can be achieved in the more general training emphasis areas, then it should be highly likely that ratings in a more specific, clearly defined situation (dealing with only one course) would also be reliable and potentially useful.

The scenario and inventory of ABGD tactics and equipment were developed during September and early October 1978. An initial list was developed by the Combat Skills staff at Camp Bullis, working with a team of occupational analysts. This list was reviewed and embellished in a conference session with the staff of the Security Police Academy at Lackland AFB. The scenario was developed by the senior staff of the Academy and was reviewed and approved by the staff of the Office of Security Police, Kirtland AFB NM, which is the highest Air Force management level for police.

The ABGD Tactics Inventory was validated in a two-step process. First, approximately 50 senior Security Police representatives at the October 1978 World Wide Security Police Symposium reviewed the list in detail and critiqued its approach and content. In addition, major command representatives took copies of the instrument back to their commands for further analysis and review. MAJCOM comments were funneled back through the Office of Security Police (AFOSP) at Kirtland AFB, to insure the differences of opinion could be resolved, and to demonstrate that this was an official AFOSP project. Command comments were analyzed and forwarded to the USAF Occupational Measurement Center for use in developing the final survey instrument. A final version of the ABGD Inventory was constructed in January and February 1979.

During April 1979, 1,100 copies of the ABGD Inventory were mailed to Security Police units worldwide. A letter of instructions was included which emphasized the importance of the project and requested an immediate and complete response. Brigadier General William R. Brooksher, the Chief of Security Police (AFOSP, Kirtland AFB NM) signed the letter, which may account for the exceptional level and speed of response. One thousand and twenty-three completed booklets (a 93 percent return rate) was received by the USAF Occupational Measurement Center within 37 days of mailing, when the administration was closed and data entered into the computer. This initial data base was subjected to a number of quality control "clean up" procedures before being entered into the primary analysis programs.

Sampling in a special study, such as the ABGD project, is a critical part of the study. In this case, one of the basic issues to be addressed was whether there was a general consensus as to what should be given emphasis in ABGD training. In order to define whether there were such a consensus, it was necessary to define a variety of groups of personnel who might have differing opinions as to what should be trained. If there was general agreement among such groups, then it would be possible to assert that the underlying rating policy had been tapped, and that the resulting ratings could be used in decisionmaking. If there was no general agreement, then it might still be possible to identify groups likely to have a realistic opinion as to what should be trained (e.g., those with knowledge of the war plan, those with combat experience, etc.) and use their ratings for making decisions as to the ABGD curriculum.

In discussion with the Security Police Academy Commander, it was decided to include a variety of groups so that the degree of their agreement could be assessed. Raters assigned to the air staff or on major command staffs would be used as one group, since they could be expected to have the greatest knowledge of security contingency plans. All Security Police officers above the grade of first lieutenant were included to insure that the overall sample had some officers with combat experience, such as in Southeast Asia. Background questions were included to insure that those with combat experience could be identified and contrasted with those who had no combat experience. It was also considered important to include some way to identify officers and NCOs who were part of the training establishment to ascertain if their opinions were in agreement with operational personnel.

The opinions of flight commanders might be different from those who are staff personnel. Thus, all first lieutenants and captains with at least 18 months experience as shift commanders were also included. In addition, a random sample of 200 senior enlisted managers were included, as well as a random sample of 150 fully qualified technicians in the security or law enforcement field. Finally, it might be considered that those who had received "Safeside" training, which was specialized training in perimeter defense, might also represent a different perspective. It proved difficult to identify personnel with Safeside experience and it was necessary to ask for nominations from each major command of personnel with Safeside experience. All commands cooperated in this effort and provided nominations by immediate return message.

A total of 1,023 booklets were returned and processed. The rater sample represented a very significant proportion of the senior leadership of the Security Police career field. Using the reliability program (REXALL) of the Comprehensive Occupational Data Analysis Programs (CODAP), 906 cases were processed. These 906 cases were found to be in substantial agreement as to the recommended training emphasis for the items in the ABGD inventory. The interrater reliability (as assessed through components of variance of group means analysis) among these 906

raters was .996. Because the high interrater agreement might be attributed to the number of cases in the sample, we did further analysis in which we divided the 906 cases up into 12 subgroups. Interrater agreement within each subgroup exceeded .90. We then correlated the responses of the individuals in the groups with those in each of the other groups, as well as the ratings of the total sample. The 13 intercorrelations range from a low of .90 (rounded) to .99. The N for the 12 subgroups ranged from 54 to 400.

These correlations indicated a very high agreement among Security Policemen who are knowledgeable of Air Base Ground Defense tactics. As a result, the Security Police Academy was able to determine the curriculum for ABGD training that is now given at Camp Bullis, Texas.

We have extended our methodology in an extensive study of the leadership, management, and communication tasks officers perform. This project is to provide data for the development of Professional Military Education (PME). In addition to collecting task performance data in the usual format, we also collected officer judgments about the emphasis that should be placed on various topics now taught in PME courses. While these data are not yet reported, preliminary analysis indicates that the extension of the technology has again produced reliable results.

In summary, then, occupational analysts, at least in the Air Force program, are in a position to answer criticisms about "what should be." It should never be forgotten that in the face of such criticism, the starting point is always what the real world of work is. Assessments of "what should be" then can be made, and the methodology exists to address the question of "what should be" when it is legitimately different than "what is."

References

Mitchell, J.L., and Driskill, W.E. Variance Within Occupational Fields: Jobs Analysis Versus Occupational Analysis. Paper presented at the 21st Annual Conference of the Military Testing Association, San Diego, Calif, 1979

Mitchell, J.L. A Scenario Analysis of Air Base Ground Defense Training Requirements. Presentation at the Third Annual International Workshop on Occupational Analysis, Randolph AFB TX, 1980

Ruck, H.W., Thompson, N.A., and Thomson, D.C. The Collection and Prediction of Training Emphasis Ratings for Curriculum Development. Paper presented at the 20th Annual Conference of the Military Testing Association, Oklahoma City, Okla, 1978

DRUCKER, Eugene H., Human Resources Research Organization, Fort Knox Kentucky, & EATON, Newell K., U.S. Army Research Institute for the Behavioral and Social Sciences.

CONSISTENCY OF UNIT PERFORMANCE RATINGS BY ARMOR OFFICERS AND NCOS
(Thu P.M.)

Although the operation of a tank requires coordination among its crewmembers, tank crews are rarely fully-manned. Both training and maintenance suffer as a consequence. In 1977 the Tank Forces Management Group recommended as a solution that additional crewmen be assigned to tank units and that the effects of their assignment be evaluated. As part of the evaluation, additional tank crewmen were assigned to 12 different battalions in USAREUR. Their effects on unit performance were determined by comparing the effectiveness of these 12 battalions with that of 6 battalions to which no additional crewmen were assigned. Since unit performance tests could not be administered, the comparisons were made using officer and NCO ratings. While these data produced the expected differences, response biases favoring unit augmentation were considered possible. The internal consistency of the data was examined to determine if such biases occurred, and if so, to determine the extent of their influence on the results.

CONSISTENCY OF UNIT PERFORMANCE RATINGS BY ARMOR OFFICERS AND NCOS

Eugene H. Drucker

Human Resources Research Organization

and

Newell K. Eaton

U.S. Army Research Institute for the Behavioral and Social Sciences

INTRODUCTION

The TO&E for armor units specifies that the 17 tanks in a tank company be manned by 68 crewmen, which is equivalent to 4 crewmen per tank. While this number of crewmen should enable each unit to fully man its tanks, the 17 tanks are, in fact, rarely fully manned. This is due in part to absenses resulting from leave, sick-call, in-processing, out-processing, and assignment to details. In part it is due to the assignment of tank crewmen to special duty, and in some cases to the failure to assign the full complement of 68 men to a tank company.

The successful operation of a tank, however, requires the coordination of the full four-man crew. Many duties performed during training involve close interaction between and among the crewmen in all four positions. These duties cannot be taught or practiced in the absense of one or more members of the crew.

The shortage of armor crewmen can also interfere with tank maintenance. Routine maintenance is frequently delayed because there are not enough men available to complete the required number of maintenance tasks.

In 1977, the Tank Forces Management Group issued a report dealing with the shortage of authorized tank crewmen in armor units.[1] The report states that the loss in the effectiveness of a Tank Weapon System resulting from the absense of just one crewman is greater than 50 percent. The report also notes that the full complement of four crewmen does not provide for replacements of crewmen lost in combat, and that it does not provide sufficient manpower for the efficient operation or servicing of the equipment. Consequently, the group recommended that the number of crewmen in armor units be increased by one crewman per tank and that the effects of this increase be assessed.

In response to these recommendations, the Department of the Army ordered that a test be conducted of the augmented tank crew or"fifth crewmen" concept. The purpose of the test would be to determine how the assignment of additional crewmen would affect the quality of training and maintenance in armor units, to determine how their assignment would affect the operational capabilities of these units, and to assess

[1]Kalergis, J.G. Tank Forces Management Group Study Report, Fort Knox, Kentucky: Office of Armor Force Management (OAFM), April, 1977.

the severity of any problems that the additional crewmen would cause in the areas of administrative reporting, transportation, logistics, and command control.

Since the test of the fifth crewman concept was to be conducted in USAREUR using TO&E armor units, an important restriction was imposed upon the test design. No data collection procedures could be used that would interfere with the combat readiness of the test units. This meant, in effect, that objective test procedures could not be used to measure the effects of unit augmentation. The assessment would be limited, instead, to information that could be collected through the use of questionnaires or interviews. As a consequence, the results of the test were susceptible to the types of biases that are present in any subjective evaluation. It is the purpose of this paper to describe the test of the "fifth crewman" concept, to summarize the major findings, and to discuss the effects of the use of subjective data.

METHOD

Eighteen armor battalions in USAREUR participated in the "fifth crewman" test. Twelve of the battalions were each augmented in strength with 54 additional tank crewmen, one crewman for each tank in an armor battalion. The remaining six battalions filled to 100% of strength authorized by the TO&E, but were not augmented beyond this level.

The number of additional tank crewmen assigned to the test battalions and the organization level to which they were assigned are summarized in Table 1.

TABLE 1

Number of Additional Crewmen Assigned to Test Battalions and Organizational Level of Assignment

Number of Additional Tank Crewmen Per Battalion	Organizational Level of Assignment	Number of Battalions
54	Company	6
54	Platoon	6
0	N/A	6

(281)

In six of the twelve augmented battalions, the additional tank crewmen were to be assigned by the battalion to company headquarters for control and administration. In the other six augmented battalions, the additional crewmen were to be assigned by the battalions to platoons. The purpose for assigning some of the additional crewmen to company headquarters and others to platoons was to compare the effects of these two levels of assignment. It was observed during the survey, however, that most of the additional crewmen were assigned to platoons regardless of the intended level. Consequently, any differences found between battalions augmented at company level and those augmented at platoon level can probably be attributed to chance.

The six nonaugmented battalions were baseline or control battalions to assess the effects of augmentation.

To determine the impact of the additional tank crewmen on unit performance, questionnaire and interview items were prepared pertaining to various test issues. These items were administered to battalion commanders and representative samples of subordinate leaders and crewmen. The data were collected twice in order to assess the effects of augmentation over time. The additional tank crewmen were scheduled to be assigned to augmented units by December, 1978, and the first data collection phase occurred during January, February, and March, 1979. The second data collection phase occurred during April, May, and June of 1979.

All battalion commanders, company commanders, first sergeants, battalion S1s, battalion S4s, and battalion maintenance officers in the test units participated in the survey; the other participants included samples of platoon leaders, platoon sergeants, and tank crewmen from these battalions. A total of 540 respondents participated during the first data collection phase, while 519 respondents participated during the second phase.

None of the questionnaire items required a respondent to directly judge the effects of the additional crewmen on unit performance. Instead, each question was designed to assess the perceived quality of performance within the unit to which the respondent was assigned. The impact of the additional tank crewmen on unit performance would be assessed by comparing the responses obtained from the two types of augmented battalions with those obtained from the nonaugmented battalions.

The questionnaires contained two types of questions--primary and secondary. Primary questions were written in general terms to obtain an overall judgment pertaining to an issue, while secondary questions were written in more specific terms to provide more detailed information. For example, the primary question dealing with the quality of maintenance was, "How would you describe the overall quality of the maintenance performed on tanks by the crews in your battalion?" This was followed by a seven-point rating scale with responses ranging from "extremely good" to "extremely bad." There were four secondary questions pertaining to the quality of maintenance. One was, "How would you describe the quality of the quarterly services performed on the tanks in your battalion?" Another was, "How adequate is the number of tank crewmen that are available in your battalion for performing day-to-day crew maintenance on tanks?" Only the

responses to the primary questions were analyzed statistically. The responses to the secondary questions were used primarily to help interpret the results obtained on the primary question.

Items were scored by assigning numerical values to each response alternative. In all cases, high values are favorable and low values are unfavorable. The most unfavorable response was always assigned a value of 1.0; the most favorable response was always assigned a value equal to the number of response alternatives contained within the item.

Comparisons between augmented and nonaugmented battalions on the primary questions were made using fixed-model analyses of variance. Whenever a significant main effect was obtained for Battalion Type, a contrast was made between augmented and nonaugmented battalions. Whenever a significant interaction occurred between Battalion Type and Data Collection Phase, simple effects analyses were conducted for each data collection phase.

RESULTS

Ratings of unit performance were obtained for training, maintenance, sustained operations, and unit readiness. The mean ratings are presented in Table 2.

TABLE 2

Mean Unit Performance Ratings by Battalion Type

Type of Unit Performance	Battalion Type			
	Augmented Company Level	Augmented Platoon Level	Nonaugmented	P
Training	5.75	5.88	5.21	<.01
Maintenance				
Phase 1	5.82	5.46	5.36	NS
Phase 2	5.85	6.08	5.44	<.01
Sustained Operations	6.25	6.33	6.00	.051
Unit Readiness	5.97	5.99	5.61	<.05

Looking at the results for training, the ratings were higher in the two types of augmented battalions than in the nonaugmented battalions. A significant main effect was obtained for Battalion Type, and a contrast between augmented and nonaugmented battalions showed that the overall quality of training was rated significantly higher in augmented battalions than in nonaugmented battalions.

Since the analysis of variance conducted on the maintenance data yielded a significant interaction between Battalion Type and Data Collection Phase, the data are presented separately for each of the two phases. Simple effects were conducted for each data collection phase separately. The simple effect for Battalion Type was significant during the second data collection phase, but not during the first. A contrast between augmented and nonaugmented battalions during Phase 2 showed that the quality of maintenance was rated significantly higher in the augmented battalions than in the nonaugmented battalions.

Looking at the rated capability to perform adequately during a 72 hour sustained operation, the ratings were again higher in the two augmented types of battalions than in the nonaugmented battalions. The main effect for Battalion Type, however, was only of borderline significance. A contrast between augmented and nonaugmented battalions showed that augmented battalions were rated significantly more able to perform adequately during a sustained operation.

Finally, the ratings of combat readiness were also higher in augmented than in nonaugmented battalions. The main effect for Battalion Type was significant at the .05 level, and a contrast between augmented and nonaugmented battalions showed that the augmented battalions were rated significantly higher in combat readiness than the nonaugmented battalions.

The next table shows data pertaining to the types of problems that were reported as a result of assigning additional tank crewmen to battalions.

TABLE 3

Mean Problem Severity Ratings by Battalion Type*

Type of Problem	Battalion Type			
	Augmented Company Level	Augmented Platoon Level	Nonaugmented	P
Administrative Reporting	5.30	5.14	5.07	NS
Transportation	4.72	4.25	4.17	NS
Logistics	5.69	5.51	4.99	<.01
Command & Control	4.27	4.31	4.05	NS

* High values imply low severity.

There were no significant differences among the battalion types in the reported severity of administrative reporting problems, transportation problems or command and control problems. However, there was a significant main effect for battalion type for the reported severity of logistics problems. Surprisingly, logistics problems were reported to be more serious in the nonaugmented battalions than in the augmented battalions. A contrast between the two types of battalions showed that this difference was significant at the .01 level.

DISCUSSION

The results obtained during the test of the "fifth crewman" concept suggest that the assignment of additional tank crewmen to armor units not only improves unit performance, but that it does so without causing problems for the units to which the additional crewmen are assigned. But are these valid results, or should they be attributed primarily to the use of subjective data? The respondents in both the augmented and nonaugmented battalions were told that they were participating in a test of the "fifth crewman" concept. It could certainly be assumed that they would have a vested interest in the outcome of the survey since they would have liked the additional tank crewmen to be permanently assigned to their units.

The best way to determine whether or not the ratings were valid indicators of unit performance would have been to correlate them with objective performance measures. But performance tests could not be given due to the restrictions described earlier. Moreover, the validity of the items could not be established on other armor units since there would have been less reason in nonparticipating units to exaggerate their level of unit performance or to understate the severity of problems that were being experienced in these units.

The strongest evidence for the existence of a response bias occurred in the data dealing with the severity of logistics problems. These problems were described as being significantly more severe in nonaugmented battalions than in augmented battalions. Since there were 25% more crewmen in the augmented battalions, logistics problems should have been more severe in the augmented battalions instead.

If there were either conscious or unconscious efforts to conceal the existence of logistics problems in augmented units, the same efforts should have been made on both the primary and secondary questions dealing with logistics. After all, the respondents did not know which were the primary questions nor even that they were primary and secondary questions.

Table 4 contains the mean responses on the secondary questions dealing with logistics.

The data seem to suggest that some logistics problems were described as being more serious in nonaugmented battalions; this corresponds to the results on the primary question. In particular, respondents in nonaugmented battalions reported a more serious shortage of TA-50, a more serious insufficiency in mess facilities, and a more serious shortage of storage facilities. On the other hand, respondents in augmented battalions appear to report a more serious shortage of station property, and no clear differences appeared in the reported seriousness of the shortage of troop billeting or in the seriousness of inadequate installation support. The fact that there were these differences suggests that the data could be valid.

But how could any logistics problem be more severe in nonaugmented battalions? The most likely answer to this question is that a greater effort was made to solve these types of problems in the augmented units. It should not be forgotten that additional tank crewmen were also assigned to the control battalions in order to bring them up to 100% strength. If additional support were to be given to the test battalions, it would

TABLE 4

Mean Logistics Ratings on Secondary Questions
by Battalion Type

Secondary Question	Battalion Type		
	Augmented Company Level	Augmented Platoon Level	Nonaugmented
Shortage of Troop Billeting	3.12	2.65	2.62
Shortage of TA-50	3.71	3.47	3.30
Shortage of Station Property	3.91	3.22	4.04
Insufficient Mess Facilities	4.27	4.57	4.10
Inadequate Installation Support	3.29	3.41	3.32
Shortage of Storage Facilities	3.67	3.53	2.82

probably be given to the augmented test battalions since these battalions experienced the greater increase in personnel strength. With this additional support, however, logistics problems in augmented units could be less severe despite the greater number of additional crewmen. In other words, both augmented and nonaugmented battalions should have experienced serious logistics problems, but these problems were more likely to be solved in the augmented battalions. Thus, the results obtained on the primary question do not necessarily imply that the responses were biased.

Another source of possible bias that may have affected the results of the "fifth crewman" test is the halo effect. This would manifest itself by a tendency to make similar judgments across a number of different scales regardless of the aspect of unit performance that was being judged.

Reexamining the ratings of unit performance presented in Table 2, it appears that the ratings given to these different aspects of unit performance were in fact quite similar. Before infering the existence of the halo effect, however, other data need to be examined also. Table 3 contained mean ratings of the different types of problems that were experienced in the test units. Reexamining these ratings, it is apparent that they did differ from problem to problem. Since the primary question dealing with command and control problems had a five-point rather than a seven-point response scale, it should not be included in the comparison. Nevertheless, in all three types of battalions, transportation problems were described as being more serious than those pertaining to administrative reporting or logistics. Next, the ratings made on the secondary

items dealing with logistics can be reexamined. These appear in Table 4. A five-point rating scale was used for all of the items. Once again the data suggest that the respondents made independent judgments across items. For all three types of battalions, the shortage of troop billeting was described as being the most serious problem, while insufficient mess facilities was described as being the least serious problem.

Table 5 presents the mean ratings on the secondary items dealing with combat readiness.

TABLE 5

Mean Combat Readiness Ratings on Secondary Questions by Battalion Type

Secondary Question	Battalion Type: Augmented Company Level	Augmented Platoon Level	Nonaugmented
Amount of MTOE Equipment	6.10	6.26	6.00
Condition of MTOE Equipment	5.83	6.07	5.50
Number of Tank Crewmen	5.58	5.55	3.29
Number of Tank Crewmen Adequately Trained	4.41	4.47	4.04
Organic Logistical Support	4.93	5.12	4.70

The respondents, once again seem to have responded differently to the different items. The number of tank crewmen who were adequately trained was described as being less adequate than the amount or condition of the MTOE equipment. Also, it is important to notice that the number of tank crewmen was described as being an especially serious problem in the non-augmented battalions, as would be expected.

While these data cannot prove the absense of a halo effect, the fact, that the respondents did respond differently to the different items suggests that its influence was small. And while it might be assumed that the halo effect operated primarily on those items on which the respondents lacked sufficient information, the process by which respondents were selected should have precluded this possibility.

Next, it should be noted that there may have been a tendency by respondents in all types of battalions to describe their units favorably in case they themselves were being evaluated. The respondents were told

that the purpose of the survey was to test the "fifth crewman" concept and not to evaluate their performance, but there is no assurance that this was believed. However, if there were such a tendency to respond favorably, this tendency should have appeared in both the augmented and nonaugmented units. If so, this source of bias should not have affected the observed differences between the two types of battalions. Moreover, it should not be assumed that the respondents described their units more favorably than was warranted. Many of the officers and NCOs took advantage of the survey to express dissatisfaction with their units.

In summary, while the data were obviously susceptible to the types of bias described in this paper, and to other types of bias that were obviously present, the existing evidence suggests that the assignment of additional tank crewmen to armor units did result in improved unit performance and caused few, if any, serious problems.

DUNCAN, R.E., USAF Occupational Measurement Centre, Randolph AFB, Texas.

AN APPROPRIATE NUMBER OF MULTIPLE CHOICE ITEM ALTERNATIVES: SWANSON (1976) REVISITED (Wed A.M.)

Swanson (1976) has made claims concerning the preferability of three versus four alternative multiple-choice test items. Reanalysis of data presented by Swanson reveals possible discrepancies in procedure, data interpretation, and conclusions. Reanalysis showed no significant statistical nor practical differences between three and four alternative items. Data seems to indicate continued use of the four alternative item method when classical test theory is employed.

An Appropriate Number of Multiple-Choice Item Alternatives: Swanson (1976) Revisited[1]

R. Eric Duncan

USAF Occupational Measurement Center
Randolph AFB, TX 78148

There may have been a misapplication of test theory when discussing the pros and cons of three-alternative test questions versus the conventional four-alternative test question format. Swanson (1976) conducted a study to verify the results obtained by Tversky (1964) and Grier (1975). Grier made claims of maximizing "(a) expected test reliability, (b) the power of a test, (c) the discrimination capacity of the test, and (d) the uncertainty index" (Swanson, 1976). Swanson's results were similar.

This paper examines Swanson's procedures and results in order to determine (1) if they correspond to classical test theory, (2) if classical procedures would produce similar results, and (3) if there were significant and/or practical differences between three and four alternative test question data.

Method

Subjects and Instruments

Data accumulated, analyzed, and reported were obtained from Swanson's subjects (Swanson administered a multiple-choice end-of-course exam to 489 US Air Force students at the Air Force Academic Instructor School, Maxwell AFB, Alabama).

Procedure

Table and figure data were reexamined to determine significant and/or practical differences between three and four choice alternatives. Swanson's (1976) first figure was placed in tabular form to more easily evaluate differences (see table 1). Swanson's second table was analyzed to determine whether obtained reliability indices were similar when expanded or reduced using the Spearman-Brown formula to determine reliabilities of expanded or reduced test lengths. Reduced or expanded reliabilities were compared against obtained reliability values to determine possible discrepancies. Swanson's third and fourth tables were non-statistically examined to evaluate practical differences between three and four choice alternatives.

Results

The reliability indices of three and four alternative tests, presented in Swanson's first figure, were compared at each "c" level (an increase in gross number of items (n) to compensate for fewer alternatives (c), i.e. c=na). The results indicated that no obtained z exceeded the critical $.95z_{489}$ of 1.96, as can be seen in Table 1. Practical differences were also examined.

[1]The views expressed in this paper represent those of the author and do not necessarily reflect the views of the United States Air Force or the Department of Defense

Table 1

Tabular Form of Figure 1 as Presented in Swanson, 1976 and Grier, 1975

Number of Alternatives	Number of Questions	"c"	r_{xx}	z
3	40	120	.575	1.23
4	30	120	.520	
3	50	150	.655	1.04
4	37.5	150	.615	
3	66.67	200	.745	1.01
4	50	200	.715	
3	100	300	.830	.95
4	75	300	.810	
3	200	600	.915	.90
4	150	600	.905	

All z-values were nonsignificant

At the conventionally acceptable level for a meaningful reliability index (r_{xx}=.80) interpretation, there were no practical differences, even though three alternative test item reliabilities exceeded those of their four alternative counterparts.

Obtained reliabilities for each test (3 and 4 alternative) were examined after expansion or reduction with the Spearman-Brown formula as shown in Table 3.

Table 2

Summary of Test Reliabilities

Test	76-A	76-B	76-C	76-D
75	.82	.83	.81	.58
100	.86	.88	.85	.76

(From Swanson, 1976)

Reliabilities for each test and corresponding class appear in Swanson's second table (see Table 2). Rationale for expansion or reduction involves increased or decreased total test variance (obtained by increasing or decreasing the number of test items) and its affect on the unexpanded or non-reduced reliability values. Results of expansion or reduction showed that there were no significant differences between the obtained reliabilities and their corresponding expanded or reduced values.

Table 3

Reliabilities After Obtained Reliabilities were Expanded or Reduced Using Spearman-Brown Expansion Formula

Class	Test	Obtained r_{xx}	Expanded or Reduced r_{xx}
76-A	75	.82	E to .859
	100	.86	R to .822
76-B	75	.83	E to .867
	100	.88	R to .846
76-C	75	.81	E to .850
	100	.85	R to .810
76-D	75	.58	E to .648
	100	.76	R to .704

Swanson's third table (see Table 4) was examined only so far as the "average differentiation index" was concerned. There were no statistically significant differences among average differentiation indices for each class. No decision was reached about the practical differences of the differentiation indices due to a lack of individual item data.

Table 4

Test Data Summary Table

Class	Test Version	Mean Score	Standard Deviation	Average Ease Index	Average Differentiation Index
76-A	75	46.98	10.70	62.64	.28
(N=109)	100	64.51	11.66	64.49	.36
76-B	75	54.60	9.17	72.80	.31
(N=142)	100	70.49	12.94	70.49	.34
75-C	75	48.66	9.26	64.91	.32
(N=123)	100	65.20	12.12	65.10	.34
76-D	75	56.98	5.66	75.96	.20
(N=115)	100	76.60	8.49	76.58	.24

(From Swanson, 1976)

Since no control group was used in Swanson's study, comparisons of mean time differences to a control group were not made. This makes the results and conclusions questionable.

Discussion

After examination of these results, it seems possible that Swanson misapplied principles of classical test theory in his data interpretation. Further explanation will point-out these errors.

The first figure presented by Swanson tends to cause confusion. Reliability curves may be inappropriate in this instance, especially without an explanation that most reliabilities below .80 are not useful in establishing test validity. Reliabilities of .80 or greater tend to lend credibility to the assumption of random selection of items from a universe of items pertinent to the knowledge area. Since the greatest reliability differences appeared at levels below .60, meaningful differences were not found. The data served to confuse rather than support the hypothesis of a better test question possessing three instead of four alternatives.

Swanson failed to examine the results obtained when computed reliabilities are expanded or reduced, using Spearman-Brown, to the number of items in the test to which each is being compared. The results show that no significant differences between obtained and expanded or reduced values were found. In fact, all classes except 76-D possessed almost identical values when obtained values were either expanded or reduced. This suggests that the 100, three-alternative-item test could have been reduced to 75 items and not lose any true variance in the process. This would theoretically have saved more time in construction and administration. The test with 75, four-alternative items could have been expanded to incorporate more similar material and not change the reliability index. In class 76-D, the reliabilities obtained suggest test problems not correctable with expansion or reduction. Swanson's suggestion of better reliabilities with the test composed of three alternative items does not appear to be warranted by the data.

One claim Swanson made was that of better discriminatory power with use of three alternative test. Even though average differentiation indices were larger for the test with three alternative items, these indices were not significantly different. Without item data, no "better than" statements should have been made. A non-parametric test, which could have shown a significant difference in the number of items with acceptable versus unacceptable differentiation capabilities, should have been conducted. A test of this type would have answered the question in a much better fashion than by simple mean comparison.

With no baseline or time series data, time savings cannot be claimed. Swanson seems to believe that a long test (100 items) given in the same period of time as a test with fewer items (75) indicates time savings. This is not always the case. Under a proper design, three and four alternative item tests would have had both 75- and 100-item versions. When compared, the differing administration times would have yielded data sufficient to make "better than" conclusions.

It must be remembered that use of a test item format should never be limited to an "easier to construct" philosophy. If this were the case, true-false questions would suffice. Since this is not the case, frustration over inability to devise a "good" fourth alternative should not be a consideration of test constructors when choosing format. It has yet to be demonstrated that the three alternative item is any better than the four alternative item. If the same knowledge areas are tested, the four alternative item test is suggested. A change in present Air Force testing procedure is not warranted based on the data presented by Swanson.

BIBLIOGRAPHY

Grier, J. G. The number of alternatives for optimum test reliability. Journal of Educational Measurement, 1975, 12, 109-113.

Swanson, Ronald G. Multiple Choice Tests; How Many Alternatives? Published by the Academic Instructor School, Maxwell AFB, Alabama, 1976.

Tversky, A. On the optimal number of alternatives of a choice point. Journal of Mathematical Psychology, 1964, 1, 386-391.

EBENRETT, Heinz-Jurgen, German Armed Forces Psychological Services Research Institute, Bonn, West Germany.

INTELLECTUAL PERFORMANCE DEPENDING ON NEUROTICISM AND INTROVERSION (Thu P.M.)

The question: Personality variables, especially the dimensions of stability-neuroticism and extraversion-introversion, are said to show a proven value in mediating predictions and an important interaction potential in all types of incellectual functioning. According to the EYSENCKs psychophysiological theory extraverts are regarded as low in arousal, introverts as high, and it is further assumed that neurotic subjects (high scorers on N scales) are characterized by higher drive than stable subjects. There are thus four main groups of subjects which differ in order of their drive level: stable extraverts (SE); neurotic extraverts (NE) and stable introverts (SI); neurotic introverts (NI).

The question: It is believed that the optimum of drive level for complex and difficult tasks like those involved in intelligence tests lies below the high level reached by the NI subjects and above that reached by the SE subjects.

The method: 597 pupils of German secondary schools were tested in each of the four personality groups and in their performance of intelligence scales representing a number of seven primary intelligence factors.

The results: The findings fit the prediction only to some extent. The NI subjects scored significantly less on three performance scales (fluency; originality; number) but higher on two scales including complex reasoning tasks (verbal reasoning; figural processing), the comparison groups differ by each other only in measures of originality. Scores of introversion determinate individual differences in intellectual performance more than that of neuroticism.

INTELLIGENCE PERFORMANCE DEPENDING ON NEUROTICISM AND INTROVERSION

- Heinz-Jürgen Ebenrett* -

1. Question

Famous researchers generally agree that personality features such as neuroticism and extraversion - introversion interact with intelligence performance in complex, though meaningful ways. Significant relations can regularly be obtained when laboratory learning tasks are used (CATTELL 1934; EYSENCK 1967; M. W. EYSENCK 1976), but they often fail to appear when orthodox intelligence tests are employed (GIBSON 1975; GREIF et al. 1977; SEDDON 1977).

The present paper is aimed to proof the assumption that different degrees both of neuroticism and extraversion-introversion go hand in hand with individual differences in orthodox test performance.

2. Methods/Basic Data

The present data originate from a study named "Productive Thinking and Problem Solving", which was carried out in 1975 by several members** of the Institute for Psychology of the Free University of Berlin.

Within this study 545 high school students*** have carried out 98 different orthodox paper-pencil intelligence tests within three days, as well as some questionnaires concerning temperamental traits, motivations and interests.

---

* Research Psychologist, Dezernat Wehrpsychologie, Streitkraefteamt, P.O. Box 20 50 03, D-5300 Bonn 2

** The team mentioned is still organized as "Forschungsprojektschwerpunkt Produktives Denken/Intelligentes Verhalten" (JAEGER et al. 1977). Their leader is Prof. Dr. A. O. JAEGER. The present author has been their speaker for several years.

*** Average age: 17.6, ranging from 15 to 21 years of age; of those: 289 male and 256 female students

2.1 Variables concerning dimensions of intelligence

By means of factoranalysis (SPSS, VARIMAX-rotation) the total member of the 98 orthodox intelligence tests could be structured into the following factors of intelligence (VIEWGRAPH 1):

- Memory ("Gedächtnis")
- Fluency ("Einfallsreichtum U,X")
- Verbal Reasoning ("Verarbeitung komplexer verbaler Information")
- Speed of Handling ("Bearbeitungsgeschwindigkeit einfachstrukturierter Aufgaben")
- Figural Reasoning ("Verarbeitung komplexer figuraler Information")
- Number facility ("Rechenfertigkeit)
- Originality ("Originalität")

For further purposes we composed seven equivalent scales, each representing one of the previous factors and including those tests, which mark the respective factor best. By means of these scales we could manage to attach standardized scores to each subject, which stand for its individual localization on each of the referred dimensions of intelligence.

2.2 Variables concerning differences in neuroticism and extraversion-introversion

For measuring temperamental features we used the "Freiburger Persönlichkeitsinventar (FAHRENBERG, SELG und HAMPEL 1973), which includes a number of 212 questionnaire items and allows to obtain individual scores for 9 first order traits of temperament as well as scores for the following second order traits:

FPI-N: Emotional Lability vs. Emotional Stability (Neuroticism)
FPI-E: Extraversion vs. Introversion

Depending on whether a subject achieved high or low scores in neuroticism (FPI-N) and/or extraversion-introversion (FPI-E) respectively, eight different subgroups have been formed (VIEWGRAPH 2):

Subgroup	FPI-Staninescores	N
E=Extraverts	FPI-E:7-9	237
I=Introverts	FPI-E:1-3	75
S=Stabile	FPI-N:1-3	92
N=Labile	FPI-N:7-9	124
SE=Stabile Extraverts	FPI-E:6-9 as well as FPI-N:1-4	130
NE=Labile Extraverts	FPI-E:6-9 as well as FPI-N:6-9	124
SI=Stabile Introverts	FPI-E:1-4 as well as FPI-N:1-4	59
NI=Labile Introverts	FPI-E:1-4 as well as FPI-N:6-9	54

(We had expected that the stanine scores 1-3 or 7-9 respectively would approximately include a quarter of the total sample. However, the empirical data show significantly higher frequencies of extraverts and labile subjects. [EYSENCK 1972 gave similar results] . We have pointed out that these results could be typical for german high school students and that selection modes and stress could have caused the slope).

In order to proof our basic assumption we compared the levels of intelligence tests performances of the previous subgroups with one another by means of t-test analysis.

3. Results

3.1 Performance of extraverts vs. introverts (VIEWGRAPH 3)

	Extraverts N=237		Introverts N=75		
	X	(s)	X	(s)	X
Verbal Reasoning	49.98	(5.49)	50.59	(6.12)	-0.64
Figural Reasoning	49.50	(5.63)	50.69	(5.40)	-2.27*
Memory	50.08	(6.26)	49.80	(7.38)	0.29
Speed of Handling	50.10	(6.21)	49.66	(6.48)	0.63
Number Facility	50.29	(6.60)	49.18	(8.18)	1.07
Originality	52.10	(6.73)	47.21	(7.29)	5.15**
Fluency	51.35	(7.45)	47.08	(5.27)	5.49**

The present viewgraph contains a comparison of the T-standardized average scores (in brackets: standard deviations) of the extraverts with those of the introverts with regard to intelligence performance. (The t-values represent the

degree of significance of the respective score difference tested: Those marked by a double asterix are significant at the 1 %-level, those marked by a single one at the 5 %-level).

Based on the viewgraph the following statements can be made:

1. In total the empirical group differences are only of slight importance. The average scores of both subgroups vary little as against the average scores of the total sample ($\bar{X}$ = 50).
2. In all but the two reasoning scales, which include tasks of high complexity, the extraverts performed better than the introverts. With regard to the two productivity scales ("Fluency" and "Originality") the group differences are significant at the 1 %-level.
3. With regard to the figural reasoning scale, introverts reached significantly (5 %-level) higher scores than extraverts.

3.2 Performance of stabile vs. labile subjects (VIEWGRAPH 4)

	Stabile N=92		Labile N=124		
	$\bar{X}$	(s)	$\bar{X}$	(s)	t
Verbal Reasoning	49.51	(6.13)	50.14	(5.57)	-0.77
Figural Reasoning	50.80	(6.04)	49.98	(5.47)	1.02
Memory	50.23	(5.30)	49.47	(6.69)	0.93
Speed of Handling	51.47	(5.97)	49.35	(6.24)	2.53*
Number of Facility	50.77	(5.39)	49.73	(6.75)	1.25
Originality	50.22	(6.61)	50.40	(7.53)	-0.18
Fluency	50.23	(5.30)	49.47	(6.69)	0.93

In the same way as the previous viewgraph, the present one shows average scores of stabile subjects contrary to labile subjects. The data once more reflect only small differences between the compared groups. Nevertheless, two further statements can be made:

4. In all but two scales stabile subjects performed better than labile subjects. The average scores with an opposite tendency ("Verbal Reasoning" and "Originality") differ the least

5. Only with regard to the speed scale ("Speed of Handling") stabile subjects performed significantly better than labile subjects. (5 %-level)

3.3 Performance of the "combined" groups
(VIEWGRAPH 5)

	Stabile Extraverts N=130 $\bar{X}$	Labile Extraverts N=124 $\bar{X}$	Stabile Introverts N=59 $\bar{X}$	Labile Introverts N=54 $\bar{X}$
Verbal Reasoning	49.86	49.91	49.59	51.24
Figural Reasoning	49.77	49.31	50.26	50.81
Memory	50.30	49.58	50.77	48.79
Speed of Handling	50.73	49.67	50.75	48.38
Number Facility	50.38	50.19	51.69	46.92
Originality	50.74	52.39	47.28	46.96
Fluency	51.03	50.62	48.75	45.22

	Stabile Extraverts N=130		Labile Introverts N=54		
	$\bar{X}$	(s)	$\bar{X}$	(s)	t
Verbal Reasoning	49.86	(5.64)	51.24	(5.51)	-1.53
Figural Reasoning	49.77	(5.68)	50.81	(5.25)	-1.18
Memory	50.30	(5.76)	48.79	(7.67)	1.72
Speed of Handling	50.73	(6.13)	48.38	(5.71)	2.48*
Number Facility	50.38	(6.36)	46.92	(6.84)	3.18**
Originality	50.74	(6.61)	46.96	(8.50)	2.92**
Fluency	51.03	(7.19)	45.22	(5.93)	5.93**

Viewgraph 5, in the upper part contains a synopsis of average scores of the four subgroups SE, NE, SI and NI with regard to the referred scales of intelligence performance.
The lower part of the present viewgraph shows the results of t-test-analysis concerning the two subgroups SE and NI, which differ most in the variables referred.

The results allow the following statements:

6. Scores of intelligence performance of the "combined" subgroups SE, NE, SI and NI differ to a greater extend than those of the previous subgroups E/I as well as S/L.

7. Significant deviations essentially concern the critical subgroup NI while the other groups differ less as against the expected average scores of $\bar{X} = 50$.

8. The subgroups SE and NI differ most in the present scores of intelligence performance. SI and NE subjects take up intermediate positions.

9. In all but the two reasoning scales SE subjects perform better than NI subjects; usually to a significantly greater extend (exception:"Memory").

10. NI subjects performed best in the two reasoning scales (including tasks of high complexity!) though not to a significant extend.

4. Discussion

In total the present results confirm the hypothesis that the (second order) temperamental traits extraversion-introversion as well as neuroticism interact with intelligence performance in meaningful ways - even when orthodox intelligence tests are employed -; though not constantly in the way we could have expected on the basis of literary research.

All in all the following two main results correspond with previous results rather well:

- Introverts performed better than extraverts with regard to figural and - with some reservations - verbal reasoning, but less to speed tests as well as verbal productions and simple arithmetical tasks.
  Due to EYSENCKs theory extraverts show a more functional arousal but generate reactive inhibition more strongly and more quickly; they therefore "opt for speed, introverts for accuracy" (EYSENCK 1967, p.92)
- NI subjects significantly showed less performance in all but two intelligence scales.
  This is due to the assumption (EYSENCK 1967)that NI subjects are characterized by a high arousal (because of being introverts) as well as by a high drive (because of being neurotics). They therefore reach a summarized drive level, which lies above the functional optimum for adequate intelligence performance.

(312)

Contrary to these results the following do not correspond with theoretical expections:

- Stabile subjects, especially stabile extraverts, performed better than the other groups. Due to the Yerkes-Dodson-law it could have been expected "that the optimum drive level for complex and difficult tasks like those involved in an intelligence test lies below the high level reached by high N subjects, and above that reached by low N subjects" (H.J. EYSENCK 1967, p. 92). This assumption does not correspond with the present data. With regard to the most complex tasks ("Verbal Reasoning" and "Figural Reasoning") the high N subjects and the high NI subjects even performed best.

We do not believe that these unexpected results can defeat EYSENCKs basic assumptions. They can be exhausted by two facts: First, the test situation was not exceptionally stressing but similar to ordinary lessons; secondly, the degrees of neuroticism probably did not reach clinical levels (otherwise the respective students would have failed during their previous careers).

On the other hand, during further steps of analysis we found weighty reasons that the A.O. construct U.I.21 "Exuberance" (CATTELL and WARBURTON 1967) was a better predictor for an adequate intelligence performance than the measures of neuroticism and extraversion-introversion (EBENRETT 1980).

5. References

CATTELL, R.B.: Temperament tests: II. Tests. Brit. J. Psychol. 24, 1934, 20-49

CATTELL, R.B. & WARBURTON, F.W.: Objective personality and motivation tests. Urbana (University of Illinois Press) 1967

EBENRETT, H.-J.: Intelligenz und Temperament: Eine Untersuchung bereichsübergreifender Beziehungen auf unterschiedlichem Variablenniveau. Wehrpsychol. Untersuchungen (Bonn, BMVg - P II 4) 1980 (in preparation)

EYSENCK, H.J.: Intelligence assessment: a theoretical and experimental approach. Brit. J. Educ. Psychol. 37, 1967, 81-98

EYSENCK, H.J.: Personality and attainment: an application of psychological principles to educational objectives. Higher Educ. 1, 1972, 39-52

EYSENCK, M.W.: Extraversion, verbal learning and memory. Psychol. Bull. 83, 1976, 75-90

GIBSON, H.B.: Relations between performance on the advanced matrices and the EPI in high-intelligence subjects. Brit. J. Soc. Chin. Psychol. 14, 1975, 363-369

GREIF, D., GREIF, S. & LIEPMANN, D.: Beziehungen zwischen Eysencks Persönlichkeitstypen und Intelligenzleistungen. Z.Klin. Psychol. und Psychotherapie 1, 1977, 29-42

JAEGER, A.O., FABER, J., KOENIG, F. & SCHMIDT, J.U.: Zwischenbericht des FPS "Produktives Denken und Problemlösen", Berichtzeitraum 1. 11. 75 - 31. 10. 77.
Zwischenbericht an die FNK der Freien Universität Berlin vom 15. 11. 1977

SEDDON, G.M.: The effects of chronological age on the relationship of academic achievement with extraversion and neuroticism: a follow up study. Brit. J. Educ. Psychol. 47, 1977, 187-192

EDDOWES, Edward E., and DEMAIO, Joseph C., Operations Training Division, U.S. Air Force Human Resources Laboratory, Williams AFB, Arizona.

IDENTIFICATION, DEFINITION AND MEASUREMENT OF CRITICAL FLYING SKILLS (Wed P.M.)

The objective of this study was to determine if flying skills could be identified, defined, and measured. It was part of a program to develop quantitative, objective procedures for the efficient management of aircrew training.

Fighter pilots were interviewed to select sample tasks, specify pilot actions required to perform them, and identify and define the skills involved in their performance. Analyses of the pop-up weapon delivery and low altitude tactical formation tasks identified six skills: planning, recheck, discriminating, anticipating, deciding, and controlling. Skill measurement procedures in which pilots rated their bombing and formation flying performances were developed. Skill ratings were collected to evaluate the measurement procedures.

Contingency Chi Square analyses disclosed significant relationships between skill ratings and bomb scores. Multiple regression analyses of formation ratings indicated that position keeping and visual lookout were significant components of formation performance. The results were interpreted as evidence of the validity of the skill measurement approach.

IDENTIFICATION, DEFINITION, AND MEASUREMENT OF CRITICAL FLYING SKILLS

Edward E. Eddowes and Joseph C. DeMaio
Operations Training Division
Air Force Human Resources Laboratory
Williams Air Force Base, Arizona

A flying training research program was developed to identify and define critical combat skills of mission-ready aircrews, and to develop a practical technology for measuring these skills. The goal of the research is to develop and validate comprehensive, quantitative, objective procedures for the efficient management of individualized flying training which will provide aircrew mission readiness at minimum cost. The specific objectives of the first phase of the program reported here were to: 1. Provide an early evaluation of the key concepts and methods of the whole program, 2. Identify and define selected critical flying skills, 3. Develop procedures for measuring these skills, and 4. Evaluate and refine the skill measures.

TECHNICAL APPROACH

Previous efforts to improve flying training have focused on task requirements. This research, however, focuses on identifying, defining and developing procedures for measuring the critical skills underlying aircrew mission readiness. Research reported by Meyer, Laveson, Weissman, and Eddowes (1974), and Meyer, Laveson, Pape, and Edwards (1978), indicated that a small, manageable number of skills could be identified and defined which cut across the range of task performances required of mission-ready U.S. Air Force Tactical Air Command (TAC) pilots.

Research systematically addressing Air Force continuation training has been minimal. Therefore, this project was initiated with a study designed to develop skill definition and measurement procedures based on study of a small sample of critical flying tasks. During the first part of the program, research efforts were aimed at developing, evaluating, and refining skill measurement procedures designed to assess pilot skills exercised in accomplishing the pop-up weapon delivery (pop-up) and the low altitude tactical formation (LATF) tasks. These measurement procedures were developed in coordination with a mission-ready A-7 squadron and an F-4 combat crew training squadron through a series of iterative refinements based on pilot self-assessment records collected during the flying training operations of these two squadrons.

Following development and initial test of the skill measurement procedures, it was determined that they should be evaluated further and cross-validated. Consequently, a TAC-wide test of skill measurement procedures for the pop-up and LATF tasks was designed and implemented.

METHOD

Participants Pilots who contributed to the results of the Preliminary Evaluation included those of the 354th Tactical Fighter Squadron (TFS)

(A-7), Davis-Monthan AFB, Arizona; the 311th Tactical Fighter Training Squadron (TFTS) (F-4), Luke AFB, Arizona; the 4th Tactical Fighter Wing (TFW) (F-4), Seymour Johnson AFB, North Carolina; the 23rd TFW (A-7), England AFB, Louisiana; the 347th TFW (F-4), Moody AFB, Georgia; the 354th TFW (A-10), Myrtle Beach AFB, South Carolina; and the 474th TFW (F-4), Nellis AFB, Nevada.

Materials Research materials consisted of pop-up and LATF self assessment forms.

Procedures The TAC preliminary evaluation was initiated with the cooperation and participation of the 354th TFS, Davis-Monthan AFB. Semi-structured interviews with eight fighter pilots of the 354th TFS were recorded and analyzed. Initial interviews were conducted to identify the critical tasks involved in fighter mission scenarios. Having identified the critical tasks to be studied, the research pilots analyzed each task into its real-time components and described the behavioral actions required to perform each component. These interviews used generalized maneuver diagrams as guidelines. Pilots were interviewed both individually and in groups.

Critical task breakdown summaries were constructed using the data obtained from the interviews. The summaries were reviewed and refined in further interviews to obtain a consensus on their completeness and accuracy. In addition to the original research pilots, interviews to review the task breakdown summaries were conducted with members of the 4444th Operations Squadron (Operational Training Development), A-7 Division, Davis-Monthan AFB, and with the staff of the Fighter Weapons School, 162nd Tactical Fighter Training Group(TFTG), Arizona Air National Guard, Tucson, Arizona, International Airport.

Subsequent interviews focused on validation of the skill identification/definition analyses. Interviews were supplemented by photographing A-7 pop-up weapon delivery maneuvers at the Gila Bend Gunnery Range, AZ, which included audio recordings of pilot commentary during the maneuver to provide additional data on available means for skill measurement. These data served as the basis for development of practical techniques for assessing the critical flying skills identified in the analyses of the pop-up and LATF tasks. Once developed, the skill measurement procedures were implemented in routine flying operations to collect data to use in evaluating and revising them.

RESULTS

Identification and Definition of Critical Tactical Flying Skills Following completion of the analyses of the pop-up and LATF tasks, the tactical taxonomy of Meyer et al. (1978), was used to identify the flying skills exercised at each point during the pop-up maneuver and in accomplishing the major performance requirements in the case of the LATF task. The skills identified and their definitions were reviewed and refined through coordination with participating pilots of the 354th TFS and the 162nd TFTG.

Initially six skills were found to accommodate the performance requirements of the pop-up and LATF tasks: Planning, Recheck,

Discriminating, Anticipating, Deciding and Controlling. Their brief definitions are shown below:

Planning	-	Describing and updating mission requirements.
Recheck	-	Repetitive seeking and filtering of information from within and outside the cockpit.
Discriminating	-	Evaluating fine differences amoung cues/cue patterns.
Anticipating	-	Predicting what aircraft control actions will be required.
Deciding	-	Selecting from among alternative aircraft control actions.
Controlling	-	Achieving and maintaining a series of aircraft system and subsystem states.

Pop-Up Measurement Development Analysis of the pop-up disclosed that the maneuver consisted of a sequence of pilot actions required to control the aircraft to a point in space from which a bomb could be released so as to hit the intended target. The initial analysis was summarized by stage of the maneuver and the major skills exercised during each stage were identified. A self-assessment form was developed on which pilots rated their performance on each stage of the pop-up thus permitting specification of the skills involved. Preliminary evaluation research continued with collection of pop-up weapon delivery skill data for use in refining and revising the self-assessment procedures and forms. Bombing skill data were acquired using revised versions of the pilot self-assessment form for pop-up maneuvers flown on both tactical and controlled ranges. The pop-up self-assessment data were processed and the resulting skill scores were analyzed to determine their relationship with bomb impact scores.

Pop-up data were collected systematically in a study using F-4 student pilots of the 311th TFTS. In this case, the evaluation forms were completed by Instructor Pilots. Analyses of the data confirmed the validity of the measurement techniques, (Pierce, DeMaio, Eddowes and Yates, 1979). Another similar evaluation was performed subsequently at the request of the 311th TFTS, (Pierce, Demaio and Yates, 1979).

LATF Measurement Development Analysis of the low altitude tactical formation task indicated that it involved the continuous performance of four major task elements, formation position keeping, low altitude flying, mutual support (visual lookout), and navigation. Further analysis of these component performances led to identification of the major skills involved. A self-assessment form for recording pilot ratings of performance on the LATF components provided for measurement of skills in the same manner employed with the pop-up. Skill measurement data generated through the use of the pilot self-assessment forms were studied in a series of multiple regression analyses to provide the basis for revision and refinement of the skill measurement materials and procedures.

Formation performance data were collected from pilots of the 354th TFS and subsequently evaluated to determine the utility of the skill measurement procedures for the LATF task. Multiple regression analyses of the LATF data indicated that formation position keeping and mutual support were the key

components of LATF performance. The results were interpreted as confirmation of the adequacy of the measurement technique, (DeMaio and Eddowes, 1979).

TAC-Wide Test The pop-up and LATF skill measurement materials and procedures were tested further during the TAC-wide test. Five tactical fighter wings participated in the test, the 4th, 23rd, 354th, 347th, and 474th.

During the TAC-wide test, more than 1200 pop-up and LATF forms were collected from participating pilots. An analysis of these data demonstrated the validity of the measurement procedures and their generalizability in applications with A-10 and F-4 aircrew personnel, (Lyon, Eubanks, Killion, Nullmeyer and Eddowes, 1980).

CONCLUSION

These findings clearly indicate that a refined methodology for identifying and defining critical tactical fighter pilot skills and for measuring them in routine flying operations has been developed. Additionally the results of the TAC-wide test confirm the validity and generalizability of the measurement techniques across aircraft systems.

While the present results are not the final product of this research, they provide a substantial baseline of proven research procedures and a data base to support and stimulate further development, evaluation, and refinement of the objectives and techniques of Project SMART. During the next phase of the program, critical air combat maneuvering and air-to-ground attack skills not yet studied will be identified and defined, and procedures for their measurement developed and validated.

REFERENCES

DeMaio, J.C., and Eddowes, E.E. Airborne performance measurement assessment: Low altitude tactical formation in two operating environments, AFHRL-TR-79-44. Williams AFB, AZ: Flying Training Division, AFHRL, February, 1980.

Lyon, D.R., Eubanks, J.L., Killion, T.H., Nullmeyer, R.T., and Eddowes, E.E., Critical components of the pop-up maneuver: An analysis using self-assessment data, AFHRL-TR-80-33. Williams AFB, AZ: Operations Training Division, AFHRL, August 1980.

Meyer, R.P., Laveson, J.I., Weissman, N.S., and Eddowes, E.E. Behavioral taxonomy of undergraduate pilot training tasks and skills: Executive summary, AFHRL-TR-74-33(I). Williams AFB, AZ: Flying Training Division, AFHRL, December 1974.

Meyer, R.P., Laveson, J.I., Pape, G.L., and Edwards, B.J. Development and application of a task taxonomy for tactical flying, AFHRL-TR-78-42 (Vols I, II, III). Williams AFB, AZ: Flying Training Division, AFHRL, September 1978.

Pierce, B.J., DeMaio, J.C., Eddowes, E.E. and Yates, D. Airborne performance measurement methodology application and validation: F-4 pop-up training evaluation, AFHRL-TR-79-7. Williams AFB, AZ: Flying Training Division, AFHRL, June 1979.

Pierce, B.J., DeMaio, J.C., and Yates, D. Validation of an in-flight measurement methodology: F-4 ground attack training evaluation. Paper presented at the annual meeting of the Human Factors Society, Boston, October 1979.

GEORGE, E.L., LAVERNE, J.E., SNODGRASS, L., NEAL, G.L., and TOVAR W.R., US Army Training & Doctrine & Command (TRADOC) Systems Analysis Activity, White Sands Missile Range, New Mexico.

COST AND TRAINING EFFECTIVENESS ANALYSIS (CTE) FROM PRINCIPLE TO APPLICATION (Wed A.M.)

In the recently implemented US Army TRADOC Training Effectiveness Analysis (TEA) System (TRADOC Regulation 350-4), the Cost and Training Effectiveness Analysis (CTEA) is performed on newly developing hardware systems to ensure training developments are initiated early and are developed in coordination with hardware development. The CTEA for the PATRIOT air defense missile system is one of the first CTEA's conducted under the TEA system. Test methodology utilized to assess the training subsystem prototype at Operational Test (OT) II is the focus of this study. The use of attitude scales, skills and knowledge tests, and hands-on performance tests to make these assessments is discussed. A unique feature of the CTEA methodology is the development of soldier profiles to assess soldier capability to operate new systems. The construc of such profiles for the PATRIOT system and their implications to system development are summarized. Illustrative results, problems encountered, and lessons learned for future CTEA studies are presented.

COST AND TRAINING EFFECTIVENESS ANALYSIS:

FROM PRINCIPLE TO APPLICATION

Edward L. George, Ph.D.
US Army TRADOC Systems Analysis Activity
White Sands Missile Range, New Mexico 88002

A Cost and Training Effectiveness Analysis (CTEA) is a systematic study to assess developing training subsystems during the hardware acquisition process. It involves the application of formal analytical procedures and empirical methodologies. The overall goal of the CTEA is to insure that once a new hardware system is fielded, it can be operated and maintained effectively by soldiers in the field.

It is imperative that the users be kept in mind while the hardware system is being designed and developed so that efficient and effective systems do not become relatively ineffective because soldiers can not use them to full advantage. Given the current military manpower situation and increasing hardware complexity, it is unwise to develop hardware under the assumption that sufficient users with the necessary characteristics to adequately adapt to the hardware needs will be available. The situation dictates that potential users be considered from system conception, and furthermore that the intended users' needs and capabilities be defined as clearly as possible. When hardware developers make use of such information, the risks of ending up with hardware of limited efficacy are reduced.

In addition to soldier-hardware interface problems, the CTEA is also concerned with how the users can be most effectively trained. Assuming the hardware is designed with user capabilities in mind, there is still the problem of how to train soldiers to proficiency in the most cost efficient manner. The CTEA examines variable costs, as well as variable effectiveness, to arrive at the mix providing the most effectiveness for the cost.

In summary, the CTEA insures that soldier capabilities are considered along with combat development factors when hardware is in the developmental phase. Because of the very nature of these types of studies, CTEAs force developers and trainers to deal with very difficult training issues before the system is deployed.

This paper reports the results of applying this philosophy to one particular developing hardware system. The basic methodology is discussed as well as selected study results in summary form. The CTEA was a complex three part study conducted by a multidisciplinary team. Emphasis will be on the air defense proficiency measures, and the development of selection criteria for PATRIOT operators.

PATRIOT AIR DEFENSE MISSILE SYSTEM CTEA

The US Army is developing the PATRIOT missile system to enhance air defense capability against a 1980-1990 threat characterized by defense suppression tactics using saturation, maneuver, and electronic countermeasures.

The CTEA was conducted in three separate parts:

- Operational Test (OT) II Training Subsystem Analysis
- Training Subsystem Equipment Alternative Analysis
- MISSILE MINDER AN/TSQ-73 Analysis

This paper will present a portion of the first and third parts of the CTEA. From the first part, the determination of soldier proficiency and measurement problems encountered will be discussed. From the last part of the CTEA, the problem of developing selection criteria for PATRIOT operators will be covered. Those interested in more detailed methodology and results are referred to the CTEA.[1]

The CTEA was conducted during the full scale engineering development and Development Test (DT)/Operational Test (OT) II phases of the PATRIOT material acquisition cycle so that training development processes could be measured and shortfalls rectified parallel to and in coordination with combat development processes.

TRAINING PROFICIENCY ASSESSMENT

One of the primary purposes of this study was to determine the effectiveness of collective training. Table 1 shows the type of proficiency measures used in the evaluation. Both hands-on and written tests were used to measure different aspects of proficiency. Due to space limitations, only the operator hands-on test, the Air Defense Mission (ADM), will be discussed here.

The purpose of the ADM test was to determine soldier proficiency in using the PATRIOT system for air defense. The study question was: Is actual effectiveness (E_A) equal to design effectiveness (E_D)?

---

[1]The PATRIOT CTEA (8-80, ACN 56244, April 1980, and August 1980) was published in four volumes by TRADOC Systems Analysis Activity (TRASANA), White Sands Missile Range, New Mexico 88002. Distribution of the report is limited. Requests for the report should be made to HQ, US Army Training and Doctrine Command, ATTN: ATTNG-AE-A, Fort Monroe, VA 23651.

TABLE 1. COLLECTIVE TRAINING PROFICIENCY ASSESSMENT

PATRIOT PROFICIENCY DATA
° HANDS-ON TESTS
°° OPERATOR (ADM)
°° MAINTAINER
°° MARCH ORDER, INITIALIZATION, EMPLACEMENT
° WRITTEN SKILLS/KNOWLEDGE TESTS
°° OPERATOR
°° MAINTAINER

Procedure

Soldier proficiency in operating the PATRIOT in an air defense situation was assessed with the Engagement Control Station (ECS) Troop Proficiency Trainer (TPT). The TPT is built into the ECS, and provides for the introduction of simulated air activity with specially designed cassette tapes. Each soldier was tested individually. The test scenario was a simulated 10 minute air raid based on a North Atlantic Treaty Organization (NATO) approved threat for the 1980-1990 time frame. Players received a group orientation briefing before the test. At this briefing they were provided with an instruction sheet for their use during the actual test. The instruction sheet was discussed in detail and all questions were answered. The test was identical for all players. During the actual test the player was observed by an experienced instructor who recorded observations on a prepared checklist. No verbal interaction between players and evaluator was allowed.

E_D Determination

E_D was determined by presenting the PATRIOT with the test scenario while operating in the automatic mode. This provided an accurate assessment of system capability in the specific situation being tested. For the purpose of analysis, system performance was set equal to 1.0, so comparisons could be reported in an unclassified form.

E_A Determination

The TPT provides a score at the end of each test. This score was originally designated E_A.

E_A vs E_D

The evaluation strategy was to determine E_D, and then make a direct comparison with soldier performance to see if the two were relatively equivalent. After the test scenario was defined, E_D was determined without difficulty. As a test validation measure, an experienced instructor who was familiar with the PATRIOT but not with the test scenario took the ADM test. The instructor's score was identical to the E_D score. In other words, $E_A = E_D$ in this case.

Several problems were encountered with E_A. The TPT score represents the percent of assets successfully defended. For example, a score of 70 means that 70 percent of the defended assets were successfully protected. Table 2 shows the PATRIOT player scores on the pre-OT II ADM test. Nineteen of the 32 (59%) players received a perfect score, indicating that they had successfully defended all of their assets during the test. Only one player had more than 30 percent of his assets damaged. This indicates that the players were proficient, and therefore well trained.

TABLE 2

PRE-OT II ADM TEST SCORES

TPT SCORE	NUMBER OF PLAYERS
100	19
94	1
88	1
70	10
0	1
	32

Unfortunately, a closer look at individual player performances revealed that the TPT score did not accurately reflect player performance. The two following actual cases illustrate the problem:

PLAYER	MISSILES LAUNCHED	KILLS	MISSES	TPT SCORE
X	11	9	2	100
Y	32	26	6	100

Obviously these two players did not perform equally on the test, yet they both received a score of 100, indicating all assets were successfully defended. A target per target comparison with the benchmark (E_D) revealed that player X was far from successful in defending his assigned assets.

There are four criteria often used in evaluating air defense system effectiveness. One is amount of damage to defended assets. The other three are:

- Damage to the air defense system
- Missiles expended
- Enemy aircraft losses

An indepth examination of these three remaining criteria revealed that they posed difficulties as well.

The PATRIOT system continuously evaluates all aircraft in its operating zone and evaluates them for threat according to Army tactical doctrine. When operating in the automatic mode, the PATRIOT prioritizes hostiles and automatically engages those determined to be highly threatening to the defended assets. The highest priority threats are engaged first, and engagements are made at a time calculated to maximize kill efficiency and minimize asset damage.

Those targets the PATRIOT identified as highly threatening and engaged in the automatic mode were identified, and used to evaluate player performance. It was found that total engagements was not meaningful, because players engaged targets that were not priority engagement targets. The E_A score was derived by dividing the number of priority engagements a player made by the number made in the automatic mode. The E_A score is therefore a ratio expressing the proportion of possible high priority engagements each player made.

Figure 1 shows the results of the ADM test at the pre-OT II and post-OT II administrations. The E_D line at the top of the figure shows the performance of the PATRIOT system on the test scenario while operating in the automatic mode. The distance between the top of each bar and the E_D line reflects the difference in performance when the system is operating in the automatic mode (E_D) and when operated by soldiers in the semi-automatic mode (E_A). This difference between E_A and E_D for each crew was tested statistically (Chi-square) and found to be highly significant for both the pre- and post-OT II scores ($p < .001$).[2]

In both ADM tests, the typical or average player engaged approximately half of the hostile targets highly threatening to defended assets. Analysis of player actions during the test revealed several training inadequacies which contributed to the relatively poor player performance. Several procedural errors were made, and the soldiers had not yet learned how to use efficiently the highly sophisticated PATRIOT capabilities. There was no substantial improvement in proficiency from the first to the second test, as can be seen from the overall mean scores of .48 and .49.[3]

ADM Summary

All crews and all individuals except one on the post-OT II test scored significantly below E_D. Several training inadequacies surfaced from the test analysis.

ADM Evaluation Problems

Attempting training evaluation of systems in the developmental cycle poses many problems. One of the most difficult problems was gaining access to the hardware. Since the hardware was being used for training and testing, and also being modified as a result of developmental test findings, there was not enough to "go around".

Another difficulty was determining how to meaningfully score the ADM test. The TPT score was used during training to evaluate soldier proficiency. Contractor representatives recommended using the score because it had been used in training to evaluate soldier proficiency. Use of the score could have resulted in saying the training was adequate when it was not. A recurring problem in CTEA studies is finding meaningful ways to evaluate soldier proficiency.

---

[2]Extensive statistical analysis of these data is reported in the CTEA report. Since the purpose of this paper is to illustrate the application of methodology, detailed results are not included.

[3]This result may not be immediately obvious from Figure 1; however, extensive statistical analysis for both crew and individual scores clearly showed no change in overall performance on the second test.

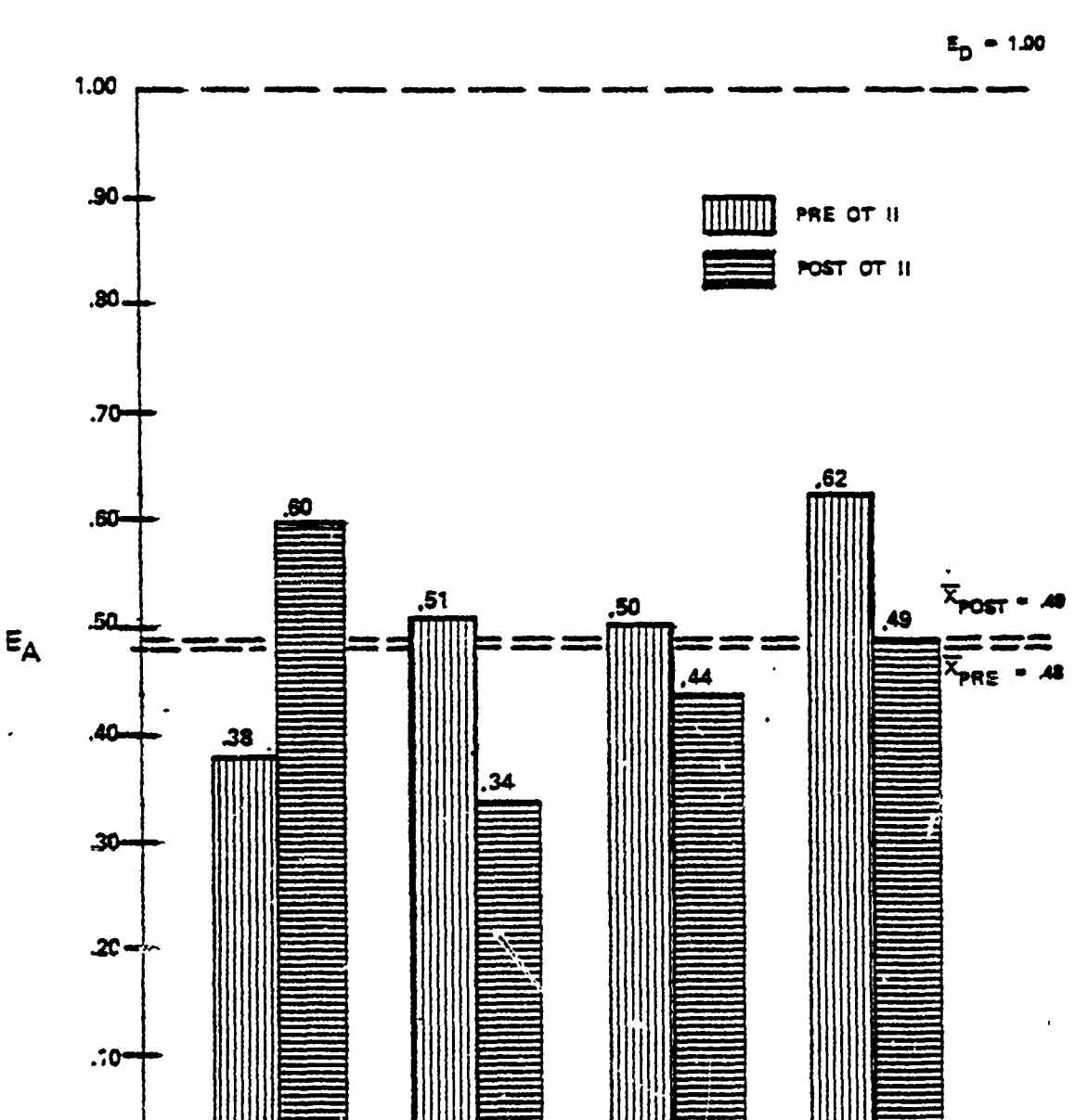


FIGURE 1

PRE- AND POST-OT II ADM SCORES BY CREW (E_A)

GE-7

One of the most important questions to be answered in any CTEA is whether the intended user's soldiers can be trained to operate the hardware effectively. This question presents a unique problem, since soldiers available during OT II may not be "typical". The soldiers who were PATRIOT OT II players were definitely high quality troops. The fact that select troops achieve proficiency cannot be construed to mean that the typical user will also achieve proficiency from the same training.

PATRIOT Operator Selection Criteria

An inherent part of a CTEA is determination of soldier capability to perform the tasks required by the hardware. Implicit in this problem is the question of what qualifications soldiers need to have the required capability. Since the PATRIOT players were not found to be proficient, the problem in this study was to find soldiers who were currently performing similiar tasks in the field. An analysis of the capabilities of soldiers successfully performing similar tasks would provide needed insight into the selection problem.

The MISSILE MINDER (AN/TSQ-73) system currently fielded in Europe has characteristically similar console operations, and thus was selected for study. Records from five classes of students attending the AN/TSQ-73 training course (MOS 25L10) at the US Army Air Defense School (USAADS) were obtained. Data were also collected from the 25L soldiers in Europe.

When the data from USAADS were received, it was noted that approximately 40 percent of the students attending the school failed to graduate. Discriminant analysis was used to determine if the attrition rate could be susbstantially reduced by using different course selection criteria.[4] The analysis revealed that the combination of two different Armed Services Vocational Aptitude Battery (ASVAB) composite scores provided a potential for reducing course attrition. The selection criteria recommended for testing were:

- High school diploma or GED
- Mechanical Maintenance (MM) score over 100
- Skilled Technical (ST) score of 100 or higher

Analysis indicated that if these criteria had been applied to the five classes studied, the 40 percent attrition rate would have been reduced to 16 percent.

---

[4]Selection criteria in effect for the course were: high school diploma or GED, ASVAB Electronics (EL) score of 90, and a Clerical (CL) score of 90.

A comparison of MOS 25L course graduates and PATRIOT operators revealed no difference in aptitude test scores. The PATRIOT players were older, more experienced, and had a higher educational level than the course graduates. It was concluded that the MOS 25L course graduates have the same capability as the PATRIOT players based on the aptitude scores. Given additional years experience, there is no reason to believe the other differences will not disappear as well.

The next question was job proficiency of the MOS 25L soldiers. Supervisor ratings of the soldiers indicated that they were proficient. Task analysis information also revealed that the soldiers had a high level of confidence in their ability to perform MISSILE MINDER tasks which had been determined to be similar to PATRIOT tasks.

At this point, several pertinent study findings were synthesized to derive tentative selection criteria for PATRIOT operators. These findings were:

- PATRIOT operators were not proficient
- Lack of proficiency was due to training inadequacies
- AN/TSQ-73 soldiers were proficient
- AN/TSQ-73 soldiers and PATRIOT players were not different in aptitude scores
- MOS 25L10 course selection criteria can be improved

The best tentative selection criteria for the PATRIOT operator course was thus determined to be the same as the improved MOS 25L10 course selection criteria:

- High school graduate or GED
- MM score of over 100
- ST score of 100 or higher

CTEA Impact

The PATRIOT CTEA concluded that soldiers can be trained to operate the system, contained recommendations for rectifying training inadequacies, and presented insights for PATRIOT personnel selection and training. One specific impact from the study was review of the training planned for the initial two PATRIOT battalions. Other recommendations were made dealing with aspects of the study not discussed in this paper.

STATE OF THE ART

CTEA methodology is currently in an evolutionary phase. There are still many problems to be solved. The methodology will be further refined with each study. Based on experience with studies to date, the following original assumptions are now considered to be closer to facts than assumptions:

- It's possible to determine/predict how effective a piece of equipment is/will be in the hands of the troops (E_A)
- It's possible to determine how effective the same piece of equipment ought to be (E_D)
- It's possible to explain differences in the above and to rectify unsatisfactory ones

It is anticipated that with further studies and refined methodology, the problems currently experienced in trying to determine E_A and E_D will be reduced.

SUMMARY

A CTEA is a systematic empirical study performed on developing hardware systems to insure that user needs and capabilities are considered along with combat development factors. The PATRIOT Air Defense Missile System CTEA was used to illustrate the application of this methodology, and possible benefits to be derived from CTEAs.

GILBERT, Arthur C.F. Ph.D., U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia.

CHARACTERISTICS OF HIGH ACHIEVERS IN ARMY OFFICER BASIC COURSES
(Thu A.M.)

A sample of Army officers who attended Officer Basic Courses (OBC) was divided into two groups on the basis of course achievement. One group consisted of those officers who performed better than predicted on the basis of aptitude measures predictive of performance in these courses, the other group consisted of those officers who performed lower than was predicted on the basis of their aptitude. The two groups of officers were compared on the basis of motivational measures, peer ratings, and on performance measures in subsequent early duty assignments. The results and implications of these results for future research are discussed.

Characteristics of High Achievers in Officer Basic Courses

Arthur C. F. Gilbert, Ph.D.

US Army Research Institute for the Behavioral and Social Sciences[1]
Alexandria, Virginia 22333

Earlier research indicated the predictive utility of the Officer Evaluation Battery (OEB) in predicting final course grades in Officer Basic Courses (Gilbert, 1978). Resulting from this research a group of officers were identified as receiving final course grades higher than would be predicted on the basis of aptitude as reflected by scores on the cognitive subtests of the Officer Evaluation Battery. Obviously, this group of officers can be viewed as doing well in their first post-commissioning Army Experience. A basic issue is if this type of performance continues to be exhibited in early duty assignments. Another question of merit is if these officers differed on measures of interest related to the Army from their contempories and that their performance in Officer Basic Course was enhanced by this factor. Still yet another question of significance is how these officers who were high performers were viewed by their contempories while in the Officer Basic Course.

The first objectives of this research was to compare the performance of those officers who received Officer Basic Course final grades higher than predicted on the basis of aptitude with that of their Officer Basic Course contempories in early duty assignments. The second objective was to compare these two groups of officers on interest measures purportedly related to success as an Army officer. A third objective was to determine if those two groups of officers would receive different ratings from their associates.

Procedure

A sample of 1,048 officers who had data on all of the pertinent Officer Evaluation Battery scales and for whom Officer Basic Course (OBC) final course grades were used as subjects. The aptitude measures consisted of the cognitive scales of the Officer Evaluation Battery (OEB); these are the Combat Leadership (Cognitive), Technical-Managerial (Non-Cognitive), and the Career Potential (Cognitive) scales. The composition of these scales and their predictive utility has been described in a previous paper (Gilbert, 1978).

The regression equation for predicting final Officer Basic Course grades from the three cognitive scale scores was computed. The regression weights were then applied to the three scale scores to determine a predicted OBC final course grade. Next, a comparison was made between the predicted OBC final course grade and the actual final course grade. Subjects were then classified into two groups on the basis of this comparison. When the obtained OBC final

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

course grade was greater than the predicted final course grade, subjects were classified as higher achievers. On the other hand, when the obtained final course grade was less than the predicted course grade subjects were classified into a second group or "other" group. By use of this method 539 officers were classified as being in the high achievers group, while 509 officers were classified in the "other" group.

Analyses were performed by use of t-tests to determine if statistically significant differences existed among the two groups on the measures of duty performance. These measures were ratings obtained on a rating scale, the Performance Evaluation Form, which is described by Gilbert and Grafton (1978). Another series of analyses using t-tests were performed using the non-cognitive scales of the Officer Evaluation Battery and peer ratings as the criterion. These OEB scales are the Combat Leadership (Non-Cognitive), Technical-Managerial (Non-Cognitive), Career Composite (Non-Cognitive), and Career Intent scales. In addition, an analysis was performed to determine if there was a statistically significant difference between the two groups of officers on the final peer ratings that they received in the Officer Basic Courses. For the purposes of this research, all measures used were first converted to Army Standard Scores.

Results and Discussion

The means of the two groups of officers on the duty performance measures are shown in Table 1. Statistically significant differences were found to exist between the two groups on each of the measures at the .01 level. In each instance, the group of officers who achieved better than predicted final course grades in the Officer Basic Course had a higher group mean than did the other group of officers.

The average scores for the two groups of officers on the non-cognitive scales of the Officer Evaluation Battery are shown in Table 2. Statistically significant differences (.01 level) were found to exist between the two groups of officers on the Combat Leadership (Non-Cognitive) scale and on the Career Intent Scale. The group of officers who achieved higher than predicted Officer Basic Course final grades had the higher group mean on these two scales. There were not any statistically significant differences between the two groups on the other non-cognitive scales of the OEB. A significant difference was also found to exist between means of the two groups of officers on the final peer ratings received in the Officer Basic Course. The mean final course peer rating for those officers who received final course grades that were greater than predicted was highest.

The results of this exploratory research clearly indicate that officers who receive Officer Basic Course final course grades greater than predicted on the basis of aptitude receive significantly higher ratings on measures of performance during the early part of their active duty tour. Their average overall performance is greater than for other officers as reflected in the overall Duty Performance scale of the Performance Evaluation Form and Officer Efficiency Report scores. On specific dimensions of officer performance as measured by the Performance Evaluation Form, they also received higher ratings.

Table 1

Performance Measures

Variables	Mean: High Achievers (N=539)	Mean: Others (N=509)
Duty Performance Measures		
Total Duty Performance	102.30	97.57**
Combat Leadership	102.17	97.70**
Technical Managerial Leadership	102.22	97.65**
Tactical Knowledge	102.22	97.65**
Understanding Mission	102.15	97.70**
Making Decisions	102.82	97.02**
Defining Subordinate Roles	101.79	98.07**
Planning & Organizing	102.21	97.68*
Motivating Troops	101.54	98.34*
Logistical Knowledge	102.24	97.63**
OER Scores		
First Year	101.49	88.72**
Second Year	102.53	97.32**
Third Year	102.04	97.84**
Three-year Average	102.49	97.36**

*Indicates a significant difference between groups at the .05 level

**Indicates a significant difference between groups at the .01 level.

Table 2

Non-Cognitive Measures and Peer Ratings

Variables	Mean	
	High Achievers (N=539)	Others (N=509)
Non-cognitive Measures:		
Combat Leadership	110.91	107.84*
Technical Managerial Leadership	105.29	103.33
Career Potential	108.15	105.92
Career Intent	117.45	113.74**
Peer Ratings	105.09	94.62**

*Indicates a statistically significant difference at the .05 level

**Indicates a statistically significant difference between groups at the .05 level

Officers who receive Officer Basic Course final course grades better than expected on the basis of aptitude scores, displayed a greater interest in becoming an Army officer as reflected in the Career Intent subtest of Officer Evaluation Battery. These officers also displayed interest in those activities related to success as a combat leader as measured by the Combat Leadership (Non-Cognitive) scale of the Officer Evaluation Battery. It could be postulated that these interests contributed to higher performance in the Officer Basic Course and subsequent assignments. In the Officer Basic Course, the officers who achieved well were viewed as having greater leadership potential by their classmates as is indicated by the peer ratings received at the end of the course.

Future research will be aimed at replicating the results of this investigation in other samples of Officer Basic Course graduates. Also, the performance of this sample will be evaluated to determine if the differences reported here will continue to persist over a longer period of time.

REFERENCES

Gilbert, A. C. F. Predictive utility of the Officer Evaluation Battery (OEB). Paper presented at the 20th Annual Conference of the Military Testing Association, Oklahoma City, October 30-November 3, 1978.

Gilbert, A. C. F., & Grafton, F. C. Characteristics of an officer evaluation measure. Paper presented at the Annual Conference of the Military Testing Association, Gulf Shores, AL, October 18-22, 1976. In Proceedings, 18th Annual Conference of the Military Testing Association. Pensacola, FL: Naval Education and Training Program Development Center, 1976.

841

GILBERT, Arthur C.F. Ph.D., U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia.

COLLEGE MAJOR AND ARMY OFFICER PERFORMANCE (Wed P.M.)

The purpose of this research was to determine the influence of the major field of study pursued in college on subsequent Army officer performance. A sample of officers who attended Officer Basic Courses (OBC) in the same year was divided on the basis of their college major and these groups were compared on several psychometric and performance measures. Analyses were performed separately for the three different groupings of the 13 Career Branches in the U.S. Army (i.e., Combat Arms, Combat Support, and Service Support) and the performance of officers within these three groups was assessed in terms of differential academic background. The results of these analyses are presented and the implications for assignment strategies are discussed.

College Major and Army Officer Performance

Arthur C. F. Gilbert, Ph.D.

US Army Research Institute for the Behavioral and Social Sciences[1]
Alexandria, Virginia 22333

The consideration of undergraduate academic preparation in the branch assignment of Army officers raises the question of the contribution such preparation makes to successful performance. In certain assignments of officers such a consideration is obvious. For example, an engineering degree would probably be the ideal preparation for assignment to the Corps of Engineers. The answer is not so obvious in what constitutes the prerequisite civilian education for assignment to the Infantry Branch or to the Armor Branch. Insofar as civilian academic education influences duty performance, then this preparation must be a factor in the assignment process to the degree that is possible within the constraints of the assignment system.

The purpose of this research was to explore the possible influence of college preparation on officer performance. An initial effort (Gilbert, 1978) indicated that differences in duty performance do exist among officers who pursue different fields of study in the Field Artillery Branch. The specfic objective of this research was to determine if these results would occur Army-wide and if the findings would differ in the three major groupings of the Army career branches, Combat Arms, Combat Support, and Combat Service Support.

Procedure

Measures of aptitude and performance were obtained on a sample of officers in Officer Basic Courses (OBC) and duty performance measures after approximately one year of active duty. The measures of aptitude and performance collected at the Officer Basic Courses (OBC) are shown in Table 1. The aptitude measures consisted of the seven scales of the Officer Evaluation Battery (OEB) and the three composite scales shown in the table. Peer ratings were obtained at the middle and the end of the course and final course grades were collected.

The other duty performance measures used in this research are shown in Table 2. One of these measures consisted of a specially constructed Performance Evaluation Form (Gilbert, 1975) which reflects the dimensions derived from the research reported by Helme, Willemin and Grafton (1971), Stogdill (1974), and Willemin (1965). This Performance Evaluation Form was completed by the immediate supervisor of each officer, a superior officer other than his immediate supervisor, and by two close associates. These four ratings were then averaged for each scale of the instrument. In addition, the Officer Efficiency Report scores were obtained for each of the first three years of active duty.

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

For the first analysis of the data, subjects were divided into five groups of college majors for the total sample. These groups were Business, Humanities, Engineering, Physical Science, and Social Science. Analysis of variance was used to evaluate the differences among the five groups on each of the aptitude and performance measures.

For the next set of analyses, subjects were divided on the basis of membership in the three types of career branches: Combat Arms, Combat Support, and Combat Service Support. Within each group of career branches, subjects were classified on the basis of the five kinds of college majors. Analysis of variance was used within each group of branches to explore the differences among college major groups on the different measures.

Results and Discussion

The means of the five groups of college majors on the Officer Evaluation Battery are shown in Table 1 for the total sample. Significant differences were obtained for all of the seven scales and for the three composite scales. The Engineering and Physical Science groups have higher group means on these measures. A significant difference did not exist on mid-course peer ratings in the OBC's but there was a difference on the final peer ratings. Again, a significant difference was found among the groups on the final course grade.

The overall duty performance scale of the Performance Evaluation did not reveal any difference among groups for the total sample shown in Table 2 but a highly significant (.01 level) difference among the groups did exist among the groups on four of the other nine dimensions. Differences among the means on the Officer Efficiency Report scores were significant at the .05 level with the exception of the 1976 Annual OER score where the means of five groups were significantly different at the .01 level.

The Tables 3 and 4 the results of the analyses within the Combat Arms Branches are shown. The differences among groups shown in Table 3 on the OEB scales were all significant (.01 level) but there was not a significant difference on the other OBC measures (i.e. mid-course and final peer ratings and final course grade). Differences among groups were found only on the decision making scale of the Performance Evaluation Form and this was only significant at the .05 level. Again, there were differences among the groups on the Officer Efficiency Report scores as shown.

In the Combat Support branches differences were found among the groups on all of the OEB scales and on all of the OBC performance measures as shown in Table 5. For these branches, significant differences were obtained on four of the Performance Evaluation Form scales as shown in Table 6 but there was not any differences among the groups on the Officer Efficiency Report criteria.

The analyses within the Combat Service Support Branches yielded differences among the five groups on all of the OEB scales and on the OBC performance measures as shown in Table 7. Only the Combat Leadership scale of the Performance Evaluation Form yielded a significant difference among the five groups and there were not any differences on the OER scores (Table 8).

Table 1

Means for the Five Groups of College Majors on the Officer Evaluation Battery and on Officer Basic Course Measures for the Total Sample

Variable	MEAN				
	Business (N=726)	Humanities (N=292)	Engineering (N=388)	Physical Sciences (N=1,379)	Social Studies (N=1,436)
Combat Leadership					
Composite	102.78	97.19	112.59	112.79	102.99**
Cognitive	99.35	95.91	107.95	111.25	99.41**
Non-cognitive	105.16	99.54	112.28	109.29	105.44**
Technical/Managerial					
Composite	98.12	100.99	118.50	115.84	99.12**
Cognitive	97.56	103.38	117.04	117.38	101.07**
Non-cognitive	99.51	98.25	111.92	111.45	97.55**
Career Potential					
Composite	93.97	99.59	113.73	110.31	100.92*
Cognitive	98.33	96.98	116.12	107.06	96.38**
Non-cognitive	92.24	102.43	105.45	109.12	105.08
Career Intent	114.73	116.24	111.56	112.43	117.26**
Leadership Peer Rating					
Mid Course	101.02	97.23	100.56	100.66	99.24
Final	100.80	97.18	102.36	100.20	99.32**
Final OBC Grade	99.51	99.06	106.93	100.54	98.00**

*Indicates a significant difference among groups at the .05 level.

**Indicates a significant difference among groups at the .01 level.

Table 2

Means of the Five Groups of College Majors on Measures of Duty Performance for the Total Sample

Variable	MEAN				
	Business (N=726)	Humanities (N=292)	Engineering (N=386)	Physical Sciences (N=1,379)	Social Studies (N=1,4367)
Duty Performance	101.82	99.02	99.89	100.78	98.28
Combat Leadership	98.74	93.07	98.35	104.20	97.13**
Technical/Managerial Leadership	102.90	98.58	102.21	100.30	97.80**
Tactical Knowledge	97.74	92.72	98.34	104.56	97.26**
Understanding Mission	102.11	98.38	99.58	100.83	98.28**
Making Decisions	101.24	96.48	98.53	101.99	97.95**
Defining Subordinate Roles	101.85	100.12	98.87	100.59	98.57
Planning and Organizing	102.31	99.20	100.98	100.00	98.70
Motivating Troops	102.10	98.31	98.06	100.51	99.16
Logistical Knowledge	102.60	94.06	102.39	100.50	98.49**
Annual OER Scores					
1974	101.17	100.05	97.18	101.38	98.88*
1975	99.94	97.81	98.42	101.37	99.55*
1976	100.34	99.55	99.89	101.83	98.22**
Weighted OER Scores	100.57	100.05	99.98	101.13	98.63**

*Indicates a significant difference among groups at the .05 level.

**Indicates a significant difference among groups at the .01 level.

Table 5

Means of the Five Groups of College Majors on the
Officer Evaluation Battery and on
Officer Basic Measures for the Combat Army Branches

Variables	MEAN				
	Business (N=380)	Humanities (N=113)	Engineering (N=111)	Physical Sciences (N=986)	Social Studies (N=763)
Combat Leadership					
Composite	108.90	108.46	115.74	114.29	110.81**
Cognitive	102.60	102.84	107.62	112.45	105.13**
Non-Cognitive	111.69	110.70	117.62	110.50	112.23**
Technical/Managerial					
Composite	100.76	103.89	119.78	115.98	101.21**
Cognitive	98.81	103.35	117.30	117.81	102.24**
Non-cognitive	102.40	102.87	113.76	107.24	99.64**
Career Potential					
Composite	98.41	105.43	115.76	111.40	105.87**
Cognitive	99.35	100.00	113.69	106.59	98.94**
Non-cognitive	98.19	108.62	111.06	111.31	110.27**
Career Intent	118.79	118.94	115.18	112.80	119.57**
Leadership Peer Rating					
Mid-course	98.82	99.76	101.23	99.99	100.36
Final	100.34	99.63	101.13	99.98	99.76
OBC Final Grade	102.32	100.67	106.37	99.84	101.55

*Indicates a significant difference among groups at the .05 level.

**Indicates a significant difference among groups at the .01 level.

Table 4

Means of the Five Groups of College Majors on Duty Performance Measures for the Combat Arms Branches

Variables	MEAN				
	Business (N=380)	Humanities (N=113)	Engineering (N=111)	Physical Sciences (N=986)	Social Studies (N=763)
Duty Performance	102.80	96.97	101.43	99.92	99.23
Combat Leadership	100.96	95.88	96.89	101.35	98.29
Technical/Managerial Leadership	103.83	96.01	100.45	99.97	98.94
Tactical Knowledge	99.22	96.04	97.84	101.50	98.63
Understanding Mission	103.57	96.30	101.63	100.08	98.66
Making Decisions	103.37	94.46	101.36	100.32	98.67*
Defining Subordinate Roles	99.84	102.57	102.01	100.08	98.50
Planning and Organizing	100.33	96.77	101.55	99.63	99.50
Motivating Troops	103.56	98.93	99.90	99.50	99.50
Logistical Knowledge	103.91	97.40	99.27	99.71	99.30
Annual OER Scores					
1974	101.06	100.37	98.01	101.24	98.09*
1975	100.40	93.59	99.48	101.51	98.84**
1976	100.71	98.30	100.66	101.60	97.76**
Weighted OER Scores	101.25	98.13	99.27	101.46	97.88**

*Indicates a significant difference among groups at the .05 level.

**Indicates a significant difference among groups at the .01 level.

Table 5

Means of the Five Groups of College Majors on the Officer Evaluations Battery and on Officer Basic Course Measures for the Combat Support Branches

Variables	MEAN				
	Business (N=111)	Humanities (N=97)	Engineering (N=230)	Physical Sciences (N=273)	Social Studies (N=375)
Combat Leadership					
Composite	103.41	95.12	112.27	109.73	98.81**
Cognitive	98.81	93.85	108.02	108.58	96.31**
Non-cognitive	105.71	98.31	111.71	107.05	101.86**
Technical/Managerial Leadership					
Composite	102.03	99.71	118.29	118.32	98.04**
Cognitive	99.58	103.61	116.73	118.45	100.48**
Non-cognitive	103.62	96.02	111.85	110.23	96.46**
Career Potential					
Composite	96.95	100.67	113.54	107.80	98.98**
Cognitive	101.21	96.79	117.02	107.25	94.59**
Non-cognitive	93.99	104.36	104.28	104.97	103.82**
Career Intent	112.16	117.20	112.08	112.06	115.99**
Leadership Peer Rating					
Mid-course	106.06	95.30	101.10	101.06	97.85**
Final	101.76	96.72	101.81	100.53	98.91
Final OBC Grade	94.52	97.71	107.71	102.08	95.40**

*Indicates a significant difference among groups at the .05 level.

**Indicates a significant difference among groups at the .01 level.

Table 6

Mean of the Five Groups of College Majors on Duty Performance Measures for the Combat Support Branches

Variables	MEAN				
	Business (N=111)	Humanities (N=97)	Engineering (N=230)	Physical Sciences (N=273)	Social Studies (N=375)
Duty Performance	101.55	101.78	100.67	101.71	96.94
Combat Leadership	100.91	95.88	102.20	104.59	94.49*
Technical/Managerial Leadership	100.48	99.59	103.81	102.55	95.41
Tactical Knowledge	102.31	96.23	102.01	104.25	94.63
Understanding Mission	103.93	101.43	100.99	100.83	96.85
Making Decisions	100.81	100.13	99.19	103.37	96.89
Defining Subordinate Roles	101.26	99.72	100.00	101.81	97.85
Planning and Organizing	100.91	99.32	101.31	101.60	97.45
Motivating Troops	103.28	100.04	100.02	101.34	97.51
Logistical Knowledge	101.55	93.56	103.08	102.19	96.87*
Annual OER Scores					
1974	98.85	100.92	96.17	100.79	101.84
1975	99.91	101.41	97.67	100.22	100.96
1976	99.71	102.29	100.89	101.05	98.20
Weighted OER Scores	98.29	102.23	100.33	100.51	99.35

*Indicates a significant difference among groups at the .05 level.

**Indicates a significant difference among groups at the .01 level.

Table 7

Means of the Five Groups of College Majors on the Officer Evaluation Battery and on Officer Basic Course Measures for the Combat Service Support Branches

Variables	MEAN				
	Business (N=235)	Humanities (N=82)	Engineering (N=47)	Physical Sciences (N=120)	Social Studies (N=298)
Combat Leadership					
Composite	92.59	84.10	106.70	107.52	88.25**
Cognitive	93.87	88.81	108.40	107.49	88.70**
Non-cognitive	94.34	85.65	102.40	104.50	92.54**
Technical/Managerial Leadership					
Composite	92.02	98.52	116.51	109.04	95.13**
Cognitive	94.58	103.15	117.98	111.42	98.84**
Non-cognitive	92.91	94.54	107.92	102.81	93.57**
Career Potential					
Composite	85.38	90.27	109.87	107.09	90.70**
Cognitive	95.31	93.05	117.45	110.49	92.09**
Non-cognitive	81.79	91.62	97.92	100.61	93.37**
Career Intent	109.38	111.38	100.49	110.21	112.95*
Leadership Peer Rating					
Mid-course	101.55	96.45	94.58	105.91	98.53**
Final	101.83	94.27	99.60	103.35	98.91*
Final OBC Grade	96.62	98.71	103.92	101.79	90.27**

*Indicates a significant difference among groups at the .05 level.

**Indicates a significant difference among groups at the .01 level.

Table 8

Mean of the Five Groups of College Majors on Duty Performance Measures for the Combat Service Support Branches

Variables	MEAN				
	Business (N=235)	Humanities (N=82)	Engineering (N=47)	Physical Sciences (N=120)	Social Studies (N=298)
Duty Performance	101.27	100.20	98.66	102.29	97.84
Combat Leadership	101.82	93.86	97.66	105.45	98.23*
Technical/Managerial Leadership	101.89	99.50	97.05	101.94	98.02
Tactical Knowledge	101.34	93.16	98.30	106.76	98.14*
Understanding Mission	100.47	99.45	94.30	102.56	99.75
Making Decisions	101.01	93.30	97.98	104.14	97.98
Defining Subordinate Roles	101.06	101.04	91.96	100.97	99.97
Planning and Organizing	101.54	101.81	98.13	99.85	98.31
Motivating Troops	101.19	97.75	91.45	101.04	101.08
Logistical Knowledge	101.77	92.05	102.65	104.66	98.17*
Annual OER Scores					
1974	100.10	103.25	105.46	99.05	98.68
1975	100.46	100.95	98.28	98.17	100.33
1976	101.05	99.60	95.51	99.99	99.99
Weighted OER Scores	101.47	100.98	99.39	97.18	99.80

*Indicates a significant difference among groups at the .05 level.
**Indicates a significant difference among groups at the .01 level.

(354)

In the total sample, those officers who majored in engineering and Physical Science were favored over the other groups on all of the scales of the OEB with the exception of the Career Intent Scale on which they had the lowest group means. Within the Combat Arms branches Engineering majors had a higher group mean on all scales with the exception of the Career Intent scale. For this analysis, the pattern of differences was not as clear cut as for the total sample. Within the Combat Support branches, Engineering and Physical Sciences majors were favored in terms of their group mean except for the Career Intent Scale where those groups had the lowest means. Finally, within the Combat Service Support these two groups were favored on all scales with the exception of the Career Intent scale where the Physical Science mean was third highest and the Engineering mean was fourth.

Business majors had the highest group mean on mid-course peer ratings in the Combat Support Branches while within the Combat Service Support branches the group mean for Physical Science majors was highest. For final course peer ratings the mean for Engineering majors was highest for the analysis of the total sample and the mean for Physical Science majors was highest in the analysis of the Combat Service Support branches. As was mentioned earlier, these were the only analyses yielding significant differences on the two types of peer ratings (i.e., mid-course and final).

Engineering majors had the highest mean Officer Basic Course final grade for the total sample. This group mean was also highest within the Combat Support branches and the Combat Service Support branches where significant differences were obtained.

Physical Science majors had the highest group mean on the Combat Leadership scale for the total sample, in the Combat Support branches, and in the Combat Service support branches. Business and Engineering majors had the highest group mean on the Technical Managerial Scale in the Total sample which was the only analysis on that scale that yielded significant results. However, even though those scales were meant to reflect the two major dimensions as defined by the work of Helme, Willemin and Grafton (1971), it should be noted that there are not any differences among the five groups of college majors in the total sample or within type of career branch on the overall rating of duty performance. It is interesting to note that only within the Combat Arms branches, in addition to the total sample, were differences found among groups on the Officer Efficiency Report scores.

The results indicate varying differences among the groups of college majors within the different kinds of branches, Combat Arms, Combat Support, and Combat Service Support in terms of duty performance. There are also differences among the groups on the basis of aptitude as measured by the Officer Evaluation Battery (OEB) and on interest in seeking a career as an Army officer as measured by the same instrument. The differences in aptitude and in interests will be taken into account in future research on the effect of college preparation on performance.

REFERENCES

Gilbert, A. C. F., & Grafton, F. C. Characteristics of an officer evaluation measure. Paper presented at the Annual Conference of the Military Testing Association, Gulf Shores, AL, October 18-22, 1976. In Proceedings, 18th Annual Conference of the Military Testing Association. Pensacola, FL: Naval Education and Training Program Development Center, 1976.

Helme, W. H., Willemin, L. P., & Grafton, F. C. Dimensions of leadership in a stimulated combat situation (Technical Research Report 1172). Arlington, VA: US Army Research Institute for the Behavioral and Social Sciences, July 1971. (NTIS No. AD-730 315)

Stogdill, R. M. Handbook of leadership; a survey of theory and research. New York: The Free Press, 1974.

Willemin, L. P. Criterion aspects of Army research on the prediction of officer performance (Research Study 65-6). Arlington, VA: US Army Research Institute for the Behavioral and Socaial Sciences, December 1965.

(356)

GOLDMAN, Lawerence A. Ph.D., and WORSTINE, Darrell A., US Army Military Personnel Center, Alexandria, Virginia.

WHY SOLDIERS ENLIST, REENLIST, AND SEPARATE (Mon P.M.)

An analysis was made of the relative importance of reasons why personnel joined the U.S. Army (concentrating on those items which may have influenced their actual enlistment decision) and the relationship of these reasons to their propensity to reenlist. This analysis focused on soldiers with six years or less of active service.

The relationship between reenlistment intention and final reenlistment decision was also ascertained. Having obtained a strong positive relationship, the relative importance of reasons why soldiers who definitely planned to reenlist decided to remain in the Army was analyzed. A similar study was made of the reasons why soldiers who definitely planned to separate (or retire) decided to leave the Army. The best correlates and mathematical "predictors" of reenlistment intent were also determined. This analysis of retention included personnel in all grades.

The overall study was based on data collected Armywide during late 1977 and early 1978 from a random sample of approximately 11,000 soldiers.

WHY SOLDIERS ENLIST, REENLIST AND SEPARATE [1]

OVERVIEW OF THE FINAL PHASE OF THE JOB SATISFACTION AND RETENTION PROJECT

Background. The US Army Military Personnel Center's (MILPERCEN) job satisfaction and retention project was previously described at the 19th annual conference of the MTA held in San Antonio, Texas in 1977 and at the 20th annual conference held in Oklahoma City, Oklahoma in 1978. The intent of today's presentation is to summarize the results from the third and final phase of this project. First, it examines the relative importance of reasons why soldiers enlist and the relationship of these reasons to their propensity to reenlist. Second, it examines the relationship between reenlistment intention and final reenlistment decision and covers the relative importance of reasons influencing reenlistment and separation/retirement decisions. The best correlates and mathematical predictors of reenlistment intent are also described.

Sample Composition. The above analyses were based on attitudinal data collected Armywide from a random sample of 10,877 soldiers in late 1977 and early 1978. Reliability checks of the sample composition uncovered a number of minor biases. Therefore, the data was weighted to reduce potential problems attributed to over/under representation by paygrade, sex and educational level. In examining reenlistment/separation reasons, career-force personnel were subdivided into two subsamples. Junior careerists were in paygrade E-6 or below, who had reenlisted once or twice, and generally had ten years or less of service. Senior careerists were in paygrade E-6 and above, who had reenlisted at least twice, and had more than ten years of service.

Methodology. With respect to enlistment, soldiers (excluding former draftees) were asked to rate 37 items in terms of importance to enlistment using an 8-point scale ranging from "DOES NOT APPLY" (equivalent to "UNIMPORTANT") to "EXTREMELY IMPORTANT". With respect to reenlistment, soldiers who had definitely decided to reenlist were asked to rate 36 items in terms of importance to reenlistment using the same 8-point scale. Soldiers who definitely decided to separate or retire at the end of their current enlistment or extension were asked to rate 42 items in terms of their importance to separation or retirement, again using this scale. With respect to the correlates and predictors of reenlistment intent, prediction was based on use of a combination of forward stepwise multiple linear regression and forward stepwise discriminant function analysis. Those independent variables which were examined included the following:

(1) 119 job attitude/satisfaction items (based on the results of a pilot test in early 1977 utilizing, to a large extent, the 348 item Occupational Attitude Inventory developed by the US Air Force's Human Resources Laboratory)

(2) Demographic variables thought to influence attitudes toward reenlistment (for example, age, hours worked per week, paygrade, and highest level of education attained)

---

[1] The views, opinions and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other official documentation.

FINDINGS AND DISCUSSION.

Importance of Enlistment Reasons. There was general agreement between first-termers and junior careerists with six years or less of service with respect to the influence on enlistment of the 37 factors considered.

(1) Factors impacting on individual development were consistently noted as having significantly contributed to enlistment decisions. All groups rated the chance for new experiences, civilian educational opportunities while in the Army, and the opportunity to learn a skill/trade for use in civilian life among the top five enlistment reasons. It is important to note that among personnel eligible for educational benefits under the GI Bill, this item was considered to have been the most important enlistment reason for all groups analyzed. On the other hand, those personnel who enlisted subsequent to 31 December 1976 and were therefore eligible for the Post-Vietnam Veterans' Educational Assistance Program (VEAP) considered this program, as then structured, to have been one of the least important reasons influencing their decision to enlist.

(2) With the exception of those factors which applied primarily to personnel possessing specific skills, abilities or aptitudes (e.g., enlistment options for Army Bands, Communications Command, Airborne, Combat Arms/Area of Choice) the following were generally considered to have been least important to the enlistment decision for all groups:

(a) Army advertising
(b) Influence of friends, neighbors, or relatives with Army experience
(c) Family history of Army or other military service
(d) Friends joining or already in the Army
(e) Army recruiters

While the above factors were considered to have been relatively unimportant to the actual enlistment decision, they were undoubtedly significant in the chain of events which led to this decision. Data supporting these findings are displayed in Table 1, page 5.

Relationship of Enlistment Reasons to Reenlistment Intent. The relationship between the reasons soldiers enlist and their propensity to reenlist was examined for first-termers overall and subgroups of first-term personnel less likely to attrite. It was found that soldiers who rated enlistment reasons associated with education/training opportunities and personal growth/self-development high generally intended to separate after their initial enlistment. Soldiers who rated Army advertising "extremely important" had a relatively high propensity to reenlist.

Relationship of Reenlistment Intent to Actual Reenlistment Decision. The relationship between the propensity to reenlist and the actual reenlistment decision was examined for those soldiers surveyed who had either reenlisted or separated/retired as of month-end June 1979. Among those who indicated that they planned to reenlist, 84.4 percent actually did so. For individuals who expressed the desire to separate or retire, 82.6 percent left the Army. It was also hypothesized that the relationship of reenlistment intent to reenlistment decision was stronger as the soldiers' Expiration of Term of Service

(ETS) date approached. Each sample was segregated into these three subgroups based on months until ETS: 6 months or less; 7-12 months; and 13-20 months. It was found that the closer to ETS date, the more certain it was that reenlistment intent corresponded to reenlistment decision. For example, 89 percent of all personnel within six months of ETS who definitely intended to reenlist actually did so compared to 82 percent of all other individuals. It was also found that first-termers (but not careerists) deciding to separate made this commitment considerably earlier than those who decided to remain in the Army. The mean number of months prior to ETS when first-termers intending to leave the Army was 18.8 compared to 14.6 months prior to ETS for first-termers who definitely intended to reenlist.

Importance of Reenlistment Reasons.

Considering only those soldiers who stated that they definitely planned to reenlist, there was general agreement among all personnel that the reasons which most influenced their decision to reenlist were associated with incentives and benefits. These included the following:

- Promotion chances
- Economic security
- The yearly pay adjustment to keep military pay comparable to civilian pay
- Thirty days of paid leave a year
- Medical care provided them and their dependents by the Army
- Dental care provided them by the Army

In addition, the ability to retire with 20 years of service was the most important reason why career-force personnel decided to reenlist. While first-termers rated getting the reenlistment option they wanted as highly important, this was not true for junior careerists and in fact was among the least important reasons associated with senior careerists. This may be more a result of opportunity among careerists for certain options (or options designed specifically for these soldiers) than their possible utility as reenlistment motivators. Soldiers who decided to reenlist were also motivated to do so, in part, because they were satisfied with their job and viewed serving the country through their work as being important. Data reflecting these findings are shown in Table 2, page 7.

Importance of Separation/Retirement Reasons. While there was general agreement concerning factors influencing reenlistment, this was not the case with regard to the separation/retirement decision. The data indicated that the most important reasons in this regard for first-termers and, to a lesser extent, junior careerists were associated with Army life and interpersonal relationships. However, senior careerists were more likely to leave the service due to issues related to incentives and benefits. The most important factors comprising Army life included the following:

- Amount of harassment (they encountered) in the Army
- Living conditions (housing/barracks)
- Unit Morale
- Spouse's attitude toward their reenlisting
- Amount of extra duties

With respect to interpersonal relationships, first-termers and junior careerists intending to separate rated the people for whom they work and with whom they must associate as relatively important factors tied to their intent to separate.

Other influential reasons, not associated with Army life or interpersonal relationships, included the desire to use their GI educational benefits and the amount of "real work" (these soldiers believed) there is to do in the Army.

Senior careerists rated these incentives and benefits as important influences affecting their decision to retire:

- Medical care provided them and their dependents
- Promotion chances (also important to junior careerists)
- Pay (base pay plus tax free allowances for subsistence and quarters)
- Dental care provided them and their dependents by the Army

In addition, the frequency of family separations due to Army assignments and frequent overseas or isolated assignments were among the most important factors impacting on the decision of senior careerists to retire. Data regarding these findings are shown in Table 3, page 9.

Correlates and Predictors of Reenlistment Intent. Inasmuch as the relationship between reenlistment intent and actual reenlistment decision was very strong, there was a need to determine those factors which were the best correlates and/or would best predict attitudes toward reenlistment. Important correlates and/or predictors for first-termers and careerists included:

- How well Army life provides what they want
- The feelings they get from wearing the Army uniform, including wearing it in the civilian community
- Self-respect they get from being in the Army
- The importance of making a good record in the Army
- The number of months left in their current enlistment or extension
- The interest they have in going to work each day
- Serving their country through Army service

Factors identified as significant correlates and/or predictors for subgroups comprised only of married personnel included the following:

- Satisfaction with how their family regards their job
- How their job affects their behavior toward their family
- Their family's regard for their job
- The effect Army life has had on the way they raised their children
- The quality of social life available to a soldier's family

APPLICATIONS. Through the Army Occupational Survey Program (AOSP), data for use in conducting job and training analyses are collected for the Military Occupational Specialties (MOSs) associated with enlisted force personnel. Abbreviated lists of the job attitude/satisfaction items and reasons for reenlistment and separation/retirement judged most important to analysis are now included in the questionnaire booklets used in the AOSP. This will permit examination of job satisfaction and retention at the MOS level. In addition, a recommendation has been made that the Army incorporate reenlistment intent statements as aids to predicting losses by MOS.

(362)

TABLE 1 - IMPORTANCE OF ENLISTMENT REASONS TO FIRST-TERMERS OVERALL AND CAREERISTS WITH SIX YEARS OR LESS OF SERVICE

CATEGORY/REASON	FIRST-TERMERS		CAREERISTS	
	MEAN	RANK	MEAN	RANK
1. EDUCATION/TRAINING OPPORTUNITIES				
a. GI Educational Benefits (Available before January 1977)[a]	5.88	1	5.67	1
b. To learn a skill/trade to use in civilian life	5.22	3	5.17	5
c. Civilian educational opportunities while in the Army	5.20	4	5.35	3
d. Post-Vietnam Veterans' Educational Assistance Program (VEAP)	3.03[a]	32	N/A	N/A
2. PERSONAL GROWTH/SELF-DEVELOPMENT				
a. Chance for new experiences	5.24	2	5.36	2
b. To give yourself time to think about your future	4.98	5	4.99	8
c. Chance for travel	4.97	6	5.26	4
d. Chance for adventure	4.73	7	4.79	10
e. Chance for a career with good promotion possibilities	4.61	10	5.12	6
f. Need to grow-up, learn self-discipline, and achieve independence	4.35	13	4.30	18
g. Need for a job	4.27	14	4.38	15
h. To compare Army life with civilian life	3.61	24	3.64	23
i. Personal problems	3.41	26	3.53	24
3. PATRIOTISM				
a. To serve the United States in some way	4.66	8	5.02	7
b. To become a soldier	3.73	22	4.05	22
4. ARMY PAY AND BENEFITS				
a. Army pay	4.62	9	4.93	9
b. Army dental care	4.38	11	4.61	13
c. Army medical care	4.36	12	4.67	12
d. Va. housing benefits	4.01	18	4.72	11
e. Army leave policy	3.87	20	4.08	21
f. PX privileges	3.74	21	4.12	19
g. Commissary privileges	3.71	23	4.10	20

[a]Data for eligible soldiers

(Continued on next page)

TABLE 1 - IMPORTANCE OF ENLISTMENT REASONS TO FIRST-TERMERS OVERALL AND CAREERISTS WITH SIX YEARS OR LESS OF SERVICE (cont.)

CATEGORY/REASONS	FIRST-TERMERS		CAREERISTS	
	MEAN	RANK	MEAN	RANK
5. ENLISTMENT OPTIONS/INCENTIVES				
a. Training of Choice Enlistment Option	4.20	15	4.35	17
b. Army area/station of Choice Enlistment Option	4.17	16	4.49	14
c. Unit of Choice Enlistment Option	4.06	17	4.36	16
d. Delayed Entry Program	3.88	19	3.33	28
e. Cash Bonus Enlistment Option[a]	3.12	28	3.49	25
f. Combat Arms Unit/Area of Choice Enlistment Option[a]	3.03	33	3.06	31
g. Airborne Enlistment Option[a]	2.57	34	2.61	33
h. Communications Command Enlistment Option	2.49	35	2.61	35
i. Army Bands Enlistment Option	2.26	37	2.31	36
6. INFLUENCE OF FAMILY/FRIENDS				
a. What you learned about the Army from friends, neighbors or relatives with Army experience	3.39	27	3.34	27
b. Friends joining the Army or already in the Army	3.11	29	2.98	32
c. Your family's history of Army or other military service	3.10	30	3.22	29
7. ARMY RECRUITING				
a. The recruiter you talked to	3.41	25	3.39	26
b. Army advertising	3.04	31	3.11	30
c. To avoid being drafted[a]	2.38	36	2.61	34

[a]Data for eligible soldiers

TABLE 2 - IMPORTANCE OF REENLISTMENT REASONS TO FIRST-TERMERS, JR CAREERISTS AND SR CAREERISTS

CATEGORY/REASON	FIRST - TERMERS		JUNIOR CAREERISTS		SENIOR CAREERISTS	
	MEAN	RANK	MEAN	RANK	MEAN	RANK
1. PATRIOTISM						
a. Serving the United States	6.17	2	6.22	8	6.38	9
2. INCENTIVES AND BENEFITS						
a. Your chance for promotion	6.23	1	6.56	2	6.70	2
b. Getting the reenlistment option you wanted	6.16	3	5.72	17	3.94	33
c. The yearly pay adjustment to keep military pay comparable to civilian pay	6.03	4	6.39	4	6.68	3
d. Thirty (30) days of paid leave a year	6.01	5	6.17	10	6.37	10
e. Economic security	5.98	6	6.19	9	6.45	8
f. Medical care provided you by the Army	5.98	7	6.35	6	6.46	7
g. Medical care provided your dependents by the Army[a]	5.95	9	6.44	3	6.60	6
h. Dental care provided you by the Army	5.87	10	6.13	11	6.34	11
i. Being able to retire with 20 years service	5.82	12	6.81	1	7.06	1
j. Your base pay	5.74	13	6.27	7	6.60	5
k. Dental care provided your dependents by the Army[a]	5.61	15	5.87	14	5.91	14
l. Commissary privileges	5.00	25	5.44	22	5.70	19
m. Availability of Selective Reenlistment Bonus (SRB)	4.92	27	3.96	35	2.68	36
n. PX privileges	4.74	29	5.07	26	5.10	25
o. Chance for Special Duty proficiency pay (such as Airborne, EOD)	3.99	36	3.43	36	3.14	35
3. INTERPERSONAL RELATIONSHIPS						
a. People for whom you work	5.41	17	5.51	20	5.65	20
b. People with whom you work	5.30	19	5.48	21	5.80	18
c. Morale in your unit	5.28	20	5.53	19	5.62	21
d. People with whom you must associate	5.26	21	5.33	23	5.46	22
e. Attitudes of your co-workers and friends within the Army	4.94	26	5.04	27	5.06	26

(Continued on next page)

TABLE 2 - IMPORTANCE OF REENLISTMENT REASONS TO FIRST-TERMERS, JR CAREERISTS AND SR CAREERISTS (cont.)

CATEGORY/REASON	FIRST - TERMERS		JUNIOR CAREERISTS		SENIOR CAREERISTS	
	MEAN	RANK	MEAN	RANK	MEAN	RANK
4. MOS/WORK RELATED						
a. Being satisfied with your job	5.96	8	6.35	5	6.62	4
b. Challenge and demands of your job	5.70	14	5.91	13	6.29	12
c. Amount of "real work" there is to do in the Army	5.53	16	5.72	16	5.89	16
d. Chance to work in your Primary MOS	5.01	24	5.58	18	5.86	17
e. Number of hours you must work for the Army	4.89	28	4.88	28	4.71	28
f. Chance of combat exposure	4.28	34	4.04	34	3.81	34
5. ARMY LIFE						
a. Attitude of your wife/ husband toward your reenlisting[a]	5.85	11	5.99	12	6.04	13
b. Your living conditions (housing/barracks)	5.39	18	5.82	15	5.91	15
c. Discipline in your unit	5.09	22	5.22	24	5.38	23
d. Amount of busy work	5.03	23	5.18	25	5.15	24
e. Frequent overseas or isolated assignments	4.46	30	4.69	30	4.63	29
f. Amount of harassment in the Army	4.46	31	4.57	31	4.62	30
g. Amount of extra duties	4.39	32	4.50	32	4.47	31
h. Army haircut policy	4.35	33	4.22	33	3.98	32
i. Frequency of family separations due to your Army assignments	4.25	35	4.74	29	4.87	27

[a]Data for eligible soldiers.

TABLE 3 - IMPORTANCE OF SEPARATION/RETIREMENT REASONS TO FIRST-TERMERS, JR CAREERISTS AND SR CAREERISTS

CATEGORY/REASON	FIRST - TERMERS		JUNIOR CAREERISTS		SENIOR CAREERISTS	
	MEAN	RANK	MEAN	RANK	MEAN	RANK
1. ARMY LIFE						
a. Amount of harassment in the Army	5.64	1	5.65	1	4.25	15
b. Your living conditions (housing/barracks)	4.98	3	4.73	8	4.45	12
c. Morale in your unit	4.69	4	4.98	2	4.30	13
d. Attitude of your wife/ husband toward your reenlisting[a]	4.62	6	4.73	7	4.17	18
e. Amount of extra duties	4.53	7	4.61	14	3.71	28
f. Army haircut policy	4.42	10	4.37	21	3.36	31
g. Irregular duty hours	4.38	12	4.64	12	3.86	25
h. Discipline in your unit	4.36	14	4.54	15	4.12	20
i. Amount of busy work	4.35	15	4.64	13	3.97	24
j. Amount of concern for personal appearance	4.20	19	4.31	23	4.11	21
k. Frequency of family separations due to your Army assignments	3.78	24	4.69	11	4.91	6
l. Frequent overseas or isolated assignments	3.74	27	4.42	19	4.87	8
2. INTERPERSONAL RELATIONSHIPS						
a. People for whom you work	4.64	5	4.84	5	4.47	11
b. People with whom you must associate	4.52	8	4.71	10	4.21	16
c. People with whom you work	4.21	18	4.40	20	4.11	22
d. Attitude of your co-workers and friends within the Army	4.11	20	4.16	26	4.19	17

(Continued on next page)

TABLE 3 - IMPORTANCE OF SEPARATION/RETIREMENT REASONS TO FIRST-TERMERS, JR CAREERISTS AND SR CAREERISTS (cont.)

CATEGORY/REASON	FIRST - TERMERS		JUNIOR CAREERISTS		SENIOR CAREERISTS	
	MEAN	RANK	MEAN	RANK	MEAN	RANK
3. INCENTIVES AND BENEFITS						
a. To use GI educational benefits[a]	5.55	2	4.94	3	4.77	9
b. Your pay (base pay plus tax free allowances for subsistence and quarters)	4.24	16	4.51	18	5.10	4
c. Your chance for promotion	4.22	17	4.92	4	5.26	2
d. Medical care provided you by the Army	3.92	21	4.30	24	5.19	3
e. Medical care provided your dependents by the Army[a]	3.86	22	4.52	17	5.43	1
f. Dental care provided you by the Army	3.78	25	3.81	28	4.90	7
g. Dental care provided your dependents by the Army[a]	3.59	29	4.21	25	5.05	5
h. Commissary privileges	3.07	33	3.56	29	4.52	10
i. PX privileges	2.94	38	3.24	33	4.04	23
4. MOS/WORK RELATED						
a. Amount of "real work" there is to do in the Army	4.44	9	4.83	6	4.26	14
b. Having a job which did not challenge your abilities & training	4.41	11	4.72	9	3.60	29
c. Number of hours you work for the Army	4.37	13	4.53	16	4.14	19
d. Time spent working outside your Primary MOS	3.82	23	3.83	27	3.32	32
e. Don't like my MOS and can't arrange to get one I do like	3.68	28	4.32	22	2.80	36
f. Chance of combat exposure	2.97	36	3.10	36	3.05	34
g. Reclassified into an MOS that I have no interest in and don't enjoy working in	2.38	42	2.57	40	2.26	38

(Continued on next page)

TABLE 3- IMPORTANCE OF SEPARATION/RETIREMENT REASONS TO FIRST-TERMERS, JR CAREERISTS AND SR CAREERISTS (cont.)

CATEGORY/REASON	FIRST - TERMERS		JUNIOR CAREERISTS		SENIOR CAREERISTS	
	MEAN	RANK	MEAN	RANK	MEAN	RANK
5. PERSONAL GOALS ATTAINED						
a. Joined the Army for new experiences and I've done that	3.78	26	3.49	30	3.72	27
b. Joined the Army to find myself/grow-up/mature and I've done that	3.56	30	3.36	31	3.02	35
c. Joined the Army to travel and I've done that	3.22	31	3.30	32	3.74	26
d. Found a civilian job using the skills/training I've acquired in the Army	3.20	32	3.10	35	3.13	33
e. Joined the Army for adventure and I've done that	3.05	35	2.99	37	3.55	30
6. UNATTAINABLE PERSONAL GOALS						
a. Joined the Army for adventure and I haven't had any/or enough	3.05	34	2.80	38	1.93	39
b. Joined the Army for new experiences and I haven't had any/or enough	2.96	37	2.69	39	1.91	40
c. Joined the Army to travel and I haven't	2.84	39	2.32	41	1.75	41
d. Not getting the reenlistment option you wanted	2.46	40	3.22	34	2.70	37
e. Joined the Army to find myself/grow-up/mature and found I couldn't do that in the Army	2.39	41	2.16	42	1.67	42

[a]Data for eligible soldiers

GOODGAME, D., Occupational Research Division Insustrial Engineering Department, Texas A&M University, College Station, Texas.

OPERATING AND ANALYTIC CAPABILITIES OF THE NEW CODAP SYSTEM (Wed A.M)

The Occupational Research Division at Texas A&M University (a user of CODAP since 1973) is redesigning and writing the computer software to replace the existing CODAP system which is dependent on IBM-AMDAHL computers. The resultant system will be relatively machine independent and with minimum conversion can be made to operate on any computer with a FORTRAN compiler. Results of a study are presented outlining how this CODAP system will operate. The results include sample outputs describing the various ways data can be formated for analysis and interpretation. The new capabilities of this system will be highlighted and integrated into a job-analytic investigation. The procedures for inputing control statements will be reviewed to illustrate the manner in which computer procedures are called and executed. The purpose of the four manuals used to operate and maintain the system also will be described.

OPERATING CHARACTERISTICS OF THE NEW CODAP SYSTEM 80

Doug T. Goodgame

Occupational Research Division
Industrial Engineering Department
Texas A&M University
College Station, Texas 77843

INTRODUCTION

Databases resulting from occupational surveys have three characteristics that make them somewhat unique. One, the databases tend to be very large. There may be several hundred to several thousand incumbent workers in a study with hundreds of observations recorded per worker. Two, the database contains two distinct types of data, worker profile or background data and task responses. Three, subsequent data processing and analysis tends to divide the original database into numerous groups of data each requiring separate study. This situation places unacceptable demands upon commonly used statistical packages and forces analysts, who use such packages to process occupational survey data, to restrict the scope and bounds of planned analysis. The Air Force recognized these problems in the 1960's and initiated the development of the CODAP system to process occupational survey data. CODAP is an acronym for Comprehensive Occupational Data Analysis Programs, which most of us are familiar with.

The present CODAP system was developed on a program by program basis over a number of years. Job analysts would recognize a need for displaying data in a specific fashion and have a programmer write a program to process it and display results in the desired manner. Not only did the number of programs grow, but eventually the programs had to be written for specific computers. As a result, there are three sets of CODAP programs in use. Each set functions on a different brand of computer, and reflects different operating capabilities.

As new programs for processing and displaying data were developed, and added to the present system, older programs were used less or eventually became obsolete. Still, the overhead for operating the system remained unchanged and continued to incorporate the older programs.

In addition, occupational analysis arrived at a point where a job analyst had to work in conjunction with a skilled programmer to produce minute changes in data processing and display formats. Also, in many instances, job analysts continued to rely on para-professionals to build CODAP control card decks to order routine runs. In essence, the system was dictating to the job analyst what the boundaries of an occupational analysis should be. Needless to say, the situation was ready for review.

The Navy is one of the major users of the CODAP system and in 1978, a representative of NODAC (Navy Occupational Development and Analysis Center) contacted the staff of the Occupational Research Division at Texas A&M University, who had extensive experience in working with the CODAP system, and requested that we present our views on this problem. After considerable

discussion, both parties agreed that the future of occupational analysis would be best served by a major rewrite of the system. It was agreed that the features of any new system should reflect these requirements:

- The system should be flexible: In that new processing and display requirements should be easily developed by persons without special training in programming.

- The system should be adaptive: The system should be able to process data other than relative time spent values without modification.

- It should be easy to use: Apprentice job analysts should be able to use the system to make routine runs without extensive training.

- All data in the system should be accessable: Any program or routine would be able to access any type of data in the database.

- The system should have high capacity: Data storage limits should be expanded to meet current demand.

The system should be transportable: The system should be operable on any main frame equipment with minimum modification.

In essence, any new system would, above all, have to be flexible and easy to use, without undue dependence on skilled programmers for operation. Using these criteria as a guide, ORD, in concert with a committee of CODAP users, decided an integrated database management system would be the best approach in redesigning and rewriting CODAP. A database management approach gives the job analysts the ability to access any data, manipulate it in preparation for processing, process it using a wide variety of computing routines and display results in easy to read forms to analyse work. In this manner, the job analyst thinks of the data as raw material for building data summaries. To do this, the job analyst uses English-like sentences to invoke the operation of certain procedures which pulls data from specific locations and processes for display. Key to understanding this approach is: 1) a knowledge of the conceptual arrangement of data in the computer, 2) knowledge of CODAP language statements that invoke data processing and reporting, and 3) knowledge of sample formats for displaying results. These are the three knowledge requirements which a job analyst needs in order to process occupational survey data for analysis. The remainder of this paper will address these three requirements.

CONCEPTUAL ARRANGEMENT OF DATA

Task inventory data in the new CODAP System 80 can be pictured as residing in a two dimensional matrix (see Figure 1). The items or variables from the task inventory are displayed down the left hand side of the matrix in task inventory order. Each variable in the task inventory is sequentially numbered with background variables using a prefix of H, task variables a prefix of T, and secondary variables a S. Incumbent workers are arrayed across the top of the matrix in the order which task inventory booklets are read in. Naturally, each incumbent's data is located in the respective column beneath the incumbent's identification number. After job clustering, the incumbents are resequenced and given a hierarchical sequence number which is a location identification

number that aids the analyst in referencing specific incumbents or groups of incumbents. All data in an occupational survey is arranged in this format for quick reference by the job analyst. The analyst should think of columns as containing an incumbent or an aggregate of incumbents known as a GROUP. The rows contain a variable or aggregate of variables known as MODULE. In a study, a job analyst will be working with numerous groups of incumbents and in regard to tasks, numerous modules. (A duty, as we know it, is a module of tasks.)

CODAP LANGUAGE

There are different categories of language statements which the job analyst uses to produce data summaries. Some statements locate the sets of data that will be extracted for study. Others define what will be done to the data identified for study. A third type of statement allows the analyst to print the summaries in various reporting formats. In review, the three major functions of the language are to: 1) select data for study, 2) perform computations on the data, and 3) print the data summaries. There are other functions, but these three are sufficient for an introduction to the system.

Let's review the major commands which job analysts can use to start processing data in CODAP System 80.

SELECT - The SELECT statement is used to identify subsets of rows (variables) or columns (incumbents) which the analyst wishes to isolate for study. SELECT will be used when the analyst needs to define a new module or group. It is mainly useful when it is necessary to carry out some computation on a subset of the data in the database or to identify rows or columns that meet certain criteria.

DESCRIBE - The DESCRIBE command produces statistical descriptions of existing data in the database. The DESCRIBE command is used primarily for generating descriptive analyses of incumbent responses to historical, task, and secondary questions. When given a group of incumbents and a module that defines a subset of the database, DESCRIBE can compute the percentages of incumbents that responded to the questions in the module and the averages and standard deviations of the incumbents' responses. Similarly, DESCRIBE is capable of computing statistics for modules on groups. In this case, analyses are computed for columns (incumbents), rather than rows (variables). This gives DESCRIBE the feature of symmetry.

VARSUM - The VARSUM procedure is used to generate reports of frequency counts and percentages of variable values on the database. The procedure also has the capability of producing two-way distributions of frequency counts and percentages.

AVALUE - AVALUE computes statistics for a specified row on a subset of the database. The subset of the database is defined by a group ID and a module ID. AVALUE produces a column that contains the statistics computed for the row.

CREATE - CREATE will be used whenever the CODAP user wants to add new data to the database that can be calculated from existing data via an arithmetic expression. Sometimes CREATE will be used as an intermediate step in a more complex calculation for which the CODAP system provides no predefined function to automatically carry out the desired computation. The ability to add data items to the database is one of the most powerful features of CODAP. This trait helps elevate CODAP from the role of specialized job analysis tool to that of multipurpose database management system and data analysis system.

PRINT - The PRINT statement displays information that exists in the database. In addition, various summary statistics are calculated and displayed optionally.

REPORT - REPORT displays information pertaining to variables that are within a CODAP database. The information printed for a variable contains the variable type (row, column, module, group, or constant) and the number of members in the variable (i.e., the length of the variable in computer words). If the variable is a created row or column, then REPORT will print the group or module that the variable was developed for. REPORT will optionally print the remark associated with the variable.

SAMPLE FORMATS AND LANGUAGE STATEMENTS

Let's examine how this language will be used to produce data summaries for analysis. We will assume a database from a sizeable occupational survey has been readied for production runs. One of the first data summaries an analyst produces in a study is a job description for a group of workers. There are two basic formats for producing job descriptions in CODAP System 80. One, a listing of tasks rank ordered on percent of members performing or on average percent of time as shown in Figure 2. The second format shows tasks grouped by duty field or module as we now call it. A most convenient way to do this is to order the modules on average percent time and rank order tasks in the module on percent members performing. The later format enables the analyst to refer to work time spent on modules when analysing work of a group across duty fields. Within the module, percent members performing is a more useful statistic for analysing performance at the task level. This job description format is located in Figure 3.

Language statements in CODAP System 80 are very easy to learn and give the job analyst increased flexibility in processing occupational survey data. These statements use English-like words complete with objects, verbs and prepositions in sentence structure to create a correct syntax invoking System 80 routines. We believe that a job analyst can learn this language in a workshop of not more than thirty to forty hours.

In Figure 4, we can examine a language statement for ordering a job description as described in Figure 2. In Figure 2, the job description reports three vectors of data across tasks ordered on magnitude of average percent time spent by all. The statements would read as follows:

The BEGIN word delineates the beginning of a CODAP program, informs the interpreter of the name of the database and the command to execute.

The next statement invokes the processing for a production run.

DESCRIBE is a command procedure that produces statistical descriptions of existing data in the database.

ROWS is a keyword that points to the portion of the database on which statistics will be computed.

FOR is a preposition identifying the object (G146).

The number in parenthesis (G146) refers to a system created group produced by job clustering which forms the set of incumbents on which statistics will be computed across.

ON is another preposition to illustrate the objective nature of the statement.

The word TASKS specifies the type of variables. In this instance, the word TASKS signals the interpreter that the processing will occur across all tasks which this group of incumbents reported performing.

PCNTG146 is the title of the first column of data to be reported.

:= is an assignment operator.

PCNT is a statistical function name. It defines the statistic that DESCRIBE procedure will compute.

The statement in quotes in any remark the analyst wishes to attach to this data.

In the second CODAP System 80 statement, a similar sentence invokes processing for computing and reporting average percent time spent by members performing and the third statement does the same for time spent by all.

The last statement is the command procedure for printing this information.

The general form of the PRINT statement is as follows:

- the procedure, PRINT.
- a description of which part of the database is being used to define the vertical axis.
- a description of which part of the database is being used to define the horizontal axis.
- a description of what is to be printed as a title at the top of the produced report.
- various options that define operations to be performed or the displayed information that control the appearance of the output.

PRINT is the command procedure.

ROWS (Tasks) identifies the part of the database which defines the vertical axis of the printout. Anything occurring before the slash (/) indicates elements of vertical axis.

COLUMNS (PCNTG146, AVGPG146, AVGAG146) defines the horizontal axis of the printout.

SORT DESCENDING BY (AVGAG146) specifies which data the task listing is ordered on.

CUM (AVGAG146) identifies the vector which will be sequentially summed.

HEADING:= can contain any title you want to give a report.

END. End period specifies you are finished with the statement.

This statement will invoke a task level job description of the type described in Figure 1. If an analyst wants job descriptions such as this for more than one group, the group identification number is inserted in the first line of the statement.

FIGURE 1

SAMPLE DATA BASE

	I1	I2	I3	I4	I5	I6	I7
H1	1	2	1	1	1	2	1
H2	23	*	*	41	27	19	53
H3	2	11	16	19	3	1	30
H4	5	7	3	2	4	1	6
T1	11	0	0	11	24	64	36
T2	11	0	43	44	24	9	64
T3	22	20	57	0	18	9	0
T4	56	50	0	22	0	18	0
T5	0	30	0	22	35	0	0
S1	*	*	*	2	*	*	2
S2	1	*	3	2	1	1	1
S3	2	2	3	*	1	1	*
S4	1	2	*	2	*	2	*
S5	*	1	*	1	3	*	*

FIGURE 2

TASK LEVEL JOB DESCRIPTION

D-TSK	DUTY TITLE	PCNTG146	AVGPG146	AVGAG146
O 13	LOCATE ERRORS IN A PROGRAM	72.79	1.75	1.27
H 1	CONSULT WITH PEERS TO GAIN INFORMATION TO SOLVE A SPECIFIC PROBLEM	77.55	1.45	1.13
O 14	DESIGN PROGRAM MODULES	51.02	1.96	1.00
Q 9	READ TECHNICAL MANUALS	68.70	1.29	0.89
O 30	DEVELOP PROGRAM LOGIC FLOW	48.30	1.76	0.85
I 33	OPERATE CONVERSATIONAL COMPUTER TERMINAL (OTHER THAN THE MAIN CONSOLE)	47.62	1.76	0.84
I 15	OPERATE CARD READER	53.06	1.55	0.82
P 15	DEVELOP TEST SITUATIONS FOR NEW OR REVISED PROGRAMS	59.18	1.34	0.79
P 11	TEST INDIVIDUAL PROGRAMS FOR ACCURACY AND EFFICIENCY REQUIREMENTS	52.38	1.50	0.78
B 24	READ OFFICE MEMOS AND LETTERS	74.15	1.03	0.76
P 12	VERIFY ACCURACY OF INPUT/OUTPUT INFORMATION FOR COMPUTER PROGRAM	56.46	1.34	0.75
D 4	INSTRUCT USERS CONCERNING CONCEPT AND USE OF A SYSTEM	55.10	1.29	0.71
O 31	WRITE PROGRAMS IN FORTRAN	49.66	1.44	0.71
O 34	WRITE PROGRAMS IN COBOL	39.45	1.79	0.70
D 15	INSTRUCT EMPLOYEES ON-THE-JOB	57.82	1.21	0.70
O 1	WRITE PROGRAMS TO INTEGRATE NEW SOFTWARE INTO SYSTEM	40.14	1.73	0.70
H 2	DETERMINE FEASIBILITY OF IMPROVING ANY PORTION OF A PROGRAM OR SYSTEM	62.58	1.08	0.68
P 6	IDENTIFY POSSIBLE PROBLEMS IN MODIFICATION OF A SYSTEM OR PROGRAM	53.74	1.25	0.67
Q 5	REVIEW USER COMPLAINTS	64.62	1.02	0.66
N 14	SORT PRINTOUTS FOR FILING	29.93	2.20	0.66
G 24	CONFER WITH CO-WORKERS ON SYSTEM DEVELOPMENT PROBLEMS	57.14	1.14	0.65

FIGURE 3

MODULAR JOB DESCRIPTION

ODULE-O CREATING PROGRAMS

	TASKS	PCNTG146	AVGA146
13	LOCATE ERRORS IN A PROGRAM	72.79	1.27
14	DESIGN PROGRAM MODULES	51.02	1.00
30	DEVELOP PROGRAM LOGIC FLOW	48.30	0.85
31	WRITE PROGRAMS IN FORTRAN	49.66	0.71
34	WRITE PROGRAMS IN COBOL	39.45	0.70
1	WRITE PROGRAMS TO INTEGRATE NEW SOFTWARE INTO SYSTEM	40.14	0.70
12	DESIGN TEST DATA TO TEST THE BRANCHES OF A PROGRAM	46.26	0.60
16	CREATE SIMULATED FILES OR TEST DATA	51.70	0.58
40	PERFORM PROGRAMMING FUNCTIONS WITH EASYTRIEVE	27.89	0.49
20	UPDATE ON-LINE DOCUMENTATION FILES	40.82	0.41

ODULE-P EVALUATING PROGRAMS

15	DEVELOP TEST SITUATIONS FOR NEW OR REVISED PROGRAMS	59.13	0.79
11	TEST INDIVIDUAL PROGRAMS FOR ACCURACY AND EFFICIENCY	52.38	0.78
12	VERIFY ACCURACY OF INPUT/OUTPUT INFORMATION FOR COMPUTER	56.46	0.75
6	IDENTIFY POSSIBLE PROBLEMS IN MODIFICATION OF A SYSTEM OR PROGRAM	53.74	0.67
19	RESEARCH TECHNICAL MANUALS TO DETERMINE CAUSE OF ABNORMAL TERMINATIONS	55.10	0.64
17	ANALYZE MEMORY DUMPS	51.70	0.53
18	EVALUATE PROGRAM TEST CASES	48.30	0.53
21	CONFER WITH AUTHOR OF A PROGRAM OR SYSTEM TO DETERMINE ITS FUNCTION	54.42	0.52
13	REVIEW ON-LINE DOCUMENTATION FOR ACCURACY	48.98	0.50
14	IDENTIFY FILES REQUIRED TO TEST NEW OR REVISED PROGRAMS	46.26	0.47

FIGURE 4

```
BEGIN    STUDY ID EXECUTE

DESCRIBE ROWS FOR (G146) ON TASKS
         PCNTG146:=PCNT 'PERCENT MEMBERS PERFORMING';
         ROWS FOR (146) ON TASKS
         AVGPG146:=AVGP 'AVERAGE PERCENT TIME SPENT BY
                         MEMBERS PERFORMING:
         ROWS FOR (G146) ON TASKS
         AVGAG146:=AVGA 'AVERAGE PERCENT TIME SPENT
                         BY ALL MEMBERS';

PRINT    ROWS (TASKS) / COLUMNS (PCNTG1461 AVGPG1461 AVGAG1461)
         SORT DESCENDING BY (AVGAG1461) CUM (AVGAG1461)
         HEADING:=SAMPLE JOB DESCRIPTION FOR PRESENTATION
         AT MTA.
         END.
```

GOTT, Sherrie P. Ph.D., and ALLEY, William E. Ph.D., Air Force Human Resources Laboratory, Manpower & Personnel Division, Brooks AFB, Texas.

PHYSICAL DEMANDS OF AIR FORCE OCCUPATIONS: A TASK ANALYSIS APPROACH
(Thu A.M.)

A four-stage research effort has been jointly undertaken by the Air Force Human Resources Laboratory and the Air Force Aerospace Medical Research Laboratory to develop and apply methodologies for assessing physical strength and stamina requirements in Air Force enlisted specialties. The ongoing first stage of the research involves surveying supervisory-level enlisted personnel to elicit global estimates of the physical demand posed by the more than 60,000 tasks performed Air Force wide. Follow-up surveys are administered to quantify the types of physical effort associated with each demanding task. Later stages of the research will focus on the development, evaluation, and validation of personnel diagnostics for measuring relevant dimensions of strength and stamina. Interim results from the survey of approximately 11,000 supervisors in 135 enlisted specialties will be reported in the following topical areas: interrater agreement, comparisons of type and level of physical demand across specialties, and distribution of specialties along a physical demand continuum.

PHYSICAL DEMANDS OF AIR FORCE OCCUPATIONS: A TASK ANALYSIS APPROACH

Sherrie P. Gott, PhD*
William E. Alley, PhD

Air Force Human Resources Laboratory, Manpower & Personnel Division

Brooks AFB, Texas 78235

INTRODUCTION

For several years, the Air Force Human Resources Laboratory (AFHRL) has been conducting a large-scale research program in the occupational requirements area. Primary emphasis in the past has been on determining aptitude requirements, with some secondary emphasis in the training area, but it is exploratory work in physical and perceptual/psychomotor requirements that is the precursor of the research effort I will be discussing today -- strength and stamina requirements for Air Force enlisted occupations.

Despite the general recognition that effective performance in a variety of AF enlisted specialties requires above average physical strength and stamina, there has been very little systematic research done to support definitive assignment criteria. The Surgeon General has instituted an interim screening program (Factor X) to guide assignments into heavy work specialties, but the system is highly judgmental both in the establishment of minimum requirements and the assessment of individual capabilities. In point of fact, AF managers continue to express concern that many of the current selection and assignment procedures are inadequate due to a lack of specificity, arbitrary distinctions, and the absence of a firm empirical basis. Such deficiencies are impeding the effective utilization of personnel resources at a time when the problems associated with maintaining an all volunteer force are reaching critical proportions.

There are likewise several important accession trends that bear relevance to the issue of effective utilization of personnel resources: First of all, it is estimated that over the next decade the number of qualified males available for entry into the armed forces will steadily decline. Concomitantly, the ratio of males to females entering military service is undergoing marked change. For example, between 1965 and 1978, the proportion of women in the armed services increased from 1.5% to 8.1%. By the mid 1980s, it is projected that women will represent 15% of military sevice personnel. Given these important trends, increasing reliance is being placed on the utilization of women in both traditional (medical, administrative, and clerical support) and nontraditional (mechanics, electronics, security services, and supply) assignments in order to offset anticipated difficulties in maintaining a viable workforce. The need to fully utilize the available pool of enlistees makes optimal assignment algorithms extremely important. And yet, since the inability to perform well in jobs involving heavy work applies to a certain proportion of men as well as women and because some women <u>are</u> physically capable of performing heavy work, sex group cannot be used as a definitive classification standard. Thus the problem is a general one, not a

*The views expressed herein are those of the authors and do not reflect the views of the United States Air Force and the Department of Defense.

gender-related one. The resolution, in our judgment, requires detailed analyses of work requirements in all enlisted specialties to (1) empirically derive the specific occupational demands and (2) guide the implementation of screening methods designed to insure that personnel capabilities meet or exceed on-the-job requirements.

AFHRL is pursuing this research in collaboration with the Aerospace Medical Research Laboratory (AMRL) at Wright-Patterson AFB. The broad objectives being pursued in this collaborative effort are (1) to determine the physical requirements for each Air Force enlisted specialty through survey methodology and followup field observation/validation at the task level, (2) to establish specialty-specific minimum entry standards based upon empirically derived job requirements, (3) to develop and validate a test battery to assess individual physical capabilities, and (4) to implement both the person and the job characteristics into the Air Force Person-Job Match System (PJM). PJM is the dynamic, computer-based Air Force assignment system that matches applicants for enlistment in the Air Force with Air Force occupations. AFHRL is taking the lead in the preliminary job analysis phase to assess occupational requirements, and AMRL will direct later phases to develop and validate appropriate screening procedures to measure physical capability at point of entry.

This paper reports preliminary findings regarding the assessment of task-level physical job requirements for 87 Air Force specialties. The reporting is interim in nature, and therefore no results bearing on the final determination of physical job standards will be presented. What can be reported, however, are some early trends regarding the frequency and distribution of high-demand tasks in selected specialties and some of the analysis options available to the researcher who adopts a task-level approach to job requirements research.

PROCEDURES AND MATERIALS

Overview of Approach

Task inventories developed by the Air Force Occupational Measurement Center (AFOMC) are being utilized in a two-stage assessment of all AF enlisted specialties. In the first stage, an exhaustive OMC-developed task inventory is used to elicit supervisory judgments about the overall physical demand requirements for each specialty-related task. In the second stage, multiple dimensions of the physically demanding tasks identified in the first stage are quantified with regard to type and level of effort by a second independent sample of enlisted supervisors. Information from the detailed assessment of high-demand tasks will be summarized in later reports.

Subjects

Two random samples of 60 supervisory-level NCOs each are selected to serve as subject matter experts for each specialty. To date, approximately 16,500 supervisors in 170 of the more than 230 enlisted specialties have been administered presurveys (first stage) and followup surveys (second stage). Thus far the average rate of return of usable surveys has been 64 percent.

The findings discussed herin are based on presurvey returns from the first 87 specialties studied. By career field, all aircrew operations, training devices, maintenance management systems, munitions and weapons maintenance, services, food services, fuels, supply, printing, security police, and aircrew protection specialties are included. In addition, specialties from the following career fields are among the 87: intelligence, command control systems operations, communications operations, communications-electronics systems, missile electronic maintenance, avionic systems, wire communications systems maintenance, intricate equipment maintenance, aircraft systems maintenance, aircraft maintenance, missile maintenance, vehicle maintenance, mechanical/electrical, structural/pavements, administration, and medical.

Presurvey Materials

The presurvey, or Physical Demands Survey, consists of two parts: the first is a Background section containing demographic items for the supervisor to complete, and the second is a comprehensive task list for the supervisor's specialty. The supervisory rater is asked to supply a single global estimate of the physical demand associated with each task using a 10-point (0-9) unidimensional scale. Anchors and lifting benchmarks are provided for all scale points providing for a transformation of the 0 to 9 values to 0 to 90 pounds. Verbal anchors range from "no significant demand" for the 0 point to "extremely heavy" for 9.

PRELIMINARY RESULTS

Results reported in this section are based upon responses from supervisors in the 87 specialties previously described. Of these 87 specialties, 28, or 32% have the highest Factor X designation of 1; 45, or 52%, have a middle Factor X rating of 2; and 14, 16%, have a low Factor X of 3. A disproportionate number of high-demand specialties is thus apparent in this sample; all findings should be interpreted in that context.

Reliability of Task Data

Reliability of the task-level physical demand estimates was assessed via an inter-rater agreement index (Lindquist, 1953). The range of reliability coefficients (R_{kk} values) for the 87 specialties was from -.714 to +.978, but as shown in Figure 1, the majority of coefficients fell in the +.800 to +1.000 range.

The range of R_{11} values, indexing the reliability associated with a single rater, was from -.015 to +.581. As shown in Figure 2, the majority of these coefficients fell in the +.200 to +.599 range. It is the R_{11} value that is entered in the Spearman-Brown prophecy formula to estimate the reliability of mean ratings based upon a sample of any given size. For the 70 specialties with R_{11} values in the four intervals from +.200 through +.500, sample sizes of approximately 30, 20, 15, and 10, respectively, would have been sufficient to achieve an R_{kk} of .900 assuming all raters rated all tasks. Since this is an untenable assumption, i.e., that all raters would have sufficient knowledge and experience to rate all tasks in a particular specialty, the requisite sample size would need to be augmented somewhat. With an average return rate of 64%, a sample of 60 raters per specialty, (38.4 usable returns) was adequate for providing stable results in most specialties.

FIGURE 1. DISTRIBUTION OF Rkk VALUES FOR 87 AIR FORCE SPECIALTIES

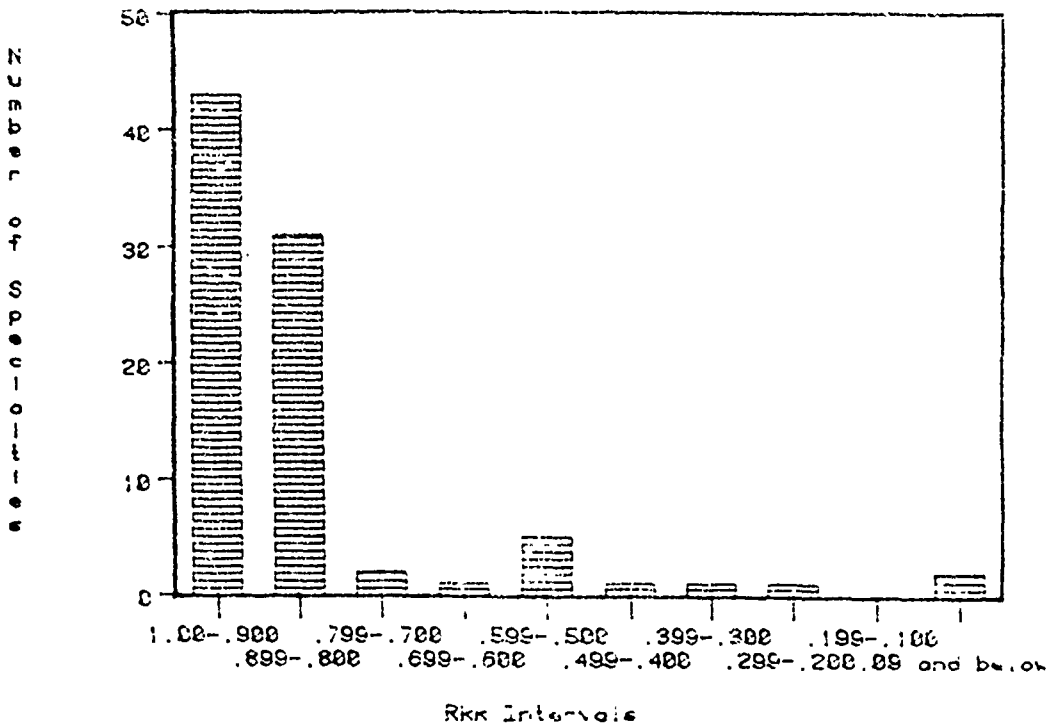
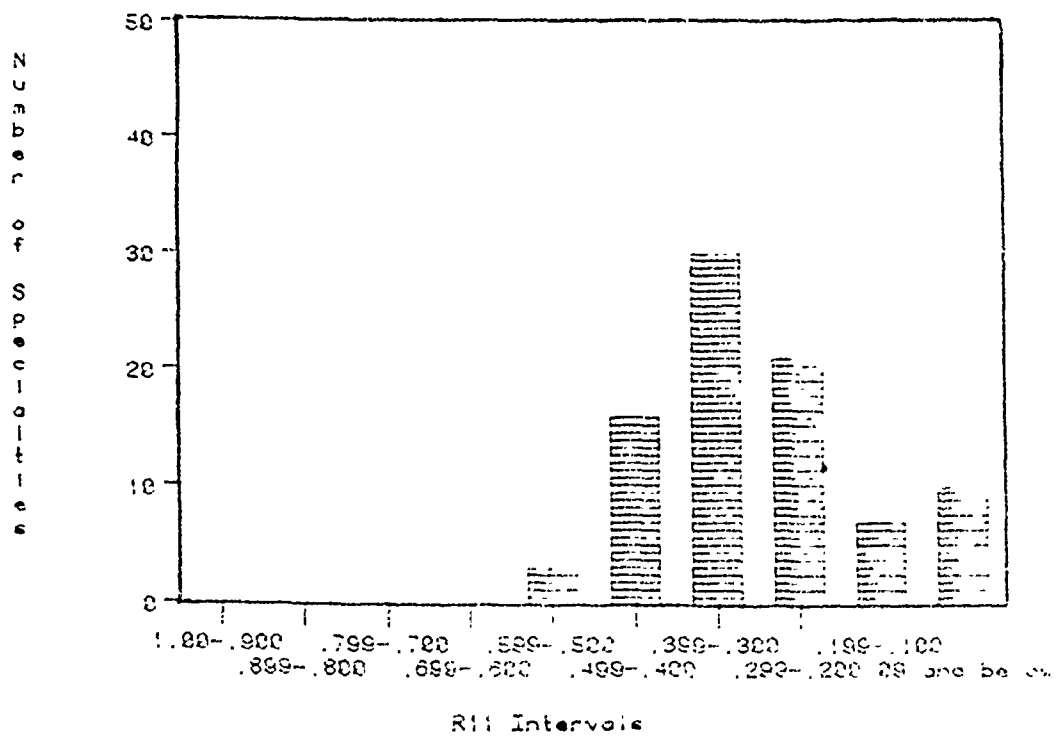


FIGURE 2. DISTRIBUTION OF R11 VALUES FOR 87 AIR FORCE SPECIALTIES



There were 13 specialties in which raters failed to agree at a minimal .800 R_{kk} level and 17 specialties where the R_{11} values were less than .200. The specialties falling below these arbitrary lower reliability boundaries (see Figures 1 and 2) were generally found to be low-demand specialties as indicated by their present Factor X ratings and the relatively low "means of all task means" that were achieved. We are presently studying all returns for these specialties and investigating background characteristics of raters as well as patterns and/or anomalies in responses that might account for the unacceptable level of agreement. One interesting finding thus far is that more than half of the task ratings produced for a large subset of these specialties are zero values. What's more, the proportion of zero ratings for some specialties is as high as 80%. Although this dominance of a single response option for a given specialty is in itself a convincing indicator of inter-rater agreement, reduced variability in ratings across tasks for these specialties serves to adversely affect the index of inter-rater agreement proposed by Lindquist (1953). We are presently researching the feasibility of alternate reliability estimates for these zero-loaded specialties.

Validity of Task Data

At this stage of the research, the validity of the Physical Demands Survey and the resultant task-level estimates can be addressed in terms of both content and construct validity[1]. A measure can be said to be content valid if it covers a representative sample of the behavior domain to be measured (Anastasi, 1976). Content validity can also be assessed in terms of the extent to which the operations measure the trait they are intended to measure as judged from the characteristics of those operations (Ghiselli, 1964).

In terms of the representativeness of the sample of behavior, the Physical Demands Survey contained an exhaustive list of all tasks performed in a given specialty as certified by the Air Force Occupational Measurement Center. Since the behavioral domain was occupational tasks, the universe was therefore included in the survey, not a sample thereof. Further assurance against task exclusion was provided by a write-in option where raters were encouraged to write in any physically demanding tasks not incorporated in the list. In the main, write-in tasks that have been provided are actually subcomponents of existing tasks expressed in more specific terms.

Concerning the content of the measurement operations and the relationship between those operations and the trait being measured, the case for content validity is as follows. The operations in this survey involved a rater's placing each occupational task along a 10-point scale according to the distinguishing characteristics or verbal anchors provided with the scale. The distinguishing characteristics were in fact concrete and behavioral in the form of lifting benchmarks ranging from 0 to 90 pounds along with accompanying verbal phrases that indicated frequency of physical effort. The case can be made that operations so defined are steadfastly connected to the trait of

---

[1]As of yet, there has been no attempt to determine if the estimated weights associated with the lifting tasks are comparable across specialties or to objectively verify the estimates with field observation data.

physical demand and that raters using such a set of operations are in fact employing the same model of physical demand if they can achieve high agreement in their task-level quantifications. In other words, high inter-rater agreement can be used to argue that no distinction exists between the definition of the trait and the set of measurement operations.

As a type of validation, construct validity is comprehensive in nature and logically subsumes the other types of validity generally considered, i.e., content and criterion-related validity (Anastasi, 1976). Conceptually, the construct validity of a measure is the extent to which it can be said to measure a theoretical construct or trait. In the test and measurement literature, convergent and discriminant validity are commonly reported as evidence of the construct validity of a particular measure or test. In the case of the Physical Demands Survey, several types of validity data have been derived thus far that approximate convergent validation. First, an interrelationship was assessed in a manner similar to the Campbell and Fiske (1959) monotrait-heteromethod process. Supervisory raters estimated the percentage of work done by first-term airmen that was heavy or very heavy in nature in addition to estimating the physical demand for each task in the inventory. A Pearson product-moment correlation was computed between the specialty means of such percentages and the average task difficulty per unit time spent by first termers for each specialty. The latter variable was computed on the basis of time-spent data obtained from routine administrations of occupational surveys by OMC by summing the products of task mean strength and stamina and time spent across all tasks within a specialty and then dividing by the number of tasks. A significant positive correlation was obtained [$r(85) = +.823$, $p < .01$]. Such a relationship between the same trait measured by differing methods provides some evidence of convergent validity in the instrument.

Secondly, if one considers the "theoretical predictions" associated with the construct being measured (i.e., physical demand), certain expectations exist regarding the kinds of occupational tasks that would logically be rated high on physical demand versus those that would logically be rated low. If the data support the theoretical expectations, then a case can be made for the measure's construct validity. For a large number of Air Force tasks, the physical requirements are common ones familiar to most people. It is therefore possible to inspect an ordering of such tasks on the dimension of estimated physical demand to see if, (1) the tasks one would expect to be high-demand tasks are in fact rated high, and (2) if a rank ordering of tasks on average demand confirms logical expectations. Table 1 shows represenatative high- and low-demand tasks for selected enlisted specialties. The most demanding task in the Fire Fighting Specialty, for example, was identified as rescuing personnel from buildings. This was judged to be equivalent to a simple lifting requirement of 84 lbs. A correspondingly low-rated task was "operate engine controls" rated at 17 lbs. In pavements/maintenance, "carry railroad track," rated as being equivalent to a lifting requirement of 80 lbs was judged to be most demanding. Although inspection reveals logical and consistent differences between tasks within specialties, the extent to which the ratings are comparable across specialties remains to be empirically determined. It is not known, for example, whether the two tasks with an average rating of 8.0 (80 lbs) in pavements/maintenance

Table 1
High- vs. Low-Demand Tasks in
Selected Specialties

Occupational Specialty	High-Demand Tasks		Low-Demand Tasks	
	Title	Mean Rating	Title	Mean Rating
Fire Protection	Rescue personnel from buildings	8.4	Operate engine controls	1.7
Pavements Maintenance	Carry railroad track	8.0	Assign vehicles or equipment to operators	.85
Survival Specialist	Fight forest fires	7.5	Operate pyrotechnic signaling devices	1.4
Security Police	Act as intruder/decoy in dog attack training	5.3	Conduct building security checks	1.3
Inflight Refueling	Secure cargo during preflight	6.0	Take inflight celestial observations	1.6
Telephone Repair	Install or remove aerial cable systems	6.4	Inspect telephone poles for climbing safety	1.4
Vehicle Maintenance	Remove or install grader hi-lo transmission	8.0	Clean or adjust automatic chokes	1.2
Masonry Specialist	Remove defective concrete with jack hammer	7.1	Clean tile surfaces	1.6
Aircraft Maintenance	Remove or replace landing gear	7.8	Ground aircraft	1.0
Carpentry Specialist	Overlay rolled roofing	6.1	Sharpen hand tools	1.3

Note: Range of scale values from 0 (no significant demand) to 9 (extremely heavy demand).

--"carry railroad track"--and vehicle maintenance--"remove or install grader hi-lo transmissions"--are exactly equivalent in terms of overall demand. Further assessment would reveal if systematic bias in the ratings requires statistical adjustment to achieve comparability across the various specialties.

Analysis Options Offered by the Task-Level Approach

Option 1: Task frequency profile. Table 2 presents a frequency and percentage distribution of task mean values for the initial 87 specialties surveyed. Each of the specialties is thus represented by a frequency profile derived from the overall estimates of task physical demand. Percentages are provided to facilitate comparisons across specialties.

The totals given at the bottom of Table 2 produce a markedly skewed distribution of tasks on the physical demand factor. Nonetheless, for these 87 specialties there are, according to supervisors' estimates, 1874 tasks for which first termers are required to manipulate an equivalent of 50 pounds or more. Table 3 shows the 15 specialties having the highest percentage of tasks means falling above 5 on the physical demand scale. Note the overepresentation of certain career fields in this rank-ordered list. There are four 55 (structural/pavements) specialties, two 54 (mechanical/electrical) specialties, two 46 (munitions and weapons maintenance) specialties, and two 47 (vehicle maintenance) specialties. On the basis of present Factor X designations, 11 of the 15 are predictable high-demand specialties since they are presently among the specialties having the highest Factor X rating of 1. However, 463X0, 472X0, 472X1A-D, and 545X0 have a lower Factor X rating of 2. As an analysis option, this approach affords the benefit of a profiled characterization of a specialty as opposed to a single index of physical demand. The format provides the opportunity for detecting the presence of outliers or high-demand tasks present in an otherwise nondemanding specialty. One deficiency in this option is the fact that quantity of tasks alone characterizes the specialty with no adjustment made for time spent on task or the percent of job incumbents who must perform the task. Option 2 offers some remedy for this deficiency.

Option 2: Average physical load index. Time spent on tasks is a frequently used moderating variable in job requirements research. By using historical occupational survey data it is possible to calculate an index that represents the "average load" on a particular group of job incumbents as a function of time spent on job tasks. For first termers in each specialty, an "average task strength and stamina per unit time spent" value can be calculated by summing the products of task mean strength and stamina and time spent across all tasks and dividing by the number of tasks. Table 4 shows a distribution of the 87 specialties on this time-weighted physical demand variable. Note the effect of time spent on the relative position of the 15 high-demand specialties identified in Option 1. Nine of the 15 retained a position in the top group, but six dropped out, several by 16 positions or more (472X0 from eleventh to twenty-seventh, 545X0 from fourteenth to thirty-second). What this comparison illustrates is the potency of the time spent factor. Given a restricted pool of enlistees for assignment to AF jobs, optimizing the person-job match might well entail considering the time spent in performing high-demand tasks across specialties.

Table 2
Task Mean Frequencies and Percent of Total Tasks
by Physical Demand Scale Interval

Air Force Specialty Code (AFSC)	Scale Interval 0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	Total
111X0	244(74%)*	51(15%)	24(7%)	12(4%)						331
112X0	114(55%)	63(30%)	20(10%)	8(4%)	1(0%)	2(1%)	1(0%)			209
113X0A/C	413(71%)	115(20%)	30(5%)	15(3%)	3(1%)	4(1%)	2(0%)	2(0%)		584
114X0	85(37%)	50(22%)	39(17%)	24(10%)	22(9%)	3(1%)	3(1%)	6(3%)		232
115X0	148(20%)	130(17%)	163(22%)	81(11%)	59(8%)	56(8%)	44(6%)	25(3%)	37(5%)	743
205X0	269(94%)	14(5%)	3(1%)	1(0%)						287
271X1	228(81%)	50(18%)	2(1%)							280
271X2	268(98%)	5(2%)	1(0%)							274
272X0 A/B/C/D	149(69%)	31(14%)	13(6%)	10(5%)	5(2%)	3(1%)	3(1%)	2(1%)		216
274X0	155(59%)	108(41%)	1(0%)							264
291X0	153(92%)	8(5%)	1(1%)	2(1%)			3(2%)			167
295X0	159(65%)	84(35%)								243
304X4	90(23%)	120(31%)	93(24%)	33(9%)	27(7%)	16(4%)	5(1%)			384
316X0F	175(34%)	211(41%)	59(12%)	49(10%)	13(3%)	2(0%)				509
316X0G/H	90(15%)	95(15%)	173(28%)	126(20%)	50(8%)	58(9%)	14(2%)	9(1%)	2(0%)	617
316X1F	110(28%)	57(15%)	67(17%)	85(22%)	50(13%)	20(5%)	4(1%)			393
316X2F	117(20%)	172(29%)	121(20%)	106(18%)	54(9%)	23(4%)	6(1%)			599
321X0K/L	51(15%)	99(29%)	111(33%)	44(13%)	12(4%)	13(4%)	5(1%)	3(1%)	1(0%)	339
322X2 A/B/C	160(26%)	201(33%)	161(26%)	55(9%)	19(3%)	11(2%)	8(1%)	1(0%)		616
325X1	87(12%)	309(44%)	207(30%)	56(8%)	19(3%)	18(3%)	1(0%)			697
328X0/A	37(17%)	101(47%)	41(19%)	21(10%)	12(6%)	5(2%)				217
328X1	213(28%)	354(47%)	105(14%)	50(7%)	19(3%)	11(1%)	4(1%)	1(0%)		757
328X3	44(17%)	63(24%)	74(28%)	33(13%)	23(9%)	17(6%)	10(4%)			264
328X4	60(39%)	23(15%)	30(20%)	23(15%)	11(7%)	4(3%)	2(1%)			153
341X1	338(66%)	109(21%)	39(8%)	15(3%)	4(1%)	5(1%)	4(1%)	1(0%)		515
341X2	384(74%)	82(16%)	29(6%)	8(2%)	6(1%)	3(1%)	4(1%)			516
341X3	348(61%)	128(22%)	50(9%)	22(4%)	9(2%)	11(2%)	5(1%)			573
341X4	462(62%)	189(26%)	44(6%)	27(4%)	8(1%)	5(1%)	6(1%)			741
341X5	270(38%)	319(44%)	71(10%)	32(4%)	12(2%)	7(1%)	5(1%)	2(0%)		718

(34)

Table 2-Continued

Air Force Specialty Code (AFSC)	Scale Interval 0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	TOTAL
341X6	627(74%)	142(17%)	43(5%)	19(2%)	5(1%)	6(1%)	8(1%)			850
341X7	370(58%)	151(24%)	62(10%)	26(4%)	14(2%)	5(1%)	5(1%)			633
361X0	16(5%)	46(13%)	47(13%)	56(16%)	45(13%)	60(17%)	61(17%)	19(5%)		350
361X1	42(11%)	97(24%)	123(31%)	88(22%)	36(9%)	7(2%)	6(2%)			399
362X3	148(39%)	154(40%)	57(15%)	15(4%)	5(1%)	4(1%)				383
362X4	96(46%)	63(30%)	22(11%)	16(8%)	5(2%)	4(2%)	1(0%)			207
391X0A/B	181(94%)	11(6%)								192
392X0	155(61%)	90(35%)	6(2%)	1(0%)		1(0%)	1(0%)			254
404X0	382(53%)	198(28%)	77(11%)	40(6%)	13(2%)	3(0%)	5(1%)	2(0%)		720
404X1	92(30%)	140(45%)	58(19%)	13(4%)	5(2%)					308
423X0	32(9%)	123(34%)	145(41%)	37(10%)	8(2%)	9(3%)	4(1%)			358
423X2	42(16%)	92(35%)	81(31%)	28(11%)	10(4%)	7(3%)	3(1%)	2(1%)		265
423X3	53(19%)	84(31%)	72(26%)	45(16%)	9(3%)	7(3%)	3(1%)	1(0%)		274
423X4	48(10%)	36(7%)	170(37%)	136(29%)	44(9%)	19(4%)	6(1%)	5(1%)		464
426X1	50(20%)	91(36%)	65(25%)	24(9%)	17(7%)	8(3%)				255
427X1	18(6%)	85(26%)	109(34%)	68(21%)	35(11%)	6(2%)	1(0%)			322
427X4	86(25%)	101(29%)	113(33%)	29(8%)	10(3%)	1(0%)	3(1%)			343
427X5	65(31%)	60(29%)	53(25%)	28(13%)	2(1%)					208
431X0C/D	68(10%)	164(24%)	186(27%)	137(20%)	66(10%)	36(5%)	24(3%)	6(1%)		687
431X1 A/C/E/F	91(15%)	240(38%)	177(28%)	53(8%)	26(4%)	18(3%)	15(2%)	6(1%)		626
443X0E	57(26%)	12(6%)	25(12%)	55(25%)	38(18%)	21(10%)	7(3%)	2(1%)		217
461X0	102(42%)	32(13%)	29(12%)	22(9%)	23(9%)	15(6%)	19(8%)	3(1%)		245
462X0	192(37%)	42(8%)	135(26%)	80(15%)	49(9%)	16(3%)	5(1%)	1(0%)		520
463X0	99(24%)	105(25%)	62(15%)	58(14%)	41(10%)	36(9%)	18(4%)	2(0%)		421
464X0	143(39%)	89(24%)	42(11%)	31(8%)	30(8%)	16(4%)	13(4%)	5(1%)		369
472X0	84(17%)	114(23%)	115(23%)	66(13%)	49(10%)	33(7%)	23(5%)	6(1%)		490
472X1 A/B/C/D	67(12%)	120(22%)	148(27%)	86(16%)	54(10%)	42(8%)	23(4%)	6(1%)		546
472X2	57(15%)	103(28%)	97(26%)	56(15%)	36(10%)	12(3%)	7(2%)	5(1%)		373
472X3	34(26%)	26(19%)	34(26%)	22(16%)	9(7%)	7(5%)	1(1%)			133
541X0F	154(35%)	133(30%)	108(25%)	27(6%)	4(1%)	5(1%)	6(1%)	2(0%)		439

Table 2-Continued

Air Force Specialty Code (AFSC)	Scale Interval 0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	Total
541X0G	71(16%)	78(17%)	169(37%)	67(15%)	29(6%)	21(5%)	11(2%)	5(1%)		451
542X1	65(17%)	42(11%)	76(20%)	66(17%)	53(14%)	28(7%)	34(9%)	14(4%)		378
545X0	52(17%)	118(38%)	65(21%)	31(10%)	9(3%)	15(5%)	14(4%)	6(2%)		310
547X0	27(9%)	112(37%)	83(28%)	34(11%)	16(5%)	17(6%)	7(2%)	2(1%)		298
551X0	8(2%)	11(3%)	87(24%)	97(26%)	75(20%)	50(14%)	31(8%)	6(2%)	1(0%)	366
551X1	10(2%)	23(6%)	44(11%)	150(37%)	107(26%)	48(12%)	23(6%)	2(0%)		407
552X0	39(13%)	41(13%)	53(17%)	59(19%)	63(20%)	39(13%)	13(4%)	2(1%)		309
552X1	63(17%)	43(12%)	84(23%)	57(16%)	66(18%)	35(10%)	13(3%)	2(1%)		363
552X4	35(22%)	48(30%)	46(28%)	23(14%)	7(4%)	3(2%)				162
571X0	107(25%)	110(26%)	64(15%)	51(12%)	47(11%)	30(7%)	10(2%)	6(1%)	3(1%)	428
611X0	34(17%)	110(54%)	32(16%)	12(6%)	10(5%)	4(2%)	2(1%)			204
611X1	1(1%)	6(7%)	38(45%)	19(22%)	12(14%)	6(7%)	2(2%)	1(1%)		85
622X0	17(8%)	67(33%)	68(34%)	31(15%)	7(3%)	10(5%)	1(0%)			201
622X1	208(52%)	121(30%)	48(12%)	17(4%)	2(1%)	2(1%)	1(0%)			399
631X0	65(33%)	63(32%)	31(16%)	15(8%)	20(10%)	2(1%)	1(1%)			197
645X0/A	152(65%)	52(22%)	16(7%)	5(2%)	2(1%)	3(1%)	3(1%)			233
645X1	90(55%)	21(13%)	13(8%)	16(10%)	9(5%)	7(4%)	7(4%)	1(1%)		164
645X2	182(82%)	33(15%)	6(3%)							221
705X0	168(88%)	22(11%)	2(1%)							192
713X0	74(39%)	110(59%)		2(1%)	1(1%)					187
713X1	161(90%)	15(8%)	1(1%)	2(1%)						179
713X2	87(37%)	142(60%)		6(3%)	1(0%)					236
811X0	66(36%)	71(39%)	34(18%)	10(5%)	3(2%)					184
811X0A/X2A	129(40%)	66(21%)	54(17%)	31(10%)	28(9%)	11(3%)	1(0%)			320
811X2	101(45%)	62(28%)	33(15%)	20(9%)	4(2%)	3(1%)	1(0%)			224
903X0	233(60%)	115(29%)	28(7%)	10(3%)	3(1%)	1(0%)	1(0%)			391
911X0	75(35%)	79(37%)	37(17%)	18(8%)	4(2%)		2(1%)			215
921X0	68(16%)	143(33%)	118(27%)	57(13%)	23(5%)	20(5%)	6(1%)	1(0%)		436
922X0	208(49%)	124(29%)	52(12%)	25(6%)	9(2%)	6(1%)	3(1%)	1(0%)		428
GRAND TOTALS	11,668 (36%)	8,333 (26%)	5,615 (17%)	3,231 (10%)	1,736 (5%)	1,062 (3%)	594 (2%)	174 (1%)	44 (0%)	32,457

*Frequency of task means (percent of total tasks in specialty)

Table 3
High-Demand Specialties Based on Percent
of Demanding Tasks

Air Force Specialty Code (AFSC)	Occupational Specialty	Percent of Tasks with Mean Physical Demand ≥ 5.0
1. 361XO	Cable & antenna systems installation/maintenance	39
2. 551XO	Pavements/maintenance	24
3. 115XO	Pararescue/recovery	22
4. 542X1	Electric power line	20
5. 551X1	Construction equipment operation	18
6. 552XO	Carpentry	18
7. 461XO	Munitions systems maintenance	15
8. 443XOE	Missile maintenance, LGM-25	14
9. 552X1	Masonry	14
10. 463XO	Nuclear weapons maintenance	13
11. 472XO	Base vehicle equipment maintenance	13
12. 472X1A-D	Special vehicle mechanics	13
13. 316XOG	Missile systems analysis, WS-133AM/CDB	12
14. 545XO	Refrigeration and air conditioning	11
15. 571XO	Fire protection	11

Table 4
Distribution of Specialties by "Average Task Strength & Stamina Per Unit Time Spent" Values

(Rank ordered within interval, low-high)

Scale Interval								
0-.5	.6-1.0	1.1-1.5	1.6-2.0	2.1-2.5	2.6-3.0	3.1-3.5	3.6-4.0	4.1-4.5
c71X2	341X1	713X0	328X0/A	304X4	423X2	423X4	542X1	551X0
713X1	705X0	541X0F	922X0	114X0	463X0	571X0	551X1	361X0
205X0	272X0A/B	903X0	911X0	921X0	461X0	427X1		
291X0	645X2	392X0	328X1	431X1A/C/E/F	811X0A/X2A	611X1		
391X0A/B	341X2	341X3	811X0	464X0	316X0G/H			
	113X0A/C	316X0F	362X4	325X1	431X0C/D			
	271X1	362X3	112X0	321X0K/L	472X3			
	111X0	404X1	322X2A-C	427X4	547X0			
	341X6	622X1	316X2F	426X1	462X0			
	274X0	404X0	631X0	427X5	115X0			
	295X0	341X5		423X3	443X0E			
	645X0	341X7		328X4	316X1F			
	341X4	713X2		423X0	361X1			
		811X2		472X2	552X1			
				272X0D	552X0			
				545X0				
				645X1				
				552X4				
				328X3				
				622X0				
				472X0				
				472X1A-D				
				541X0G				
				611X0				

Note: Full range of scale from 0 (no significant demand) to 9 (extremely heavy demand).

(597)

While this option does offer a refinement not present in Option 1, i.e., the time spent effect, it suffers, in our judgment because using a single index to characterize a specialty may mask the presence of outlying high-demand tasks. Option 3 is an elaboration of this option that reduces the concern about a singular physical demand index.

Option 3: Job description profile. Cumulative time spent across tasks can be computed to generate a job description for first-term airmen that identifies the minimum number of tasks accounting for a certain percentage of work time. In this way it is possible to profile the first termer's job using time spent to moderate the unweighted frequency distributions provided in Option 1. What results is a task-level representation of the major part of the job in terms of physical requirements.

This option offers the advantage of a profile to characterize the specialty as opposed to a singular index, while at the same time it allows the researcher to distinguish those specialties where first termers spend a disproportionate amount of time performing high-demand tasks.

Option 4: Percent members performing profile. While time spent is a job requirements factor of some consequence for physical job requirements research, percent members performing may be a more logical moderator. Unlike other job requirements, the physical demand of a job may be a property for which time spent is not as important a consideration as whether or not a task needs to be performed at all. If a high-demand task is performed by any first termers in the specialty--regardless of time spent on the task--the capability to perform should exist among job incumbents and the requirement to perform should be reflected in enlistment standards. As stated before, simple performance-nonperformance is a defensible task factor as far as physical demand is concerned. The implicit philosophy associated with this view is that one cannot ignore the requirement posed by a high-demand task even if the job incumbent spends only a trivial amount of time in task execution. This view becomes even more defensible when one considers the high-demand tasks of emergency-oriented specialties such as fire protection. The three highest ordered fire protection tasks on the physical demand factor involve rescuing personnel from buildings, aerospace vehicles, and motor vechicles. However, none of the three tasks ranks above the 115th position on a first termer job description ordered on time spent. In fact, only one of the three would be included in such a job description accounting for 75% of the first termer's time (Option 3). The consequences of allowing time spent to unduly influence a physical job requirements algorithm for such a specialty are weighty.

CONCLUSION

Further work is continuing on the initial assessment of overall task demands in the remaining Air Force specialties. The second stage of the data collection, planned for completion in 1981, is intended to further quantify the specific types and levels of effort associated with the high-demand tasks (i.e., lifting/lowering, pushing/pulling, carrying, torquing, etc.) Data from the second assessment, including information on weights, distances, duration, and frequency of task performance provides a further basis for objective validation of the summary ratings and a potential methodology for benchmarking the supervisory ratings across specialties if that becomes necessary.

In the final stages of this work, task demand characteristics would need to be aggregated at the specialty level as required for interfacing with the Person-Job-Match assignment system. Specifications for entry into each enlisted career field would then be sufficient to insure an optimum distribution of talent in a period when the Air Force can scarely afford to do less.

References

Anastasi, A. Psychological testing (4th ed.). New York: Macmillan, 1976.

Campbell, D.T., & Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Ghiselli, E.E. Theory of psychological measurement. New York: McGraw Hill, 1964.

Lindquist, E.F. Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin, 1953.

INSTRUCTIONAL SYSTEMS DEVELOPMENT

ALEXANDER M. GOTTESMAN

> The mission of a military instructional system is to determine instructional needs and priorities, to develop effective and efficient solutions to achieving these needs, to implement these solutions in a competent manner, and to assess the degrees to which the output of the system meets the specified needs.[1]

Instructional Systems Development (ISD) is an orderly approach to curriculum planning, implementation, and evaluation--based on careful analysis of the job to be performed, the duties and tasks of the job, and the elements that make up each task. The ISD process requires attending to individual differences in student abilities, achievements, motivation, and rates and styles of learning.

RELATIONSHIPS TO CURRICULUM DEVELOPMENT

Jobs, duties, tasks, and elements are carefully analyzed. To illustrate, figure 1[2] breaks down the job of a "hospital corpsman" into examples of duties such as "patient care," "emergency care," and "drug therapy." The duty of patient care is broken down into examples of tasks such as "take and record temperature, pulse, and respiration," "take and record blood pressure," "regulate intravenous flow," and "reinforce dressings." Finally, examples of elements leading to task mastery are noted, such as "demonstrate three methods of taking temperature" under the task of take and record temperature, pulse, and respiration; and "identify equipment needed for IV therapy" as an element of the task of regulate intravenous flow.

---

1. Interservice Procedures for Instructional Systems Development, NAVEDTRA 106A. "Executive Summary and Model." 1975, p 3.

2. Format for figures 1, 2, and 3 adapted from Interservice Procedures for Instructional Systems Development, NAVEDTRA 106A. " Phase I." 1975, pp 8-9.

Dr. Gottesman is Head, Curriculum Branch, Naval Health Sciences Education and Training Command, National Naval Medical Center, Bethesda MD 20014.

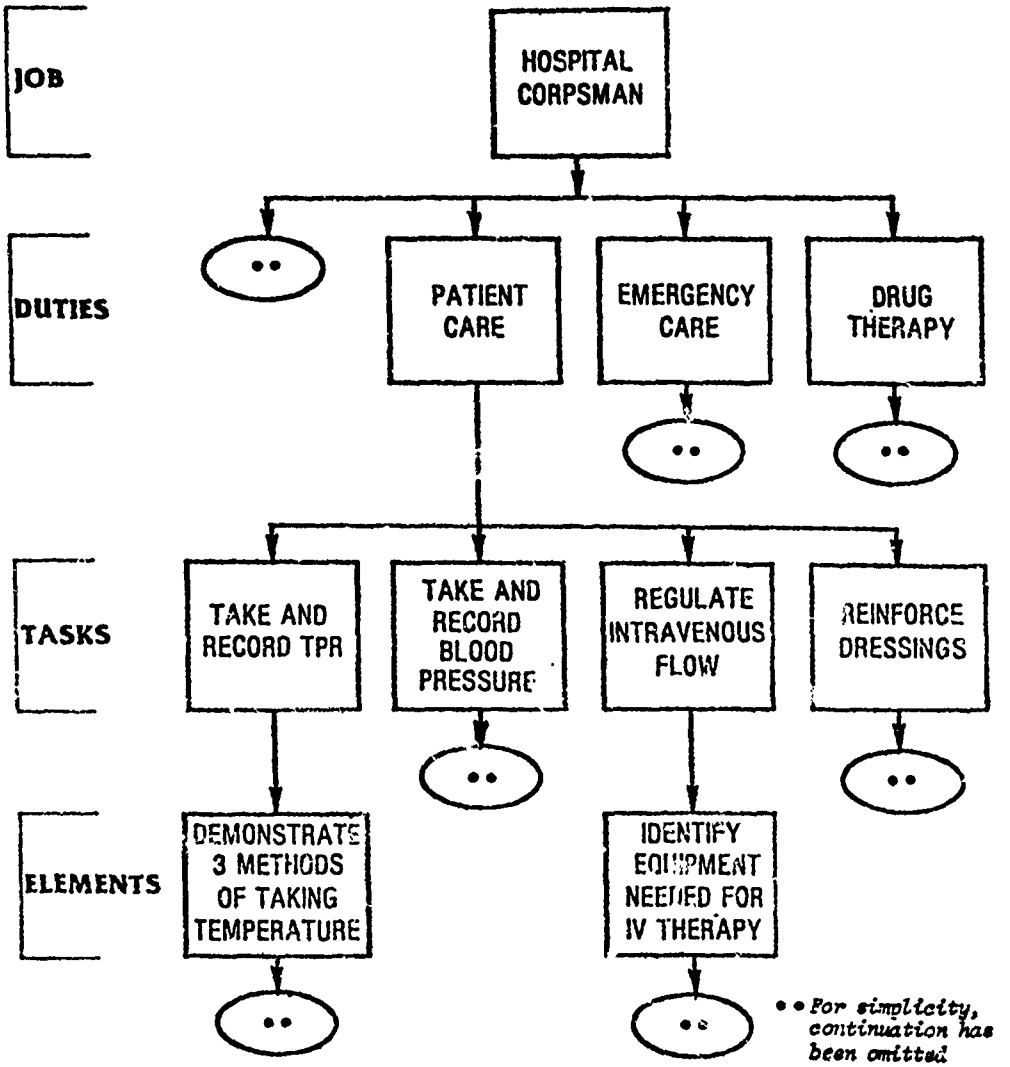


Figure 1. Relation of Job, Duties, Tasks, and Elements to Job Analysis.

Jobs, duties, tasks, and elements are also related to training mission and objectives as illustrated in figure 2. The job relates to "training mission," duties relate to "terminal objective clusters," tasks relate to "terminal objectives," and elements relate to "enabling objectives."

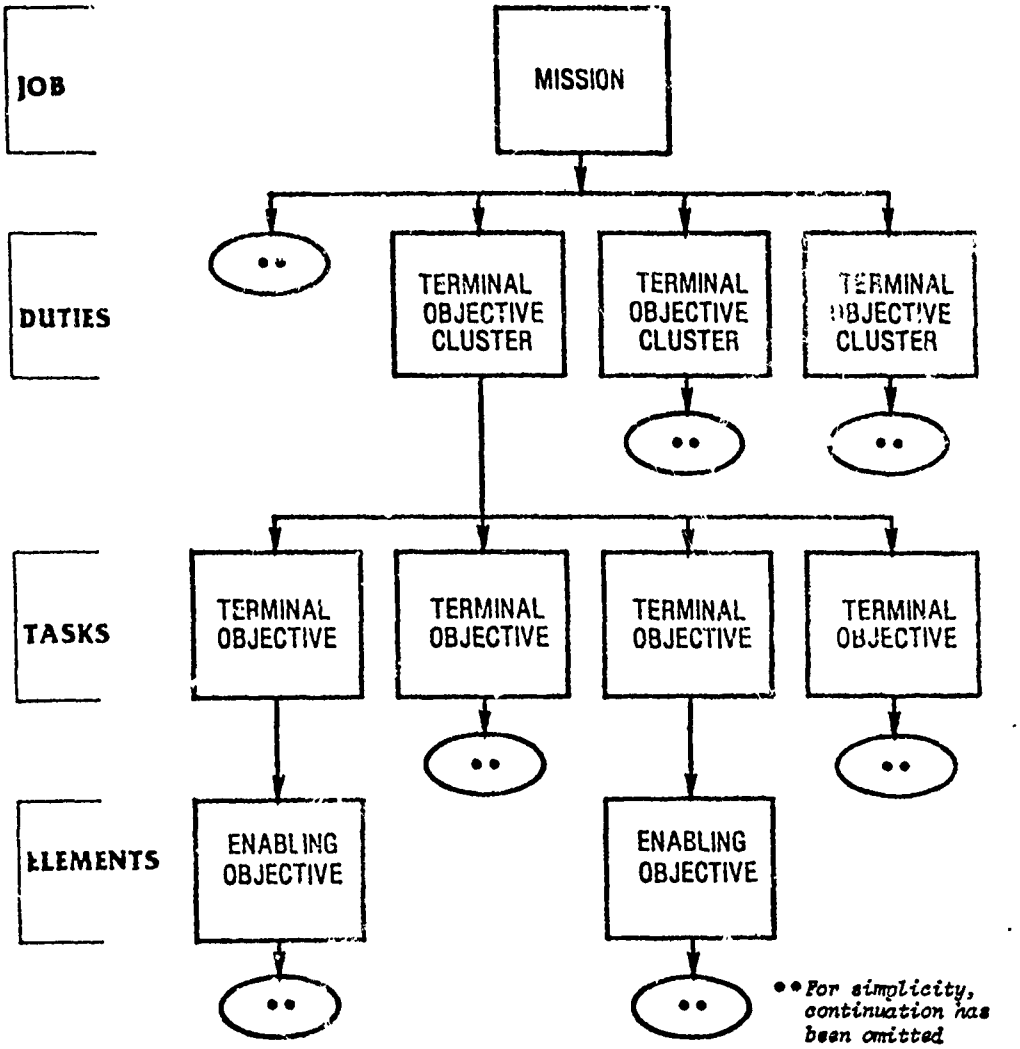


Figure 2. Relation of Job, Duties, Tasks, and Elements to Training Mission and Objectives.

Figure 3 illustrates the relationship of the job, duties, tasks, and elements to curriculum organization, in that the job determines the "course," duties determine the "units," tasks determine the "lessons," and elements determine the "events/activities." We recommend, for example, that one and only one task corresponds to a terminal objective, and there is one and only one terminal objective per lesson. A "lesson" may take 15 to 20 minutes or several days depending on how much instruction and how many learning activities are required for mastery of the terminal objective and ultimately the task.

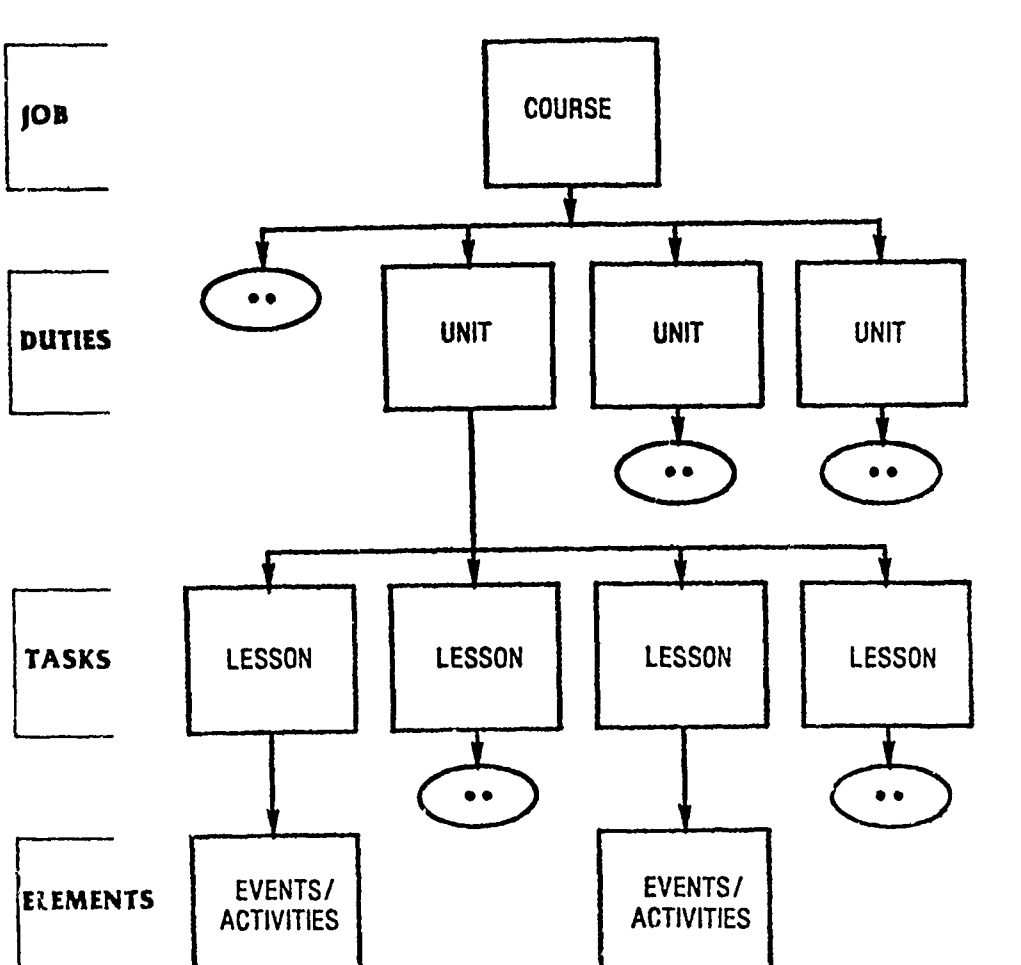


Figure 3. Relation of Job, Duties, Tasks, and Elements to Curriculum Organization.

PHASES AND STEPS IN THE ISD PROCESS

Five crucial phases and 19 interrelated steps of the ISD Model flowchart are identified in figure 4. The process would be undertaken as a result of the need to install a new system or procedure, the need for periodic updating of existing jobs, or the result of a needs assessment in any aspect of training. As with any needs assessment, a discrepancy may exist between what should be and what is. The flowchart begins with such a discrepancy as the entry to Phase I (Analyze).

K04

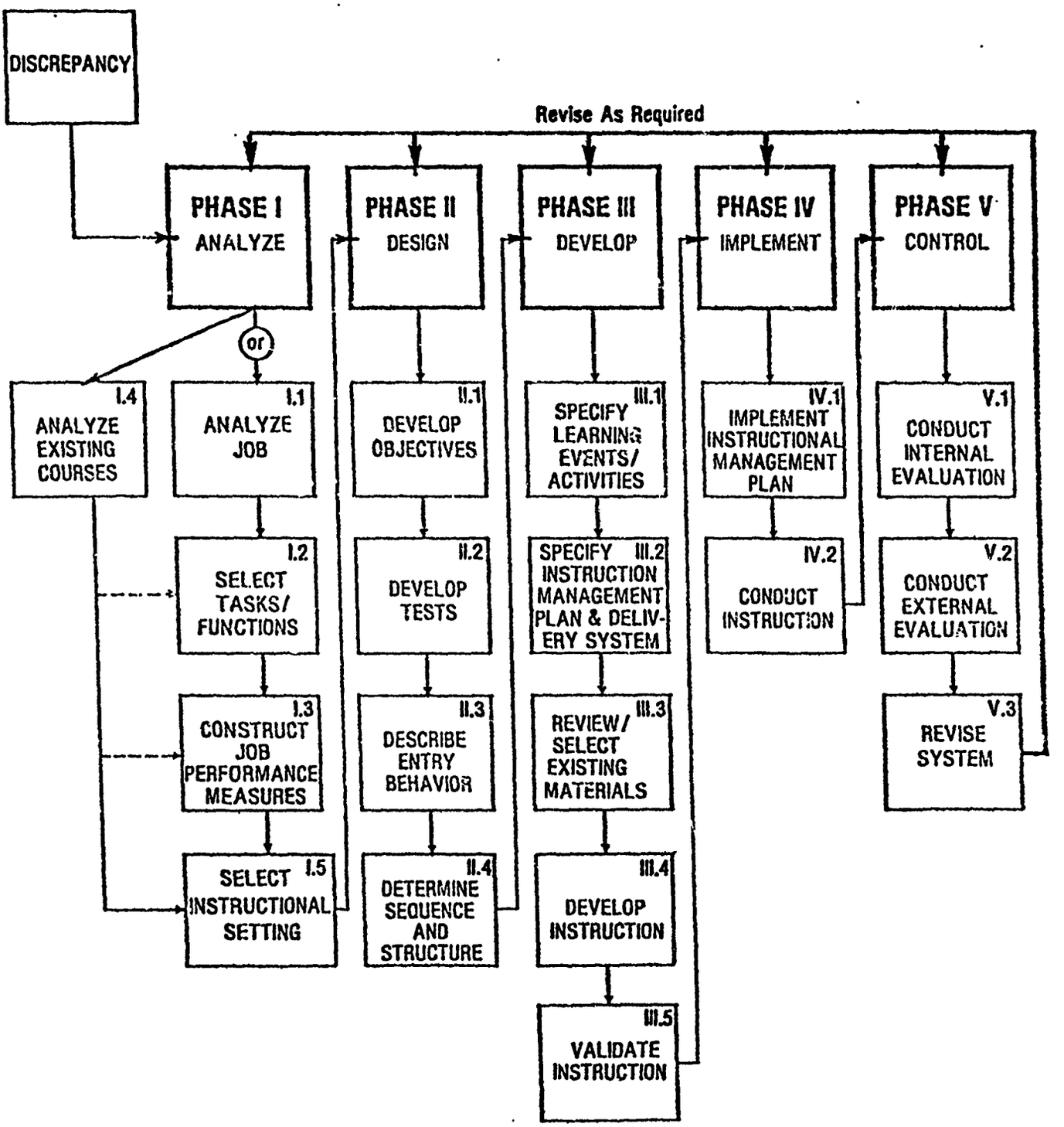


Figure 4. The US Navy ISD Flowchart.

In the US Navy's systems approach, curriculum development for any purpose, e.g., a new course, would logically begin with step I.1. The ISD model, however, permits entry at any level depending upon our knowledge of existing courses. We may begin with step I.4 (Analyze Existing Courses) and determine only minor revision in step I.5 (Select Instructional Setting) is necessary for Phase I. Further analysis of the design of an appropriate existing course could be approved with only the need to update materials of instruction--followed by the minor changes new material would require in developing and validating instruction.

Before explaining the 19 steps identified in figure 4, a brief definition of the remaining four phases is necessary. Following the Analyze Phase:

- Design the form and specifications for training based on input from Phase I.

- Develop, in turn from previous input, the actual program of instruction.

- Implement according to procedures outlined in step III.2 (Specify Instructional Management Plan & Delivery System) and continue as long as there is a need for instruction.

- Control the procedures and techniques for monitoring quality control of instruction and also continue as long as there is a need for instruction.

Formative and summative evaluations are built into Phase V (Control), but formative evaluation occurs throughout the steps of the ISD model. The feedback loop (illustrated by the heavy black line in figure 4) serves as the change mechanism at appropriate steps if deficiencies or irrelevant instruction occurs. Revision as a result of evaluation is possible at any phase in this closed-loop system.

Figure 5 briefly explains what is done in each of the steps of the ISD process identified in figure 4.

QUALITY CONTROL

Two key quality control measures are imperative in monitoring ISD. These are adequacy and consistency. Adequacy is vital throughout the five phases of the ISD process and is assured, for example, when:

- Objectives accurately reflect the intended student performance after training.

- Objectives clearly state the conditions under which student performance is expected.

- Objectives concisely specify the criteria for measuring student performance.

ANALYZE	DESIGN	DEVELOP	IMPLEMENT	CONTROL
I.1 Analyze job to be performed to determine what is done, when it is done, order done, and conditions under which job is done. I.2 Determine criteria for selection and list tasks selected for instruction. I.3 Determine performance standards for tasks by observing/interviewing job holders and verifying with subject matter experts. I.4 Analyze existing courses to see if part or all of the analysis has been done (ISD Mode). I.5 Determine most suitable instructional setting for each task (e.g., individualized instruction, OJT, formal course, trischool).	II.1 Convert each task into terminal objective to determine enabling objectives and learning steps for mastery. II.2 Develop test to match each objective consistent with conditions. II.3 Develop pretest to determine entry level and/or to bypass objectives already mastered. II.4 Determine independent, dependent, or interdependent objectives for scope, sequence, and structure.	III.1 Classify objectives according to learning categories (skills, information, attitudes) and select most appropriate learning activities (practice, media, discussion). III.2 Determine how instruction is to be packaged and presented to students by selecting appropriate media and developing a plan to allocate and manage resources for conducting instruction. III.3 Review existing materials to determine which can be used, should be revised, or can be used in developing new materials. III.4 Develop instruction (initial plan to include materials, procedures, and media for teaching/learning). III.5 Validate instruction (try out in small groups and revise above).	IV.1 Implement management plan to train or orient teaching staff and to select training managers for program administration and collection of evaluative data. IV.2 Process students, obtain resources, conduct instruction, and use feedback to improve program.	V.1 Conduct internal evaluation by analyzing learner performance and instructor feedback to determine deficient or irrelevant instruction and suggest solutions to problems. V.2 Conduct external evaluation by assessing job, duty, task performance on the job--compare to others not trained. V.3 Decision-making process to revise Phases I to V as needed.

Figure 5. Explanation of Steps in the Five ISD Phases.

- Tests accurately and fully measure performance or cognition.

- Tests contain only items that are well-constructed, clear, and unambiguous, and they do not give away answers.

- Tests include items that provide opportunities for students to make errors commonly made on the job.

- Presentations reflect sound principles of learning and teaching.

- Presentations are performance oriented, i.e., directly related to job/task mastery.

Consistency is essential through the analysis of job, duties, tasks, and elements, as well as among/between tasks, objectives, test items, and presentations. Consistency is assured, for example, when:

- Objectives accurately correspond to tasks or elements and conform to the purpose of training, i.e., preparation for a job.

- Test items are congruent with objectives through matched conditions, standards, and actions.

- Presentations parallel objectives and test items by providing learning experiences identical to conditions of the objectives/test items and correspond to on-the-job conditions as much as possible.

OUTPUT

What can be expected by developing curriculum according to the ISD process? We find that ISD provides effective and efficient training for the job and attends to cost justifications. Moreover, ISD optimizes the proportion of entering students who meet acceptable job/task performance standards by the end of training.

REFERENCES

Cottesman, AM: Applying a Model in Curriculum Planning. NASSP Bulletin pp 24-30, October 1977.

US Navy: Interservice Procedures for Instructional Systems Development, NAVEDTRA 106A. "Phases I-V" and "Executive Summary and Model." 1975.

US Navy: Procedures for Instructional Systems Development, NAVEDTRA 110. 1977.

US Navy: The Instructional Quality Inventory, NPRDC SR 79-24. Vol II, "User's Manual." 1979.

(408)

HALTRECHT, Ed., Ph.D., Ontario Hydro, Toronto, Ontario.

CODAP: INTRODUCTION AND USES IN A LARGE PUBLIC UTILITY (Wed A.M.)

Ontario Hydro is a large public utility employing close to 30,000 employees. In attempting to develop a method of establishing training priorities, we recognized the need for a viable occupational analysis system. CODAP was introduced in 1977 and used experimentally with one occupation. Problems of operationalizing and "selling" the system were overcome. Eleven additional occupations have been completed during the ensueing 3 years. The CODAP data has been used primarily for training purposes. Some work enabling trainers to easily make use of the data has been fairly successful. Work on training course analysis, managerial analysis and safety is currently underway.

This paper is not an in-depth analysis of any one particualar application but a general overview of our solution to problems encountered in obtaining, introducing, operating, selling, and actually using CODAP and CODAP-generated information. While the emphasis is on training, some discussion of future applications will be presented.

CODAP: Introduction in a Large Public Utility

Dr. Edward Haltrecht, Ontario Hydro, Toronto, Canada

INTRODUCTION

Good morning. I manage a small applied personnel research unit consisting of eight professionals in Ontario Hydro. Ontario Hydro is the largest electrical utility in Canada and second largest in North America. Its operational area extends roughly a thousand miles from east to west and covers a quarter million square miles. It is a public utility selling electrical energy at cost, earning neither a profit, nor receiving funding from the tax base. Its interconnections with other jurisdictions place the corporation in the massive North American grid.

We generate electricity from water, the burning of fossil fuels and nuclear power, in roughly equal amounts.

In 1979, Ontario Hydro employed an average of 28,385 persons. Approximately a quarter of these form the professional group, including engineers, scientists and managers. Another 20% are part of the clerical support services, with the remaining 55% in relatively few, large occupational classifications, (e.g. 1200 power linemen). In addition, there are several other occupational groups which are small in number, such as a single boat captain and newspaper editor.

ACQUISITION OF CODAP

With this as a backdrop, I would now like to tell you about our acquisition of the Comprehensive Occupational Data Analysis Program (CODAP) system - I will briefly address the reasons for wanting it; how we obtained it; how we sell it to management and users; our organization for operating it; some methodological issues, our accomplishments; and plans for the future. You are fortunate in that I only have 30 minutes to cover these areas.

We first heard about CODAP from Dr. Raymond Christal at the 1973 MTA Conference. We recognized that CODAP is a management information system, whose output addresses issues related to organization structure, training needs analysis, job re-engineering, performance appraisal, among others. We wished to obtain the system, but felt that we would need a long term strategy to ensure that the transplanting of CODAP into Hydro would be successful and that it would not be shelved. Our strategy was to acquire this complex system only after we had an initial problem to solve, along with a visible and important client. If we could demonstrate CODAP's utility in solving such a problem, we would have earned the privilege of applying it elsewhere.

* Paper presented to the Military Testing Association (MTA) Conference, October, 1980, Toronto, Canada.

CODAP's debut turned out to be a training needs analysis in the Nuclear Mechanical Maintenance trades area. It may be interesting to note out rationale. First - why training needs analysis?

1. Changes in training (particularly cut-backs) can result in substantial resource savings - a common goal of line management.

2. Many of our trainers felt that there were either excesses or deficits in our programs. They had trouble in defining these, and management was reluctant to reduce or expand training without supporting data. In this respect a real information gap existed in Hydro, and our probability of success in filling this gap was high.

3. We felt that the unions would be more receptive of occupational analysis supporting training decisions as opposed to other areas such as performance appraisal, and job re-structuring.

4. While Hydro has systems in place to define jobs, establish pay, assess performance, no such systems existed for trades training needs analysis - in other words, we were not about to step on anyone's toes.

We chose the mechanical field because:

1. The trainers felt that they understood it best and would readily recognize a valid job, and

2. The mechanical task inventory would not be too complex as compared to operator (signal detection) type tasks.

To recap for a moment, we first heard about CODAP in 1973, and subsequently kept in touch with then Commander Bruce Cormack of the Canadian Armed Forces. Bruce walked us through our first study using the IBM version. With Ray Christal's support and the extraordinary generosity of the U.S. Government, we received the UNIVAC version in November of 1976. Within 3 months and many telephone conversations with Johnny Weismuller at AFHRL we had an operating system.

Unfortunately, we were too successful in our first endeavour and were forced to spend the next 3-1/2 years establishing training needs for just about every other trade and professional group in the organization.

In short we changed from a research function to an operational one.

The size of our research group made it impossible to both provide an adequate service in terms of establishing training priorities as well as doing the necessary research to rigorously test some of the ideas that we had with respect to using the CODAP system in non-validated ways. We are pleased that the anticipated operational load seems to be reduced for 1981. This will enable us to do the basic research that we feel is required.

ESTABLISHING TRAINING PRIORITIES

In attempting to introduce CODAP to the training community we basically had two choices: (1) to focus upon the problem and show that among the various alternatives, CODAP seemed to be ideally suited to solving that problem, or (2) to focus on CODAP itself and demonstrate how it can be used to solve problems.

We chose the former option. We argued that if training resources were limitless, they did not have a real problem. They could teach everyone everything for whatever dollars it took. However, since we are held accountable for dollars, and training time, we argued that the problem was one of establishing training priorities. To do so required a model, which we could not readily locate. Using material presented by the CODAP community, we worked to the model by successive approximations. It began with an arm-chair front-end analysis. We thought that we could improve the CASUAL system by documenting what was being done by job incumbents. While this would provide a "go", "no-go" training decision based upon whether a task was performed or not, we would still need additional information to establish a priority among these "performed" tasks. Collecting data on the number of workers performing each task seemed to be the next logical step. Priorities could simply be based on this data. However, how would we determine priorities for several tasks performed by equal numbers of workers? We introduced the time dimension, arguing that many tasks performed by equal numbers of job incumbents could be ranked based on the amount of time each task was done. The model is simple - provide information to trainers which documents how many and for what percent of time job incumbents perform each task.

We then presented data from a study which indicated that 95% of a work group changed in and out of work clothes, a larger number swept the floor, while, except for practice, no one gave artificial respiration. We argued that using our model, we would have a course on how to sweep the floor, change work clothes, and would of course ignore the teaching of artificial respiration. We obviously required additional information. We noted that information sources include the job incumbents as well as supervisors and experts. Table 1 summarizes the additional types of information that we suggested might be appropriate. Based upon the consequences of inadequate performance, and task learning difficulty, we rationally set priorities for the tasks in the aforementioned example. We then considered optional job analysis techniques which could be employed to gather the required information such as observation, interview, etc. The inventory technique was selected, as was CODAP and our first client - the Nuclear Generation Division.

Our problem, now, was to implement CODAP. Among our options, two captured our attention. First, increase our research group and form a section responsible for the generation of inventory booklets, the distribution and collection of booklets, lifting of data, data analysis and final interpretation. Unfortunately, circumstances

pre-empted this. At the time that we were bringing CODAP into the organization, Hydro was attempting to limit staff growth making further expansion impossible. Secondly, we are a research unit and to expand into the operation area did not seem to match our philosophy. Our second and subsequently selected choice, was to integrate the running of CODAP with the training department user groups. In essence one of our staff became familiar with the art of writing items and taught each of the training groups in the organization how to do likewise. By working in such teams we were able in each case to generate the inventory. In a sense, we captured additional staff by working very closely with users. The development of the inventory in this manner probably took longer had we done it ourselves, but we had few choices. I might add that the experience to date by using such a system is very positive. The user groups identify with the system, are extremely familiar with all the items, and find it fairly easy to interpret the final data analysis.

In terms of running CODAP, we had an individual spend about six months in learning the system. This individual, who was not a computer specialist, did have several years of experience in the use of canned programs such as SPSS. Our team consisted of 3 individuals from research - a programmer, an inventory specialist, and myself; as well as up to 5 captured user group members. By working closely with them, we now have a large community of sophisticated CODAP users in the organization. It is little wonder that we are starting to find that community placing more demands on their data.

Incidentally, we were worried that some users, receiving a free service, would passively request our services. In the long term we felt this to be detrimental. We decided therefore, to charge for printing, travel, and computer processing. This did not limit our customers - perhaps we should have charged more.

DATA GATHERING AND PROCESSING

In most of our studies we inventory the entire population, usually consisting of 200 to 600 members. We now collect task factor data (learning difficulty, etc.) prior to gathering job incumbent data. When the 20 or so supervisors and experts scrutinize the inventory in scaling task factors, they provide an excellent critique of our items. We have often modified our inventory on this basis.

We employ the services of a local optical character reading (OCR) company to lift our data. The system is somewhat clever. It can detect if an individual placed more than one response mark for an item, and erased the unwanted ones. It then determines the most dense response and saves it for output. When such a decision is made, it provides a list of the book, page, and line number for us to verify that the proper response was selected. The system is also capable of reading hand printed and typed alphanumerics. The final output is a computer tape in the format required by the first CODAP program.

IMPLEMENTATION ISSUES

A) CREDIBILITY OF CODAP DATA

One of the most frequent questions asked by managers, is how do we know that employees are honestly completing the inventories? We referred to validation studies that have been done by others. We also graphically presented one of our studies dealing with six thermal generation plants and the same trades group (see Figure 1). The data were ordered in descending order of percent members performing across duty headings for one of the plants. That is, we forced the smooth function by changing the duty heading order on the abscissa for one of the six. We then plotted the results for each of the other five stations, on the same graph. The curves were almost identical with the exception of one station on one duty heading - welding (J). Upon investigation, it was realized that problems at that plant demanded an excessive amount of welding. With this type of analysis, we have had no further challenges as to the credibility of CODAP generated information.

B) CONFIDENTIALITY

It appeared that a demand for identification of respondents, which is recommended by most CODAP users, would almost be impossible in a union environment. We, however, asked for names on a volunteer basis. The percentage of individuals doing so ranged between 85 and 97% over our 13 studies.

C) END USER TERMINOLOGY

It became apparent that while we in the research group understood and obviously enjoyed terms such as "policy capturing" and "multiple-R", our user groups, educated in other disciplines, were uncomfortable. They had problems in making training changes on the basis of a task statement, followed by up to 8 numbers (each with two decimal places) representing such factors as percent members performing, percent time spent, task learning difficulty, and safety consequences. To aid the end users, we draw frequency distributions of number of items and scale values (see Figure 2).

The trainers divide the distribution into several categories. For the example in Figure 3, items scaled above 6.25 on the consequences ofinadequate performance were labeled "H" on HIGH. (Eighty-five such items were so labeled.)

We then list all tasks along with the category labels representing the factors of interest. For example the task - "Make a final selection decision" was found to be H, M, M (high, medium, medium) for "consequences of inadequate performance", "task learning difficulty", and "part of job" respectively. The decision table used (see Figure 4) led to a change in training - it became part

of the training program. The "X's" in Figure 4 indicate the conditions for training course inclusion. Only when a task is scored H or M on consequences, learning difficulty, and part of job will it become part of the curriculum. I understand that a new CODAP program soon to be released will do the above work for us more readily.

Using the same type of categorization, an overall training profile analysis is also generated. The horizontal axis in Figure 5 represents the notion of task criticality while percent members performing is presented on the ordinate. If a high percentage of employees perform a task that is also rated high on the other factors such as learning difficulty, the task will be a candidate for off-site training as a core-task. This means that all members of that occupational group will be expected to learn how to perform that task. On the other hand, an equally "critical" task performed by a low percentage of members would be taught off-site to only those specialists required to perform the task. Tasks with low critical rating would be taught on the job as either a core or specialized task.

Another aid that we have adopted is the plotting of percent performing histograms for tasks, across apprenticeship year. Figure 6 represents six such tasks. In each group the bars represent subsequent years from left to right over a four year training program, ending with the journeyman. In task B31, we see the probability of task performance increasing with apprenticeship year. For B31 there is a very good chance that any one individual would have performed that task upon graduation to journeyman. In task B6, on the other hand, the system is such that there is only a very slim chance that an individual will have been called upon to perform this task. The trainers were able to use this data to convince the generation station managers to alter field experiences. Task B1 is a very simple task and poses no problem in that everyone is involved, commencing early in their career.

D) OTHER SCALES

We are currently examining the use of some unique scales. For example there are two which deal with safety. The first asks for the likelihood of an accident if standard operating procedures are ignored, and the second asks for the probable severity of such an accident. An individual using a hammer carelessly will probably have an accident that is not too serious.

E) ANALYSIS OF MANAGEMENT TASKS

We think one of the biggest challenges for us is the area of managerial occupational analysis. The literature suggests a number of factors that could be used by managers to rate tasks, such as importance, frequency, relative time, complexity and so forth. We decided to try relative time and importance. In our first managerial study we combined these in a matrix (see

(416)

Figure 7). We instructed job encumbents to assign a 7, 8, or 9 to those tasks that were extremely important. Further, they should distinguish between the 7, 8, or 9 on the basis of the relative time that they spent doing those tasks. In other words a 7 was a task that was seen as very important with very little effort or time spent on it. A 3 was a task that was deemed to take quite a bit of their time and yet was low in order of importance. Percent members performing data are not affected by using the matrix. The 9 point scale, being multi-dimensional, is problematic. But ours was an "explicit" multi-dimensional scale - even an apparent "single dimension scale" called part of job may be implicitly multi-dimensional.

More recently, we had managers and professionals score each task twice - once on importance and once on relative time spent. The results are just being analyzed and appear to have some utility.

F) TRAINING COURSE ANALYSIS

In most of the Hydro trades areas analyzed to date, the training is designed for a specific and well-defined population. For example, we have a 9 month training program for the nuclear mechanical maintenance tradesmen. CODAP generated data has led to very specific changes in that particular curriculum - new tasks being taught and others dropped.

On the other hand, a centralized group serving the entire corporation offers several in-house management and supervisory courses. These are from 2 to 21 days in duration.

The content is general enough to serve a wide range of management and professionals including chemists, electrical engineers, and behavioural scientists. It is not always practical to tailor make any one course to exactly fit the needs of the heterogeneous groups loaded on that course. In addition, there are a number of outside courses into which many of our professionals wish to enroll. One problem is to establish who should be loaded on what course. Another is to establish whether or not the current curriculum offers enough training, or if there is too much overlap between courses. To address these issues, we have begun to investigate a use of CODAP which I call training course analysis. We had the trainers complete a management task inventory for each of the courses they offered. Each inventory was filled out by trainers as if the course was "a person" except the scale dealt with "current training emphasis". A task greatly emphasized in a training course was scored as a 9. A task not covered was scored as a 0. The completed booklets, one per course, were treated as people and tracked on the cluster diagram. We are now analyzing the results.

analysis of overlap among courses may provide evidence of course content redundancy. Finally, gaps in course content would be identified by difference measures between the courses as a group and the overall job description.

CHANGE

We are often asked how CODAP may help to change job content. In a parallel study to course analysis described above, we plan to provide management with inventory booklets. They would complete one booklet for each job as they would like to see it created. In other words, they will provide target job descriptions. To aid them, we plan to experiment with two techniques. In the first case, they would have a free hand and no CODAP generated data. They simply would use the 9-point scale to indicate what relative time they wish a group of employees to spend on each task. In the second case, we would provide CODAP job descriptions for a target position. We would then use a scale where, for each task, senior management would indicate their desire for more, less, on the same amount of time to be spent by job incumbents.

In each case, the data would provide a desired job description which would be treated as "people" in the cluster programs. The Cluster diagram, in addition to overlap with the theoretical positions, could indicate which job incumbents come closest to these theoretical jobs. Group differences would tell us what current tasks would have to be dropped and waht new ones would have to be picked up to bring these workers closer to the jobs as ideally described by management.

RESULTS TO DATE

I would like to close by indicating some typical training cost-savings realized to date, on the basis of CODAP data analysis. With a few recent changes, the Nuclear Generation group is saving over one million dollars annually with reductions in training to their professionals and trade groups.

Just two training decisions for a trades group in our Thermal Generation Division has led to an annual savings of $186,000.

Analysis for a group of Hydraulic Generation electricians led us to no training time savings but the re-distribution of training effort has improved their training profile.

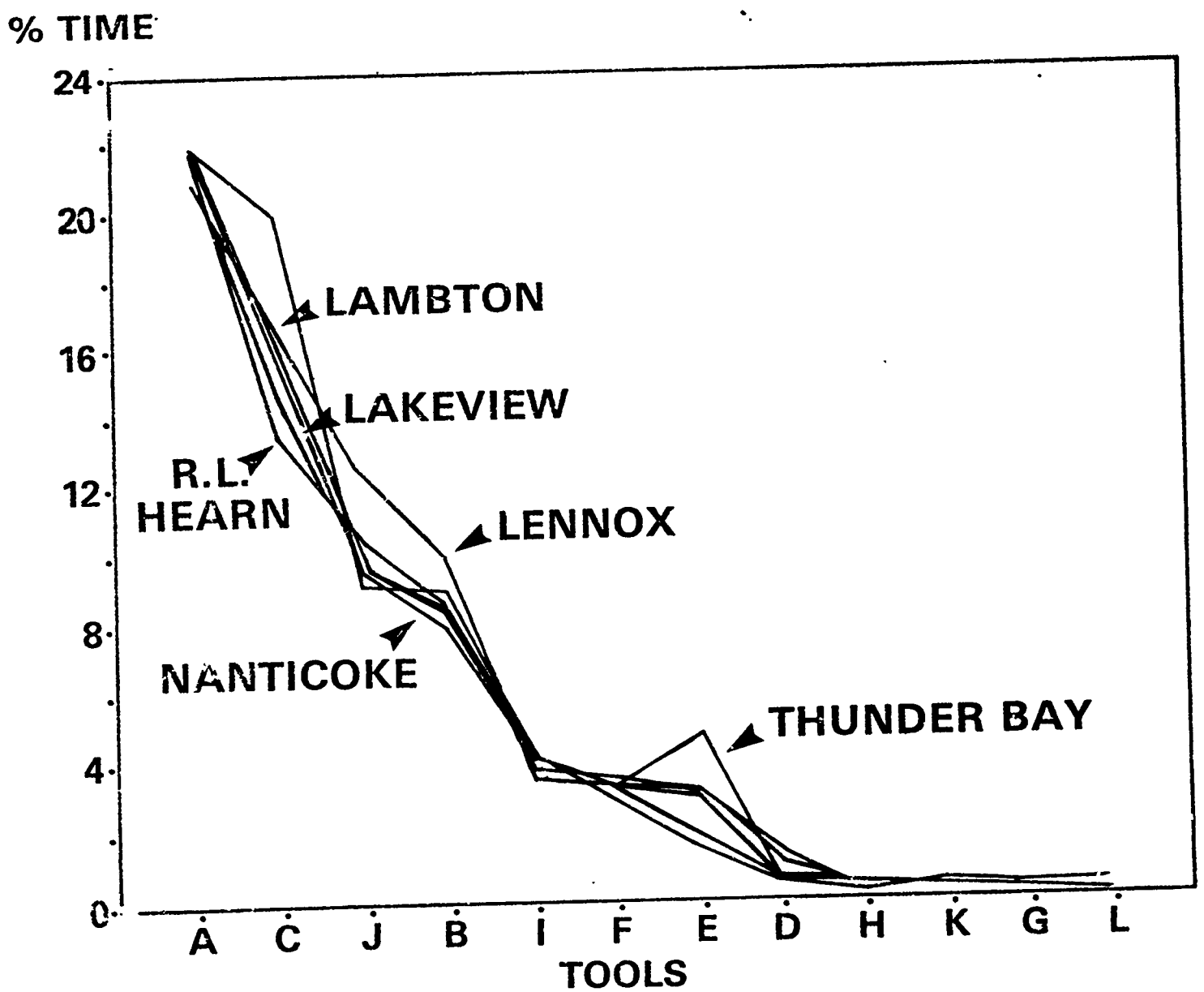
We have come a long way, and are pleased to be part of an active community of very dedicated scientists helping us go the next mile.

INFORMATION

TO DESCRIBE JOBS (FROM INCUMBENTS)	TO DESCRIBE JOBS (FROM SUPERVISORS, EXPERTS)
. RELATIVE TIME SPENT	. CONSEQUENCES OF INADEQUATE PERFORMANCE
. PART OF JOB	. LEARNING DIFFICULTY
. DEGREE OF IMPORTANCE	. NEED FOR IMMEDIATE PERFORMANCE
	. DESIRED TASK OCCURRENCE
	. SAFETY - ACCIDENT PROBABILITY
	. - ACCIDENT SEVERITY
	. CURRENT TRAINING EMPHASIS
	. TRAINING REQUIRED

TABLE 1 Types of information which may be used to establish training priorities.

FIGURE 1 PERCENT TIME SPENT USING VARIOUS TYPES OF TOOLS AND EQUIPMENT BY EMPLOYEES AT VARIOUS GENERATION PLANTS



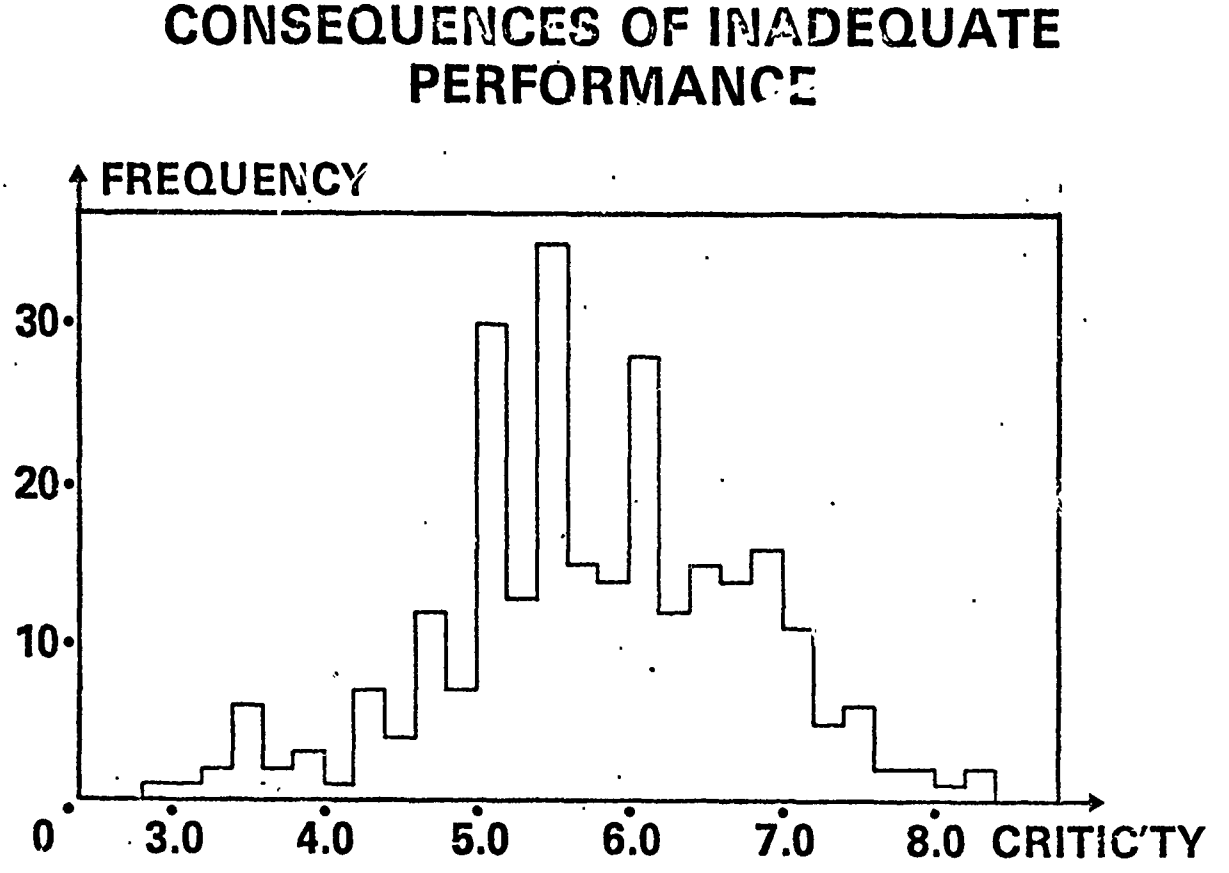


FIGURE 2 FREQUENCY DISTRIBUTION OF THE TASKS RATED ON CONSEQUENCES OF INADEQUATE PERFORMANCE

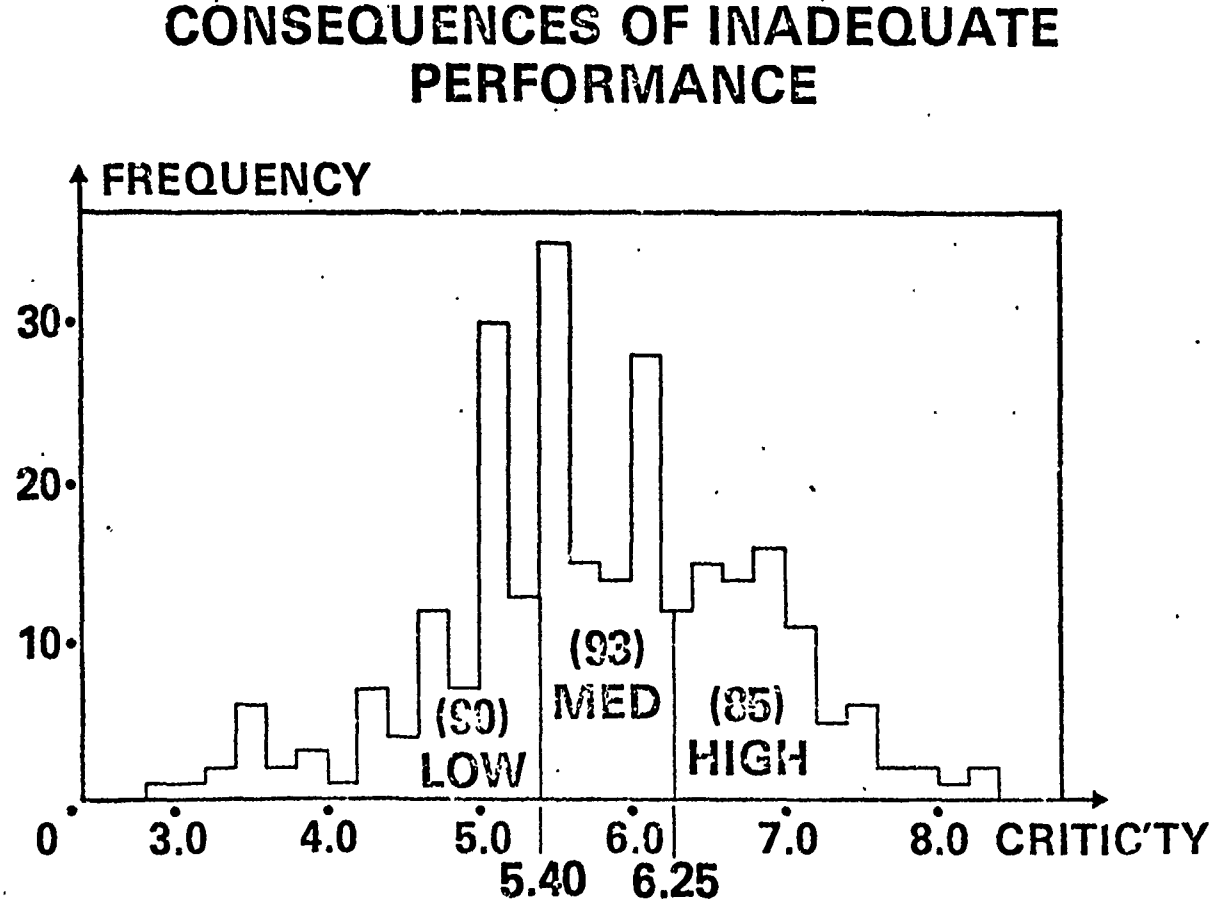
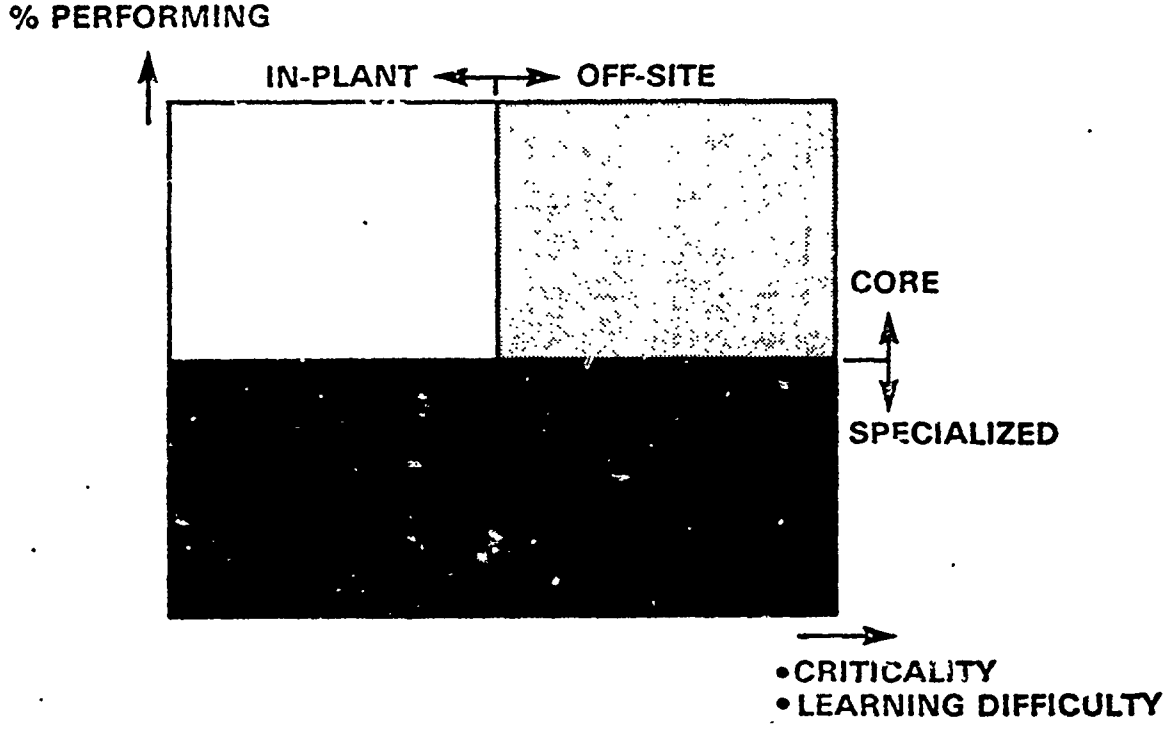


FIGURE 3 CATEGORIES FOR CONSEQUENCES OF INADEQUATE PERFORMANCE

FIGURE 4 TRAINING DECISION MATRIX



HAL-12

FIGURE 5 TRAINING DECISION MATRICES

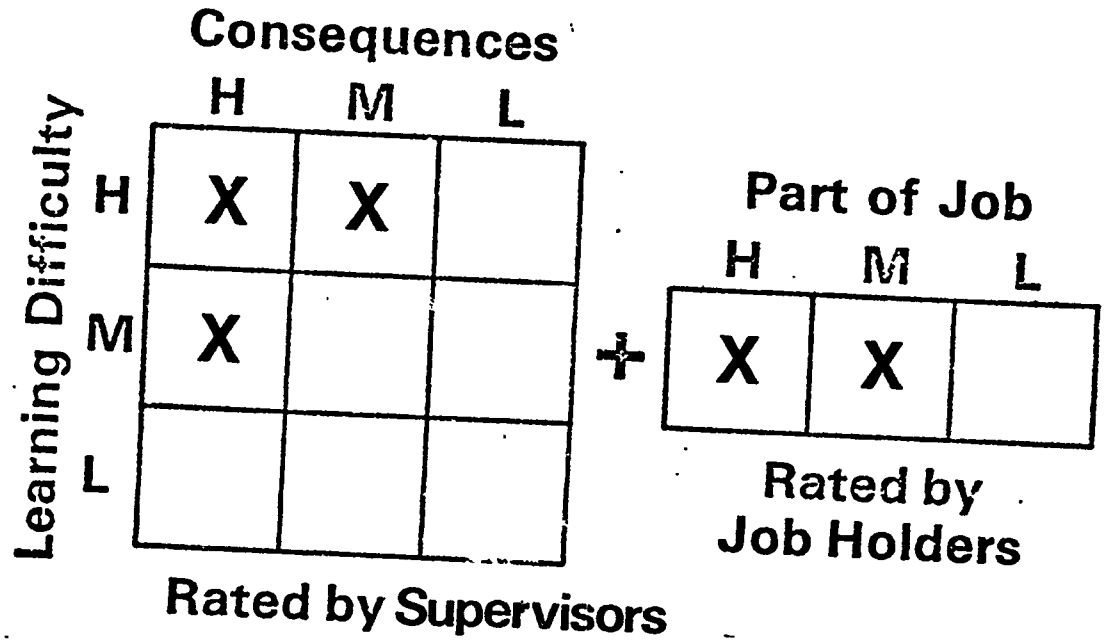
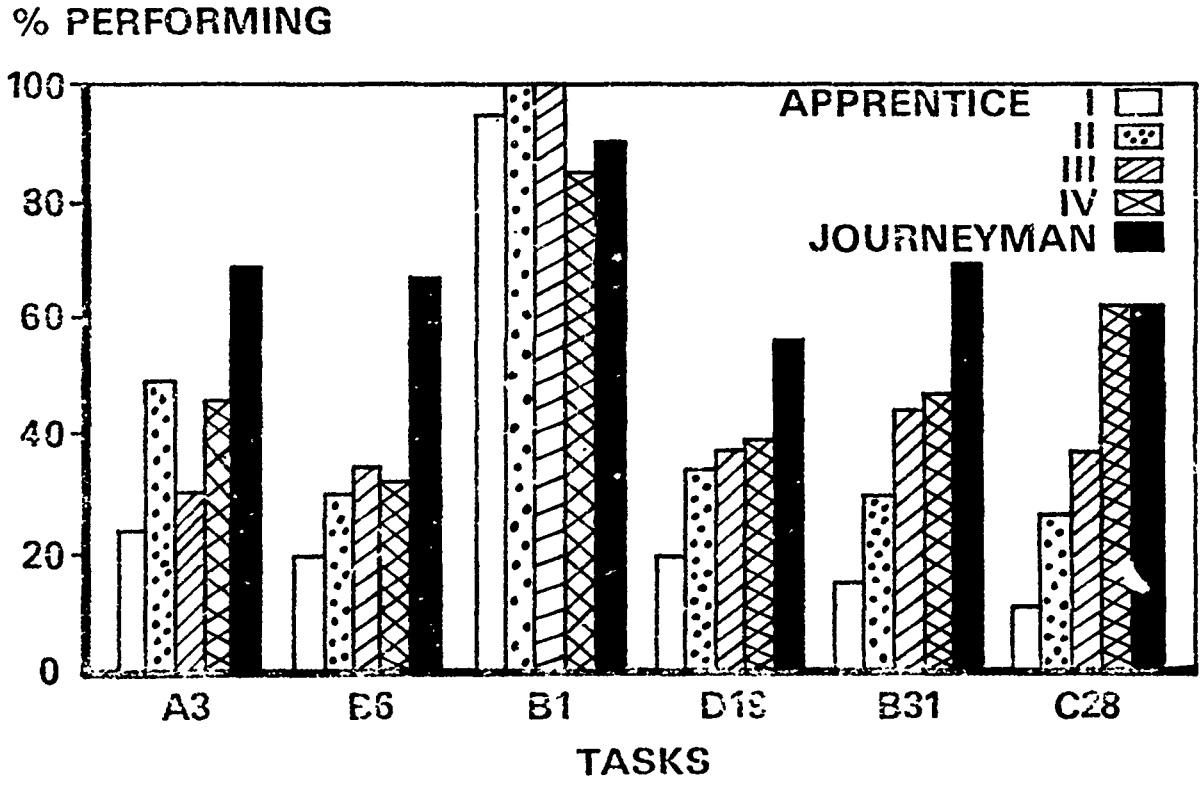


FIGURE 6 TASK PERFORMANCE HISTOGRAMS



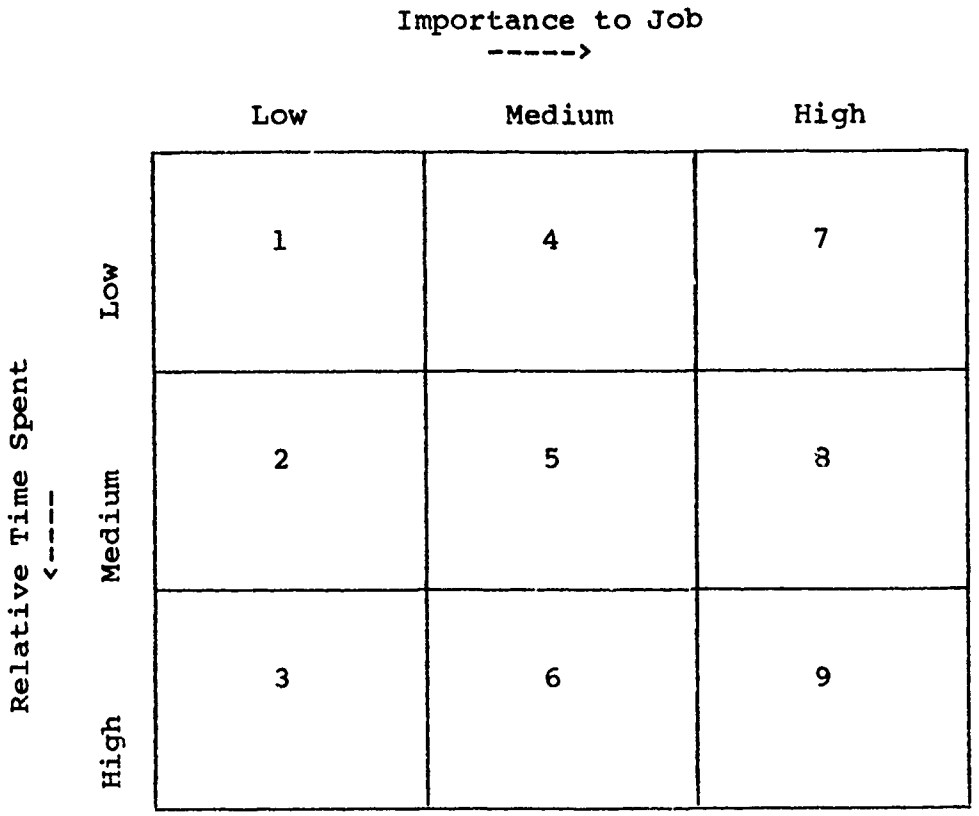


FIGURE 7 - Response Matrix Defining Part of Job.

HANSEN, K., Psych Systems, Baltimore, Maryland.

DISTRIBUTED PROCESSING CONSIDERATIONS IN COMPUTERIZED TESTING SYSTEM DESIGN (Thu P.M.)

Distributed data processing is a technique in which an application is segmented and its processing load is equitably distributed across several processors. The volume and geographic distribution of the military's testing needs suggest that certain concepts of distributed processing may be useful in design of a computerized testing system.

Adaptive testing, a form of administration likely to be adopted with computerization, requires a high level of computational power at each testing station. This may be obtained by providing a processing unit at each testing station or by accommodating several stations with a single shared processor. The processors can then be connected, in an appropriate network topology, resulting in very cost-efficient testing stations not requiring substantial mass-storage capabilities at each testing site. The individual processors can collect and temporarily store test data, forwarding the information to a more appropriate data base management system for maintenance and manipulation.

This paper discusses the rationale for a distributed approach to computerized testing. Considerations relevant to interfacing the various devices throughout a military testing network are discussed.

DISTRIBUTED PROCESSING CONSIDERATIONS IN COMPUTERIZED TESTING SYSTEM DESIGN

Kenneth E. Hansen

The utilization of computer technology for the development of automated testing systems has a number of distinct advantages. Some benefits of this approach have been illustrated in various evaluation studies (cf. Klingler, et al., 1977). Computerized testing systems are now commercially available and are being utilized in disparate sources ranging from the diagnostic screening of psychiatric patients (Johnson, Giannetti, & Williams, 1978) to the screening of drivers license applicants.

Much of the research and development that has been conducted in the area of computerized testing has direct application to military testing needs. However, since military testing applications would likely be conducted on a very large scale, with a significant number of individuals being tested at many different testing sites, the question arises as to what the most appropriate approach would be to meet this demand. One such approach would be to connect various computerized testing systems together to form a network and to distribute the processing amongst the various computers. Some of the concepts of networking and distributed processing, as well as their implications to military testing, will be discussed in this paper.

NETWORKS

The initial ideas about computer networking began with the introduction of time-sharing techniques and the development of data communications technology in the late 1950's. Time-sharing systems were typically constructed of a large central computer acting as a host to several remote terminals. As increasing demands were placed on the systems, some single host computers were replaced with multiple, connected computers, thus forming small computer networks. In these situations, the host processors performed large computations, controlled data bases, and supervised network operations. As computer technology developed, networking became more individualized and customized to the requirements of the situation. Individual processors were assigned to specific tasks and programmed to communicate with other processors and data base files. This approach resulted in easier software development and improved tolerance to system failures (Digital Equipment Corporation, 1974).

The computer network currently offers a number of significant advantages. By segmenting an application, a network of smaller computers can handle a distributed processing load more economically and can provide more processing power than would be possible in a single large computer. A network of computers also has the ability to grow relatively easily as increasing demands are made upon it and allows software to be broken down to a more manageable level, thus resulting in an overall reduction of programming time and related costs.

The primary advantages of utilizing a computer network are, then, the conveniences and economics that are achieved through resource-sharing. By linking several computers together through communication links, resources available at each location on the network are made accessible for use at any other location on the network. These shared resources might include devices, files, programs, and data.

NETWORK TOPOLOGY

The topology of a network refers to its geometric arrangement of links and nodes within the system. A link (circuit or channel) is the communication path between two nodes. A node is defined as an end point of a channel or a junction of a circuit. A node might consist of a combination of equipment including: a remote computing system, a host computer, a computer devoted to network control functions, or simply a remote terminal.

Network topology is related to network design, operations, reliability, and operating costs. A fully connected distributed network has more links for the same number of nodes than either a partially connected network or a simple star network. The major factor in determining the most appropriate topology of a network relates to the application. Several typical network configurations will now be discussed to illustrate some possible approaches.

The star configuration, or centralized system, is illustrated in Figure 1. In this approach, all users communicate with a central host that has supervisory control over its entire network. Remote users can communicate with each other only through, and by permission of, the central system or host processor.

A tree structure or heirarchical configuration is often used to supervise and control certain real-time applications. This structure, as illustrated in Figure 2, minimizes the communication links necessary in a network. However, this type of network can be separated by a failure of one communication node; thus, it should be used only where communication failures can be tolerated. The tree configuration simplifies control programming or network supervision, and is relatively inexpensive. Therefore, it is apt to be the most practical system in a medium sized or small environment.

In the loop or ring structured network, as illustrated in

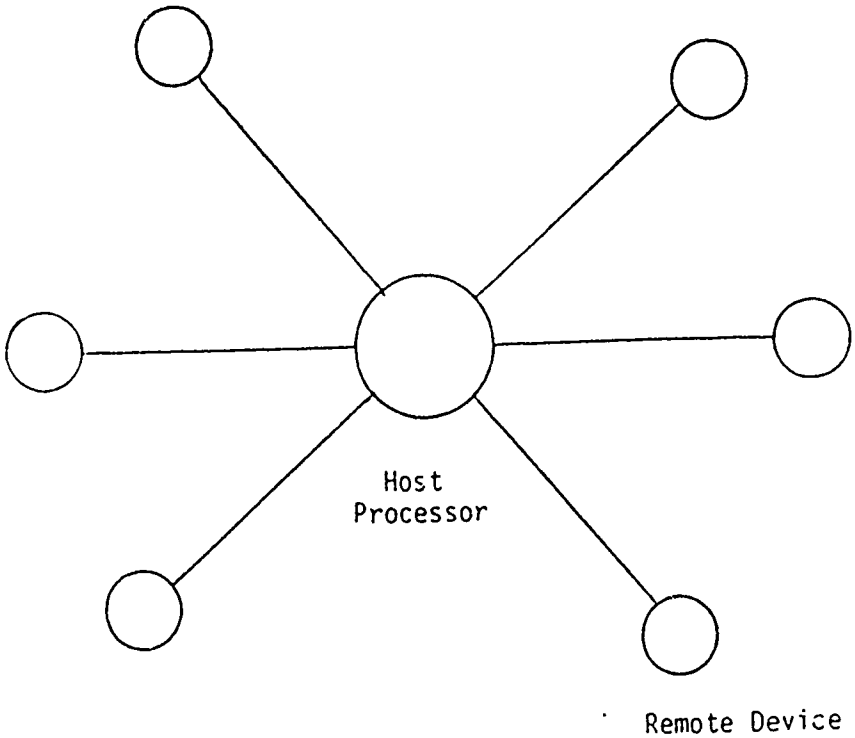


FIGURE 1
STAR CONFIGURATION

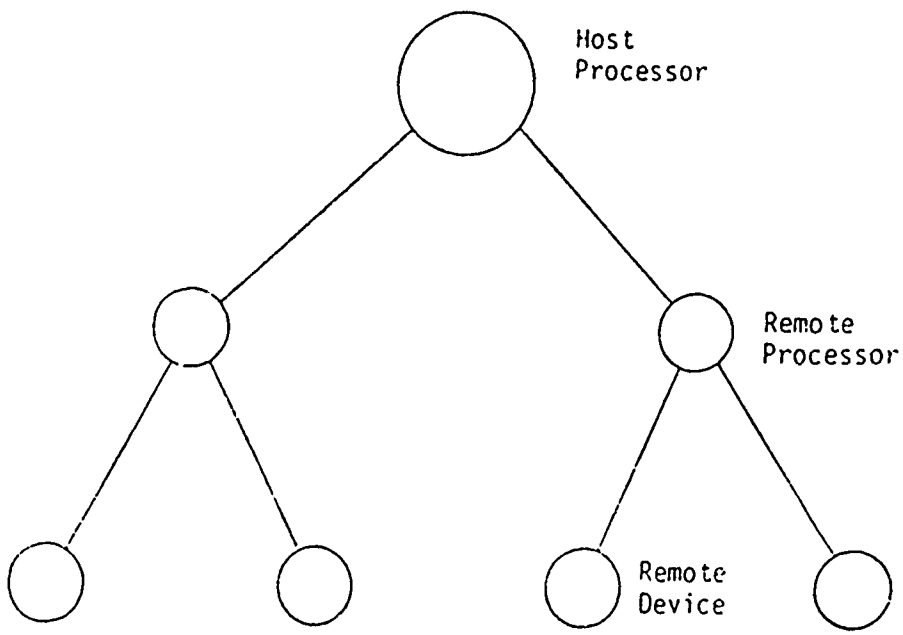


FIGURE 2
TREE STRUCTURED NETWORK

Figure 3, many of the remote stations (terminals or computers) connected to the ring do not communicate with the main site or host processor individually. Instead, the data to be transmitted is looped around the network.

The loop structured network is economical when several remote stations and host processors are located near each other. When remote stations are geographically disbursed over long distances, line costs may prove to be prohibitive.

Figure 4 illustrates a distributed system with a multi-star configuration, where there are several connnected hosts, each with its own set of users. Such a distributed structure offers considerable advantages in reducing the cost of terminal communications by permitting installations to be located near concentrations of terminals. If properly designed, distributed networks can offer significant reliability advantages, since a failure at one node does not affect the rest of the network.

In applications where the reliability of continuous communications is important, a fully distributed network, as shown in Figure 5, in which every point is connected to several neighboring points, may be preferable. The additional transmission paths provided in this type of network improve the overall performance of the structure. However, because of the redundancy in the communication paths, the associated costs are increased.

NETWORK COMPONENTS

A computer network is constructed by assembling various hardware and software components. The hardware components can be described in two general categories. First, there are the computers themselves, with associated peripheral equipment. The computers may serve as applications processing facilities, or may be devoted specifically to network control functions. The selection of appropriate computers to serve as hosts or nodal processors is of vital importance when designing the network. The goal in such design is to distribute the processing load to the various computing facilities to result in a cost effective system of greater efficiency. Different types of computers with different processing or storage capabilities can be effectively mixed in designing the network.

The second general category of hardware components consists of the various communications modules, including interfaces, modems, and the communications channels or facilities. Communications channels are lines for transmitting signals. These channels are typically obtained from common carriers and may access the public switched telephone network, or may be obtained on a private leased line basis. A modem is a device used when a computer sends data over the telephone lines. The modulator portion of the modem converts digital pulses, originated by the computer or terminal, to analog signals acceptable for transmission over telephone lines. The demodulator reverses this process, converting the analog telephone

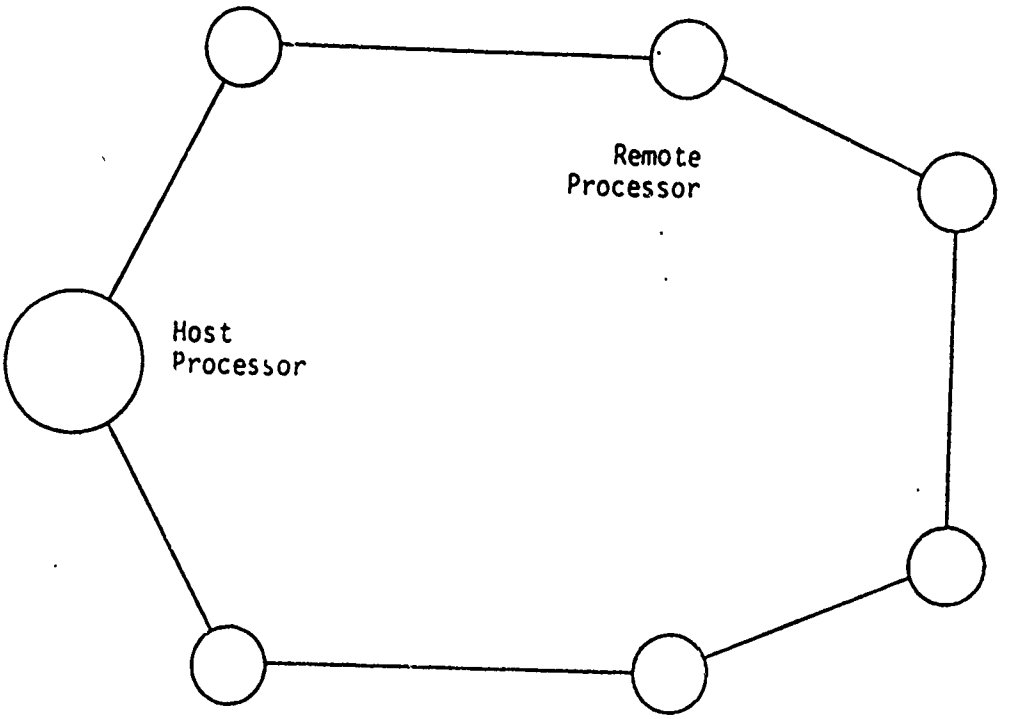


FIGURE 3
LOOP NETWORK

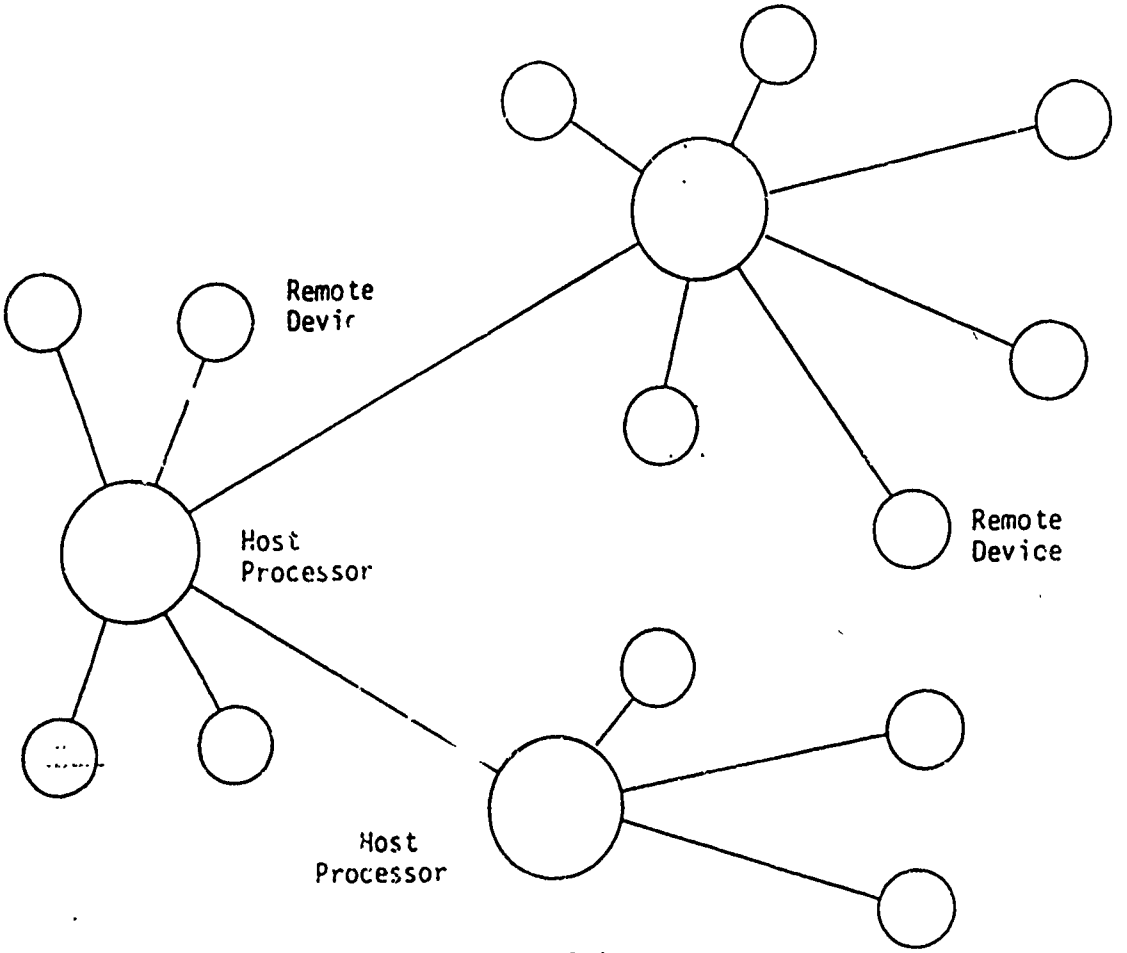


FIGURE 4
MULTI-STAR NETWORK

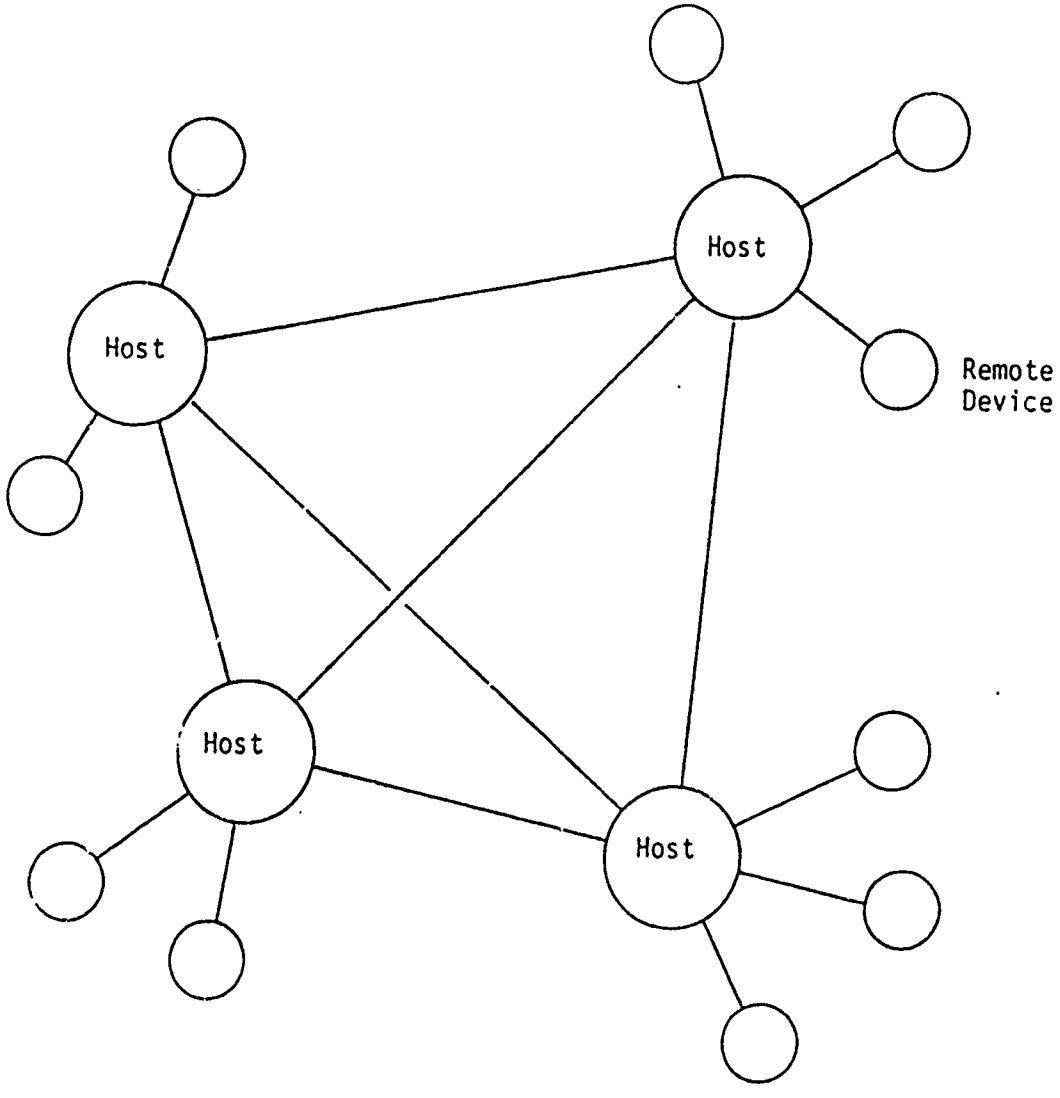


FIGURE 5
FULLY DISTRIBUTED NETWORK

signals back into digital data acceptable to the equipment. Modems transmit data in steady streams (synchronous) or in packets of data (asynchronous). Asynchronous modems tend to be somewhat less expensive, but they are also restricted to lower transmission speeds. One of the most commonly referenced characteristics of the modem is the speed or data rate, expressed in baud or bits per second. Within the Bell System, modems are referred to as Data sets or Data phones. A device similiar to the modem is the acoustic coupler. This type of device accepts serial data from computer equipment, modulates it into the audio spectrum and produces the modulation as audible tones. Acoustic couplers are equipped with fittings that accept conventional telephone hand sets and couple the acoustic signals directly into the mouthpiece. Acoustic couplers enable any conventional telephone to be used as a data terminal.

A line interface provides a means for a computer system to communicate with a peripheral device. The peripheral device, such as a Cathode Ray Tube (CRT) terminal, may either be located locally with a direct line to the interface, or at a remote site via a modem and communication line. Interfaces generally conform to the electrical and logical standards of the Electronics Industry Association (EIA).

The software elements necessary in a network can be described in three general categories, although there is considerable overlap in their functionality. The first major type of software can be described as system software. This software can be defined as an organized collection of programs designed to increase the efficiency of a computer system by providing common functions to all user programs. System software is also referred to as the "operating system". System software can vary in size and complexity, but usually provides control in monitoring program execution, management of system resources, and control of input/output (I/O) devices.

The second category of software necessary for a network application is the data communications software. These are programs and routines necessary to send data, commands, messages, and status from one computer to another or from one computer to another device. The data communications software typically consists of the line control module and the network control module. The line control module assists in providing error free communications between the various nodes in the network. The network control module supports interprocess communications in the network. It deals with creating links between processes, routing messages over the links, and acknowledging and diagnosing the message flow in the network.

Fortunately, formal sets of conventions governing the format and relative timing of message exchange between two communicating processes have been adopted. These rules for communication system operation are referred to as protocols (McNamara, 1978). Several different protocols are available, examples include IBM's Binary Synchronous Protocol (BISYNC), Digital Equipment Corporation's Digital Data Communication Message Protocol (DDCMP), the International Standards Organization High-level Data Link Controls (HDLC) and IBM's Synchronous Data Link Control (SDLC).

The final software component can be referred to as the application software. These are the programs which accomplish a specific user designed function. Application software typically utilizes higher level language compilers or interpreters, such as Fortran or Basic.

SUMMARY

A project was recently completed by Psych Systems, Inc. for the United States Air Force which serves to illustrate the planning involved for a distributed approach. The project called for a small portable system to provide vocational interest measurements on incoming recruits. The resulting prototype system consisted of a Digital Equipment Corporation LSI 11 Microprocessor, 64K bytes memory, two tape cassettes with total mass storage capabilities of 512K bytes, a 30 character per second matrix printer, and necessary interfaces. In this system, the backplane containing the processor, memory, and peripherals is located within the CRT housing. The system has the capability to communicate asynchronously with other equipment, either locally or by attachment of a modem to a remote system. By simply inserting a new interface module into the backplane and adding the appropriate communication software, this system could communicate with a majority of processors commercially available, and would support several of the communication protocols previously mentioned. Such a system also has the flexibility to increase its mass storage and computational power, while maintaining its ability to interface to a communications system.

Considering the size and scope of the military testing needs, concepts of distributed processing and networking appear to have much merit. This type of an approach could result in many benefits to the military. First, by segmenting the processing load, the application software is more manageable and more easily maintained. Second, as additional demands develop, additional testing systems can be added to the network, thus obviating the need for continual upgrades of a single computer system. Third, the military's testing requirements are diversified and continually changing. Limiting the number of terminals dedicated to any one processor would result in more processing power at that site. This would be an important fact in considering the anticipated implementation of contemporary adaptive strategies. Fourth, a network can be constructed such that adequate redundancy is integral to the system. This would result in an overall decrease in downtime cuased by any malfunctions. Thus, an appropriate network design is a very cost-effective approach.

REFERENCES

Digital Equipment Corporation. Introduction to Mini-computer Networks. Maynard, Massachusetts.

Johnson, J.H., Giannetii, R.A., and Williams, T.A. A self-contained microcomputer system for psychological testing. Behavior Research Methods and Instrumentation, 1978, 10(4), 579-581.

Klingler, D.E., Miller, D.A., Johnson, J.H. , and Williams, T.A. Process evaluation of an on-line computer-assisted unit for intake assessment of mental health patients. Behavior Research Methods and Instrumentation, 1979, 9(2), 110-116.

McNamara, John E. Technical Aspects of Data Communication. 1978, Digital Press, Bedford, Massachusetts.

HISS, Richard H., THOMASON, S., and WENGER, W., Essex Corporation, White Sands Missile Range, New Mexico.

HUMAN FACTORS DATA COLLECTION TECHNIQUES DURING FIELD TESTING (Tue P.M.)

Traditional approaches to Human Factors testing have usually relied upon questionnaire evaluations and the use of detailed checklists. Both of these approaches are more suitable for a laboratory or classroom setting where the environment is closely controlled and relatively few events are occurring at the same time. The approach to be described here has been used to perform Human Factors testing on systems in the field operating under more or less standard battle emplacements. For these types of situations many events are occurring simultaneously and it is difficult for one or two observers to monitor these events.

Our approach has been to train SOMTE (Soldier, Operator, Maintainer Test and Evaluation) personnel as human factors data collectors and use them during field exercises. This training takes advantage of the soldier's knowledge of the system, which most HFE personnel do not have; it allows for greater mobility for data collectors during testing, and it drastically reduces the need for additional HFE personnel during peak data collection periods.

This technique has been used to gather field data on the PATRIOT Air Defense System during recent tests at White Sands and has proven satisfactory and very reliable.

HUMAN FACTORS DATA COLLECTION TECHNIQUES DURING FIELD TESTING

R.H. Hiss, W. Wenger, W. Talley, and S. Thomason

ESSEX Corporation
P.O. Box 147
White Sands Missile Range, NM 88002

1.0 INTRODUCTION

During planning for the HFE evaluation of the PATRIOT Air Defense System, it was concluded that the HFE scientists assigned to the project would be severely overloaded when testing involved many player participants and pieces of equipment spread over a large area. It also appeared likely that this state of affairs could become a common occurrence in the test and evaluation of future systems. A method was therefore needed to expand the HFE data collection effort by supplementing the HFE specialists with lay personnel trained in HFE data collection. The Data Collector Orientation (DCO) was developed as a training aid for that purpose. Due to the complexity of the PATRIOT system, it became paramount that these data collectors should be experienced with the system. The effectiveness of system naive, briefly trained data collectors would be minimal. The primary requirement in using the DCO, therefore, is that data collector trainees are system experts, able to apply their HFE training to the collection of HFE data on a familiar, complex system.

2.0 PREPARING THE DCO FOR USE

The DCO is designed to meet those common needs of HFE specialists evaluating different systems. HFE tools, constructs, and procedures not relevant to the general data collection mission are not addressed. The applicability of the DCO to other systems will, of course, vary. Terminology and examples that are PATRIOT specific are presented as an example, and these terms and examples must be modified to fit a particular system. To complement the training, a data collector's handout* is distributed to each student. Ample copies, suitably modified for the applicable system should be produced prior to instruction. The DCO consists of a brief introduction to HFE, instruction in the data collection techniques to be used, and a practical exercise with at least two scenarios. The scenarios should be video taped and must be developed by the instructor to agree with the Task-Time Checklist used. Performance of the scenarios in real-time during instruction is not recommended as this may introduce a variability in student responses across multiple classes.

*Space limitations prevent including the data collector's handout in this paper, however, this material will be provided upon request.

The view, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

3.0 PREDICTING DATA COLLECTOR PERFORMANCE

An estimate of the data collector's performance may be made by collecting the completed Task-Time Checklists at the end of the practical exercise. These data may be compared to pre-established criteria or a baseline performance in terms of time variability, number of errors, and other quantified, dichotomized or scaled comments. This information can be used to assign data collectors to observational tasks which take advantage of their particular talents.

4.0 DCO VALIDITY

At its present stage of development, the DCO only has face validity. Pilot studies of training indicate that the information presented is easily learned by US Army personnel who are trained maintainers and operators for the PATRIOT Air Defense System. Most of these personnel had extensive (more than 5 years) experience in Air Defense Systems such as HAWK and HERCULES. The DCO needs to be compared to other data collection procedures under actual use conditions for ultimate validation. Some data are available now based on our experience during the Development Testing II phase of the PATRIOT evaluation. Comparisons were made between a sample of video tapes and the comments recorded by the data collectors during the March Order and Emplacement exercises. Not only did the data collectors out-perform the video support, but because of their mobility they were able to move around and visually access places the video cameras could not. Because the primary data acquisition activities were covered by these collectors, HFE personnel were able to perform video and photographic functions. We feel that the six hour training session required to train the data collectors was well spent and that the technique is applicable to future field exercise testing.

OUTLINE USED FOR DATA COLLECTOR TRAINING

INTRODUCTION

1. OBJECTIVE: A rapport is established between trainees and HFE personnel.

INSTRUCTIONAL PROCEDURE:

a. HFE personnel introduce themselves to trainees and explain their role in the program.

b. Have students introduce themselves and fill out an academic and work experience information sheet.

2. OBJECTIVE: Data Collector trainees are cognizant of the HFE mission and of their contribution as individuals, and as a group, in accomplishing that mission.

INSTRUCTIONAL PROCEDURE: Describe the role HFE has played during the development and utilization of military/civilian systems in the past.

3. OBJECTIVE: Trainees are knowledgeable in HFE fundamentals.

INSTRUCTIONAL PROCEDURE:

a. Present the film Of Men and Machines: Engineering Psychology.*

b. Describe how HFE can enhance the safety and performance of complex military systems. Ask trainees to describe any HFE related problems they have experienced in military or civilian systems. At this stage of instruction, each trainee should be able to describe a HFE problem he or she has experienced.

*A BW, 29 minute sound film from the Focus on Behavior Series available from Krasker Memorial Film Library, 765 Commonwealth Ave., Boston, MA 02215.

INSTRUCTION

1. OBJECTIVE: Trainees are indoctrinated in the specific areas to be evaluated within the system.

INSTRUCTIONAL PROCEDURE: Define and give examples of the areas to be addressed during data collection. Considerations with regard to the PATRIOT Air Defense System are:

(a) Task sequence validation.
(b) Establishment of task time lines.
(c) Error analysis.
(d) Safety.
(e) Crew proficiency.
(f) Personnel movement and posture.

2. OBJECTIVE: Data collector trainees are able to record task times and observations on the TASK-TIME CHECKLISTS.

INSTRUCTIONAL PROCEDURE: Distribute sample checklists, identify each specific area of the checklist and describe the procedures for recording data. Discuss the terms presented in the GLOSSARY.

- For each task named, the following is required:

1. Verify the sequential position of the task as shown on the checklist.
2. Circle "LMT" if task movement time is subjectively long.
3. Circle "ImpM" and/or "AwkP" when impeded movement and/or awkward posture are observed.
4. Circle "LowP" if low proficiency is observed.
5. Place a tally mark for each hazard observed in the HZ column.
6. Place a tally mark for each error observed in the E column.
7. Record task time.
8. Complete the HAZARD ANALYSIS at the conclusion of the test exercise.
9. Complete the CREW MEMBER ERROR ANALYSIS at the conclusion of the test exercise.

3. OBJECTIVE: Trainees are able to operate the stopwatch which is used for task timing.

INSTRUCTIONAL PROCEDURE: Assign each student a stopwatch and instruct trainees on stopwatch operation.

4. OBJECTIVE: Trainees are able to track player-participants efficiently.

INSTRUCTIONAL PROCEDURE: Discuss tracking procedure.

- Data collection is most accurate when the data collector has unobstructed visual and auditory access to crew member activities. The space between the data collector and the crew member shall be maximal with regard to distance and optimal with regard to sensitivity. The data collector, being familiar with the crew member's function, should anticipate the movements and activities of the assigned crewman as well as the other crew members in the area. The catchword for the data collector is best expressed as DON'T INTERFERE, meaning: (1) Don't assist; (2) Don't interrupt; (3) Don't ask or answer questions; and (4) Don't be an obstacle.

5. OBJECTIVE: Trainees understand data collection procedures.

INSTRUCTIONAL PROCEDURE: Discuss data collection procedures:

- Task Timing - The stopwatch is started when the assigned player participant begins his/her function in the test operation. Each task time is captured upon completion of the task. The task shall not be considered completed until (1) the task goal is accomplished, (2) task-related material is stowed, and (3) the task workspace is secure.

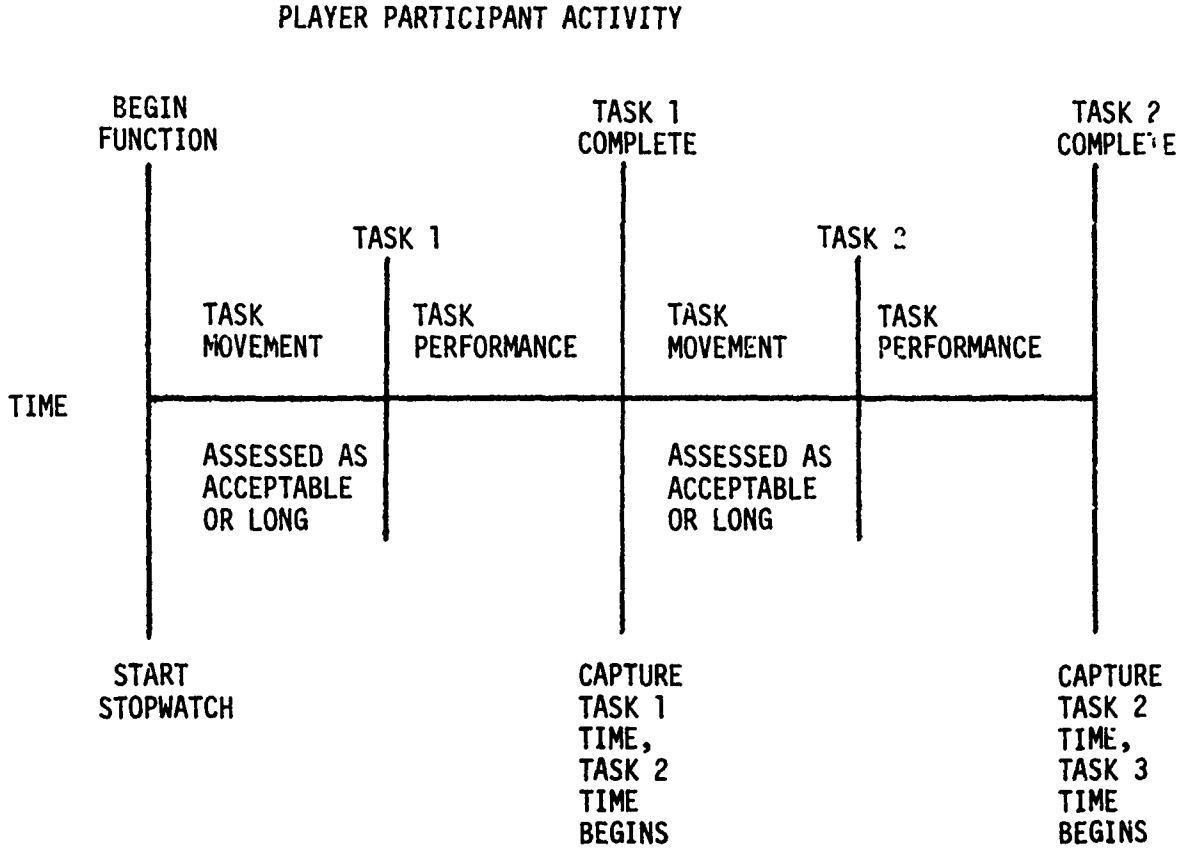
The timing of each successive task begins when the previous task time has been captured. Task time consists of two parts, task movement time

and task performance time. Task movement time will always occur before task performance time and, summed together, they will equal the captured task time. During data collection, task movement time will be estimated as:

LONG - Associated with considerable movement such as moving from one major hardware item to another; longer than 20 seconds duration.

Task movement ends when either of the following occurs:
(1) The player participant arrives at the task workspace.
(2) The player participant obtains material necessary for task performance prior to arriving at the task workspace.

The following diagram illustrates the relationship between collecting time data and player participant activities.



PRACTICAL EXERCISE

1. OBJECTIVE: Trainees have experienced the data collection process including the timing of tasks and the recording of HFE observations.

INSTRUCTIONAL PROCEDURE: Trainees are presented a videotaped sequence of tasks from which they obtain task times, make HFE observations, and record task times and observations using the TASK-TIME CHECKLIST.

- Suitable practice scenarios include the assembly and disassembly of a bicycle. Bicycle components and the associated tasks of assembly and disassembly are familiar to nearly everyone; the trainee should have no problem in collecting task/time data.

2. OBJECTIVE: Data collector trainees have been evaluated on their level of proficiency at collecting HFE data.

INSTRUCTIONAL PROCEDURE: The TASK-TIME CHECKLISTS completed during the practical exercise are graded on accuracy and thoroughness. Trainees not performing at an acceptable level of performance are given additional training in those areas in which they are deficient.

GLOSSARY

Task Not Applicable to Exercise - A task may be eliminated from a test exercise due to safety restrictions or nonavailability of task hardware. To lessen the data collector's workload, it is recommended that these tasks be removed from the checklists prior to data collection.

Task Performed Prior to Exercise - Any task performed prior to the test exercise should be noted in the "comments" area of the checklist. The HFE data collector in charge shall notify the test conductor of such events. Any task performed prior to the test exercise shall not be scored, and the number of observations of that task will decrease accordingly.

Task Performed By Other Crewmember - When a crewmember cannot perform a task because it was previously performed by another crewmember the data collector notes this under "comments" and does not score as an error.

Performed Nonassigned Task - Performance of another crewmember's task prior to when the task should be performed is an error and is recorded on the error analysis form. Performance of another crewmember's task because the task had not been completed as assigned is not an error.

Incorrectly Performed Task - Performance error of which any one of the following apply:

(1) Unsuccessful attempt to complete a step or task.
(2) A step which must be repeated before being completed.
(3) Any action resulting in damage or injury.
(4) Unacceptable end results.
(5) Forgetting to perform any step of a task.

Task Not Done - Forgetting to perform the entire task.

Taylor Timing - A stopwatch function for timing sequential events where each press of the button captures the time elapsed since the last button press and begins timing the next event.

TASK-TIME CHECKLIST

TEST FUNCTION: Crew Member #1, LS Emplacement **FU** ____ **SITE NO.** ____

DATA COLLECTOR: _____ **DATE:** _____

PLAYER PARTICIPANT: _____ **START TIME** _____ **END TIME** _____

S	TASK	OBS	HZD	ERROR	COMMENTS	TIME min	TIME sec
1	Connect ground cable	LMT ImpM AwkP LowP					
		LMT ImpM AwkP LowP					
2	Lower gooseneck work platforms	LMT ImpM AwkP LowP					
3	Start generator	LMT ImpM AwkP LowP					
		LMT ImpM AwkP LowP					
4	Roll up louver covers	LMT ImpM AwkP LowP	✓	✓			

LMT — Long Movement Time
ImpM — Impeded Movement
AwkP — Awkward Posture
LowP — Low Proficiency
(CIRCLED WHEN APPLICABLE)

When marked, an ERROR ANALYSIS form is filled out.

When marked, a HAZARD ANALYSIS form is filled out.

(448)

CREW MEMBER ERROR ANALYSIS

Instructions:

1. Complete at end of test exercise.
2. Attach to TASK-TIME CHECKLIST.
3. Data collector completes Section A.
4. Crew member completes Section B.

SECTION A - DATA COLLECTOR

DATE:_____ TIME:_____

TASK SEQUENCE NUMBER:_____

TASK TIME:_____

CLASSIFIED AS ERROR OF: ☐ OMISSION
☐ COMMISSION

ERROR DISCOVERED: ☐ BEFORE TASK WAS COMPLETED
☐ AFTER TASK WAS COMPLETED

IF AFTER, WHAT WAS THE CORRECTION TIME? _____

DESCRIBE ERROR:_____

_____

_____

_____

SECTION B - CREW MEMBER

(Completed by data collector if crew member is unavailable)

ESTIMATED CAUSE OF ERROR:_____

_____

_____

HOW WAS ERROR DISCOVERED? _____

_____

HAZARD ANALYSIS

Instructions:

1. Complete at end of test exercise.
2. Attach to TASK-TIME CHECKLIST.

DATE:_____ TIME:_____

TASK SEQUENCE NUMBER:_____

CLASSIFIED AS HAZARD ASSOCIATED WITH:
- ☐ HARDWARE
- ☐ PROCEDURE
- ☐ TRAINING

DESCRIBE HAZARD:_____

_____

_____

INJURY SUSTAINED?
- ☐ YES
- ☐ NO

DESCRIBE INJURY:_____

_____

_____

(450)

HOLLANDER, Paul L., Educational Programme and Planning Consultant, Paul L. Hollander & Associates Inc., Willowdale, Ontario.

"MIND MAPPING" - A TOOL FOR PLANNING, NOTETAKING, COUNSELLING, & INTERVIEWING (Tue P.M.)

The purpose of this workshop is to demonstrate a method of notation known as mind mapping. This method has been used for planning, meeting presentations, letter and report writing, notetaking, and preparing for interviews. The particular application which would be relevant and useful to members of the Military Testing Association would be in planning for interviews, notetaking during interviews or meetings and for later recall.

Rationale and Description of the process:

Mind maps are constructed using "key recall words"--strong nouns and verbs that trigger associations. The central idea or topic is placed in the centre of a circle or square and around the main idea the outside lines are drawn. The lines radiate form the centre of the square or circle to form a non linear organic pattern. On each line is written or printed a "Key recall word". These patterns reflect the natural way in which the mind links and associates ideas. The use of symbols and visual drawing is encouraged to stimulate recall, imagination, and creativity.

Applications:

I have worked and taught with this technique for seven years to many professional groups including nurses, doctors, physiotherapists, social workers, teachers, and accountants. Many people in the health sciences, education, industry, and business have adopted this technique for counselling and interviewing. Many have found that this type of nonlinear notation stimulates creativity and aids recall.

Workshop Format:

I propose to conduct a 90 minute experiential workshop relating the concept of mind mapping to natural memory and recall. The participants will do exercises using linear and nonlinear notation to demonstrate the usefulness of the mind mapping technique. Emphasis will be placed on its practical application to the their life and work.

ILLES, Joseph W., Army Education Center, Ledward Barracks, Schweinfurt, West Germany.

DEVELOPING LOCAL NORMS FOR PREDICTING SUCCESS ON THE GED TEST (Tue A.M.)

The purpose of this work was to develop local norms for the prediction of GED Test Scores based on ABLE III results. Those local norms can prevent placement in unnecessary classes and prevent the frustration or anxiety incurred from premature GED Testing.

Local norming populations are usually more homogenous than are those used to develop publisher's norms, and the resulting data will therefore vary from that published. This is demonstrated in the comparisons of variables selected, errors of estimate and R^2 of local and published norms.

Bi-variate expectancy tables and correlation matrices are compared. Simple linear regressions are graphed and stanines develped for "Quick and Dirty" predictions. Multiple regressions using all five independent variables as well as regressions using only the significant variables are presented.

Included are graphic representations of predicted and actual scores achieved using local and published norms. Data is presented for individual GED tests as well as total GED scores.

DEVELOPING LOCAL NORMS FOR PREDICTING SUCCESS ON THE GED TEST

Joseph W. Illes
U.S. Army Continuing Education Services

PURPOSE

This report presents the results of norming project undertaken during 1979 at a US Army Education Center in Germany. The purpose of that project was to develop local norms for the prediction of achievement on the General Educational Development (GED) test based on Adult Basic Learning Examination (ABLE) test results. The purpose here is to demonstrate that locally developed norms vary sufficiently from the publisher's norms to justify the effort.

RATIONALE

The ability to predict GED test results can be a powerful tool in the hands of a counselor. Using that tool the counselor can reduce the level of frustration and subsequent disilluisonment by examinees who would have failed to qualify for their state high school certificates through the GED test. Using this tool the counselor can recommend preparatory study where qualifying scores are not indicated. On the other side of the coin, a positive attitude toward the GED test could easily be generated in the examinee where the ABLE III test results predict qualifying scores on the GED test.

INSTRUMENT

The ABLE III published by Harcourt Brace Jovanovich, Inc., is a standardized achievement test designed to be used by adults. It consists of three levels: Level I, Level II, and Level III. These levels are intended to discriminate amongst personnel who are operating at grade levels 1 - 4, grades 5 - 8, and grades 9 - 12 respectively. Each level consists of two forms, Form A and Form B. Each form contains the following content areas which have been numbered 1 through 5 for the sake of convenience; (1) Vocabulary, (2) Spelling, (3) Reading, (4) Arithmetic Computation, and (5) Arithmetic Problem Solving. A sixth test score, Total Arithmetic, is nothing more than the combination of ABLE III tests 4 and 5. It is sometimes referred to as Test 6 in the text. In this project, only Form A of level III was used for norming.

The GED tests are designed to measure the general knowledge and achievement of non-high school graduates. The results of these tests are used by the various State Departments of Education as the basis for awarding their high school equivalency certificates and diplomas. The five content areas of the GED are as follows: (1) Correctness and Effectiveness of Expression; (2) Interpretation of Reading Materials in Social Studies; (3) Interpretation of Reading Materials in Natural Sciences; (4) Interpretation of Literary Materials; and (5) General Mathematics Ability.

DELIMITATIONS

Prior to the development of these norms, only those soldiers whose ABLE III scores fell into the fifth stanine or higher as defined on pages 36 - 41 of the ABLE III Handbook were encouraged to take the GED test. It was from this group that these norming standards were constructed. Of course, this screening had its effect on the resulting norms. If a soldier insisted, he was administered the GED test regardless of his stanine placement.

SUBJECTS

Still more homogeneity, at least on a surface level, was introduced by the test population itself. It consisted entirely of soldiers who were between 18-24 years old; had 1 - 2 years military service; and who had been in an overseas area between 6 - 18 months. No other demographic data were collected. The publisher constructed his norms from the test results of approximately 815 soldiers at US Army installations throughout the United States.

PROCEDURE

After the GED tests were administered and the results returned, the results were recorded by Social Security Account Number (SSAN). The ABLE III test results, which had also been recorded by SSAN, were then matched. Pearson product-moment correlations, bi-variate frequency distributions, simple linear regression models, multiple regression coefficients and a table of stanines and percentiles were generated from these data.

The actual computations were performed on a hand calculator, a Texas Instrument TI-51-III. Checks for accuracy were built in by the requirement to pair each of the various ABLE III test results with each other; the pairing of GED test results with each other; and finally, the pairing of each of the ABLE III test results with each of the GED test results.

The linear regression models between ABLE III and GED tests were prepared in graphic form where the correlation coefficients were the highest. The simple regression equation of the form $Y = a + bx$ was not given simply because it would be easier for a counselor to use the graphic form than to compute the predicted score. Because all the soldiers did not take all the GED tests, these samples vary in size between 110 and 113 subjects. Graphic models of each ABLE III test with Total GED tests were also prepared.

Bi-variate frequency tables were prepared in order to show what percentage of the normed population could expect to fall within a specific range of GED scores when given a particular ABLE III test score. These data were prepared using only the data of those 110 subjects who took all the GED tests. A sample of these data are presented with the publisher's data for purposes of comparison.

Multiple regression coefficients using all the ABLE III predictors were developed and compared with those of the publishers.

Multiple regression coefficients using a reduced number of predictor variables were also developed and compared with those of the publishers. Before being entered into the equation, each of these predictor variables were required to survive a t-ratio test by achieving a significance of 0.049 or less. The subsequent F-test of the entire equation then had to meet the same standard of significance of 0.049 or less. Since the area of interest was only that of predictor, the order of the variables being entered into the equation made no difference (Kerlinger and Pedhazur, p 98). For convenience they were entered in numerical order.

The predicted scores generated by the reduced predictor variable equations were then plotted against the scores actually achieved. A similar procedure was used to plot predicted versus actual scores using the publisher's data.

To develop the table of stanines and percentiles, every ABLE III test result available to the testing section was "thrown into the hopper" without regard to GED testing. The number of tests used to derive these data varied from 242 to 247 tests.

DISCUSSION

Correlations As would be expected, the highest correlations were between ABLE III and GED tests of similar content. Disregarding the correlation coefficents which involve Total GED scores, the 25 correlations of the locally normed population range from 0.27 to 0.73; those of the published norms range from 0.31 to 0.73 (See Table 1).

The majority of the correlations of the locally normed group fell between 0.40 and 0.49 (13 of 25 correlations, or 52%). The majority of the publisher's correlations was in the 0.50 - 0.59 group (9 of 25, or 36%). This would suggest a greater homogeneity in the locally normed group. The lower mode of grouped correlations in the locally normed group (0.40 -0.49 versus 0.50 - 0.59) is another demonstration of that greater homogeneity. It should be remembered that any restriction in the range of scores in the predictor and/or the predicted variables acts to reduce the absolute value of the correlation coefficient.

There is still further evidence of tighter grouping and homogeneity. The standard deviation of the GED test is 10 points, since it is a standardized score. The standard deviation of the locally normed population ranges from 6.9 to 7.7 while the publisher's population's standard deviations have a range from 5.9 to 8.6 (See Table 1).

Linear Regression A sample of the linear regression equation for one predictor score to one predicted score is demonstrated graphically in Figure 1. These data were not available in published form, so no comparisons are possible. Since many states require a Total GED score as a way to measure qualification for the state high school equivalency certificate, Figure 2 shows a sample of the regression of an ABLE III test score on the total GED score.

The argument for entry is the ABLE III score on the "X" axis, and is followed vertically upward to the solid regression line. The expected GED

test score is then read, horizontally to the left, on the "Y" axis. For example, if a soldier were to score 33 on the ABLE III Spelling test and we wished to predict his score on GED I, we would enter Figure 1 to determine that he should achieve a score of 50 (+ 5.203 in approximately 67% of the cases tested). This figure of + 5.203 is the Standard Error of the Estimate (SEe). It is graphically represented by the dotted line above and below the solid regression line, and numerically given in the marginal data.

The marginal data also includes the correlation coefficient (r), and the square of the correlation coefficient (r^2). The r^2 tells us what percentage of the deviation from the regression line is attributed to factors contained within the predictor variable; i.e., the particular ABLE III test. To continue the example above, the r^2 of 0.505 simply means that 50.5% of the variation from the regression line can be measured by ABLE III Test 2, Spelling. Yet to be accounted for is the remaining 49.5% of the variation.

Bi-variate frequencies These show that percentage of the normed population which could expect to fall within a specific range of GED scores when given a particular ABLE III test score. These data appear in Tables 2a -d, and were prepared using only the scores of the 110 soldiers who took all the GED tests.

Table 2a, for example, would tell us that of the soldiers in the locally normed group who scored 23 on the ABLE III Spelling test, 53% could expect to score 35 - 44 on GED Test 1. Similarly, the other 47% could expect to score above 45 on the same GED test. The publisher's data appears in parentheses for purposes of comparison. The population mode is shown for local and publisher populations by an asterisk(*). There are differences in the two populations, but no meaningful distinctions can be drawn from the data presented here.

Multiple regression It would seem logical that the more predictors that can "ground into the equation", the more accurate the prediction. Unfortunately, this is not always the case. The principal reason for this is that we do not always have the ideal situation of low correlations among each of the predictor variables, but high correlations with the predicted variable (Kerlinger and Pedhazur, p 73). Two sets of multiple regression equations were developed, and the user may take his choice!

One set of equations used all five of the predictor variables for each GED test. The other set used only those predictor variables which demonstrated that they could make a significant contribution to the accuracy of the prediction. The procedure used to determine this was described earlier, and is a slight variation from the "step down" method. The procedure followed by the publisher is unknown. The report states only that "These equations were generated through standard multiple regression techniques".

A striking difference in the selection of variables appears in the prediction equation for GED test 4 (See Table 3). The locally generated equation uses ABLE III tests 1, 2, and 3; while the publisher used only ABLE III tests 1 and 3. The advantage of the additional variable can be seen through a comparison of the SEe and R^2. The locally normed population's SEe is smaller,

458

and the accounted for variation (represented by R^2 in multiple regression methods) is larger. Conversely, the locally developed equation for GED Test 5, General Mathematical Ability, requires an additional variable to account for the same amount of variation as the publisher's equation.

In Figure 3 further comparisons are made. Here, the predicted scores are plotted along the "X" axis and the achieved scores are plotted along the "Y" axis. The predicted scores of 110 soldiers in the norming group were determined using the locally developed equations; these were plotted against the scores they actually achieved.

The regression line shown in these figures uses the appropriate equation constant as its starting point. In the case of the local norms, that point is connected to the intersection of the average predicted score and the average achieved score. In the case of the publisher's norms, it is connected to a point at the intersection of the predicted score and the publisher's average achieved score. This very graphically demonstrates the variation of the predicted versus the achieved score using the two norms.

Stanines and percentiles These represent only a "quick and dirty" method of screening. The stanine reduces percentile scores to a range of 1 - 9, with a stanine of 1 representing the lowest percentile scores. As would be expected, a stanine of 5 would represent a range around the 50th percentile. Since percentiles, and consequently stanines, rely upon the standard deviation for their values, the stanine on one ABLE III test cannot be applied to another.

The table of stanines and percentiles (Table 4) is read by entering at the "Raw Scores" column at the left and moving across to the right to the appropriate ABLE III test to determine the stanine placement and percentile achieved. No predictions for success on the GED are made; the ABLE III examine is merely placed in relation to his contemporaries.

SUMMARY

The data represented here was an attempt to demonstrate that locally developed norms vary sufficiently from published norms to make them worthwhile. Comparisions have been made to show those differences, and underscore the need for locally developed norms.

REFERENCES

Harris, Richard A.; A Primer of Multivariate Statistics; Academic Press; New York; 1975.

Kerlinger, Fred N.; Foundations of Behavioral Research (Second Edition); Holt, Rinehart and Winston, Inc.; New York; 1973.

Kerlinger, F.N. and Pedhazur, E.J.; Multiple Regression in Behavioral Research; Gardner Press, Inc.; New York 1973.

Popham, W.J. and Sirotuik, K.A.; Educational Statistics, Use and Interpretation (Second Edition); Harper and Rowe, Publications; New York; 1973.

Snedecor, G.W. and Cochrane, W.G.; Statistical Methods (Sixth Edition); The Iowas State University Press; Ames; 1967.

Thorndike, Robert M.; Correlational Procedures for Research; Gardner Press, Inc.; New York; 1978.

"Using the Adult Basic Learning Examination to Predict General Educational Development Test Results"; Test Department, Harcourt Brace Jovanovich, Inc; New York; 1975.

CORRELATIONS BETWEEN GED AND ABLE III TESTS

GED Test	Vocabulary (1)	Spelling (2)	Reading (3)	Arithmetic Computation (4)	Arithmetic Problem Sol Solving (5)	Total Math (4) + (5)=(6)	Mean	Std Dev
(1) Correctness and Effectiveness of Expression	.49 .59	.71 .63	.45 .55	.45 .43	.43 .51	.48 .51	45.2 42.4	7.3 5.9
(2) Interpretation of Reading Materials in the Social Sciences	.55 .57	.46 .36	.56 .59	.44 .35	.53 .53	.52 .47	50.1 45.0	6.9 8.6
(3) Interpretation of Reading Materials in the Natural Sciences	.61 .59	.42 .33	.55 .62	.46 .36	.49 .54	.52 .45	51.0 46.8	7.0 8.1
(4) Interpretation of Literary Materials	.46 .58	.41 .35	.49 .61	.30 .31	.27 .47	.33 .42	49.7 46.1	7.7 7.2
(5) General Mathematical Ability	.43 .43	.39 .32	.34 .44	.67 .67	.67 .69	.73 .74	48.1 42.9	7.3 6.2
Total	.60 .65	.58 .45	.56 .66	.55 .48	.57 .63	.56 .60	243.8 223.4	30.5 30.6
N	110-115	110-113	110-113	110-113	110-113	110-113	110	
X	42.0 42.1	23.0 22.6	42.3 41.5	22.6 22.6	26.0 25.5	47.7 49.1		
Std Dev	7.6 8.0	10.9 9.7	6.9 7.6	8.5 7.7	7.0 6.1	15.5 12.7		

Note: The published correlations appear here in this type.

TABLE 1

(1974)

TABLE 2a

GED Test 1 - Correctness and Effectiveness of Expression

Scores on ABLE III, Spelling	Score Ranges on GED Test #1 20 - 34 %	35 - 44 %	45 - 80 %	N
1 - 4	(100)			(3)
5 - 9	22 (34)	45 (62)	33 (3)	9 (58)
10 - 14	5 (18)	90 (71)	5 (11)	23* (142)
15 - 19	6 (13)	65 (64)	29 (23)	18 (149)*
20 - 24	(6)	53 (67)	47 (27)	15 (144)
25 - 29	(2)	47 (60)	53 (39)	17 (109)
30 - 34		40 (41)	60 (59)	10 (96)
35 - 39	(1)	(18)	100 (81)	7 (84)
40 - 44		(18)	100 (81)	7 (26)
45 - 49			100 (100)	4 (100)
r= .71 (r= .63)			110	(911)

TABLE 2b

GED Test 2 - Interpretation of Reading Materials in Social Studies

Scores on ABLE III, Reading	Score Ranges on GED Test #2 20 - 34 %	35 - 44 %	45 - 80 %	N
1 - 19	(50)	(17)	(33)	(6)
20 - 24	50 (33)	50 (50)	(17)	2 (6)
25 - 29	(19)	(73)	100 (8)	2 (48)
30 - 34	6 (30)	31 (46)	62 (24)	13 (90)
35 - 39	(14)	16 (49)	84 (38)	20 (160)
40 - 44	(7)	14 (33)	86 (59)	30* (219)*
45 - 49	4 (3)	4 (25)	92 (71)	24 (171)
50 - 54	(2)	5 (6)	95 (91)	10 (94)
55 - 60			(100)	9 (21)
r= .56 (r= .59)			110	(821)

TABLE 2c

GED Test 3 - Interpretation of Reading Materials in Natural Science

Scores on ABLE III, Vocabulary	Score Ranges on GED Test 3						N	
	20 - 34		35 - 44		45 - 80			
	%		%		%			
1 - 19		(33)		(50)		(17)		(6)
20 - 24		(33)		(67)	100		1	(6)
25 - 29	20	(27)	20	(56)	60	(18)	5	(45)
30 - 34		(11)	55	(55)	45	(34)	11	(88)
35 - 39		(9)	13	(38)	87	(53)	26*	(160)
40 - 44	5	(7)	18	(25)	77	(68)	23	(217)*
45 - 49		(2)	4	(16)	96	(82)	25	(177)
50 - 54			5	(3)	95	(97)	19	(94)
55 - 60				(5)		(95)		(21)
r= .61								
(r= .59)							110	(814)

TABLE 2d

GED Test 4 - Interpretation of Literary Materials

Scores on ABLE III, Vocabulary	Score Ranges on GED Test 4						N
	20 - 34		35 - 44		45 - 80		
	%		%		%		
1 - 19		(50)		(50)			
20 - 24		(33)		(33)	100	(33)	1
25 - 29		(27)	60	(60)	40	(13)	5
30 - 34		(15)	56	(48)	44	(37)	9
35 - 39		(11)	13	(39)	87	(49)	25
40 - 44	4	(5)	25	(29)	71	(66)	24
45 - 49		(2)	17	(18)	83	(80)	24
50 - 54	5			(6)	95	(94)	21
55 - 60					100	(100)	1
r= .46							110
(r= .58)							110

TABLE 2e

GED Test 5 - General Mathematical Ability

Scores on ABLE III, Total Math	Score Ranges on GED Test #5 20 - 34 %	35 - 44 %	45 - 80 %	N
2 - 19		100 (100)		1 (2)
20 - 24	(45)	100 (55)		2 (11)
25 - 29	(37)	43 (63)	57	7 (38)
30 - 34	(11)	88 (81)	12 (8)	8 (62)
35 - 39	7 (15)	29 (76)	64 (9)	14 (117)*
40 - 44	14 (6)	33 (78)	53 (16)	17 (109)
45 - 49	(4)	42 (62)	58 (33)	12 (117)*
50 - 54	(2)	19 (46)	81 (53)	16 (112)
55 - 59		10 (27)	90 (73)	10 (79)
60 - 64		(11)	100 (89)	9 (76)
65 - 69		20	80 (100)	5 (49)
70 - 74		(4)	100 (96)	2 (26)
75 - 89			100 (100)	3 (14)
80 - 84			100 (100)	4 (3)
r= .73			110	(815)
(r= .74)				

TABLE 2f

Total GED - Total ABLE

Scores on Total ABLE III	Score Ranges on Total GED Less than 175 %	175 - 224 %	225 or Greater %	N
Less than 99		50	50	2
100 - 109		100		3
110 - 119	29	42	29	7
120 - 129		25	75	12
130 - 139		20	80	10
140 - 149		50	50	12
150 - 159		28	72	19
160 - 169		13	87	17
170 - 179			100	7
180 - 189			100	7
190 - 199			100	4
200 - 209			100	4
210 - 219			100	1
220 - 229			100	4
230 - 239			100	1
r= .76				110

REGRESSION EQUATIONS

ABLE III TESTS PREDICTING GED RESULTS

TABLE 3

		Standard Error of Estimate	R	R^2
GED TEST 1	$Y = 24.095 + .224X_1 + .382X_2 - .018X_3 + .145X_4 + .030X_5$	4.80	.768	.590
	$Y = 20.388 + .174X_1 + .258X_2 + .118X_3 + .056X_4 + .209X_5$	3.90	.747	.558
Correctness and Effectiveness of Expression	$Y= 26.157 + .232X_1 + .404X_2$	4.95	.738	.545
	$Y= 24.162 + .286X_1 + .278X_2$	4.08	.719	.517
GED TEST 2	$Y = 23.040 + .206X_1 + .096X_2 + .220X_3 + .065X_4 + .210X_5$	5.22	.680	.463
	$Y = 10.927 + .267X_1 + .032X_2 + .326X_3 + .171X_4 + .490X_5$	6.39	.670	.450
Interpretation of Reading Materials in the Social Sciences	$Y= 19.924 + .236X_5 + .192X_3 + .263X_1$	5.52	.624	.389
	$Y= 14.375 + .419X_5 + .482X_3$	6.63	6.38	.407
GED TEST 3	$Y = 22.298 + .345X_1 + .047X_2 + .170X_3 + .149X_4 + .098X_5$	5.18	.695	.482
	$Y = 13.161 + .268X_1 + .034X_2 + .388X_3 + .149X_3 + .429X_5$	5.85	.696	.484
Interpretation of Reading Materials in the Natural Sciences	$Y= 27.704 + .389X_1 + .308X_4$	5.36	.654	.428
	$Y= 14.982 + .336X_1 + .426X_3$	6.11	.662	.439
GED TEST 4	$Y = 25.224 + .251X_1 + .133X_2 + .271X_3 + .097X_4 = .105X_5$	6.58	.554	.314
	$Y = 17.581 + .232X_1 + .060X_2 + .324X_3 + .154X_4 = .292X_5$	5.30	.693	.453
Interpretation of Literary Materials	$Y= 0.597 + 1.087X_1 + 0.175X_2 - 0.022X_3$	5.09	.762	.581
	$Y= 18.684 + .280X_1 + .378X_2$	5.44	.651	.424
GED TEST 5	$Y = 28.026 + .195X_1 + .045X_2 - .147X_3 + .355X_4 + .347X_5$	4.97	.753	.567
	$Y = 23.552 + .074X_1 + .006X_2 + .010X_3 + .274X_4 + .385X_5$	4.19	.731	.535
General Mathematical Ability	$Y= 24.915 + .152X_1 + .342X_4 + .347X_5$	4.99	.745	.555
	$Y= 24.487 + .493X_6 - .234X_4$	4.10	.745	.555
GED Total Score:	$Y = 124.849 + 1.207X_1 + .710X_2 + .458X_3 + .771X_4 + .600X_5$	19.92	.768	.589
	$Y = 86.562 + 1.010X_1 + .396X_2 + 1.106X_3 + .266X_4 + 1.748X_5$	19.32	.776	.602
	$Y = 135.371 + 1.564X_1 + .781X_2 + 1.118X_4$	20.13	.758	.574

(459)

TABLE 4

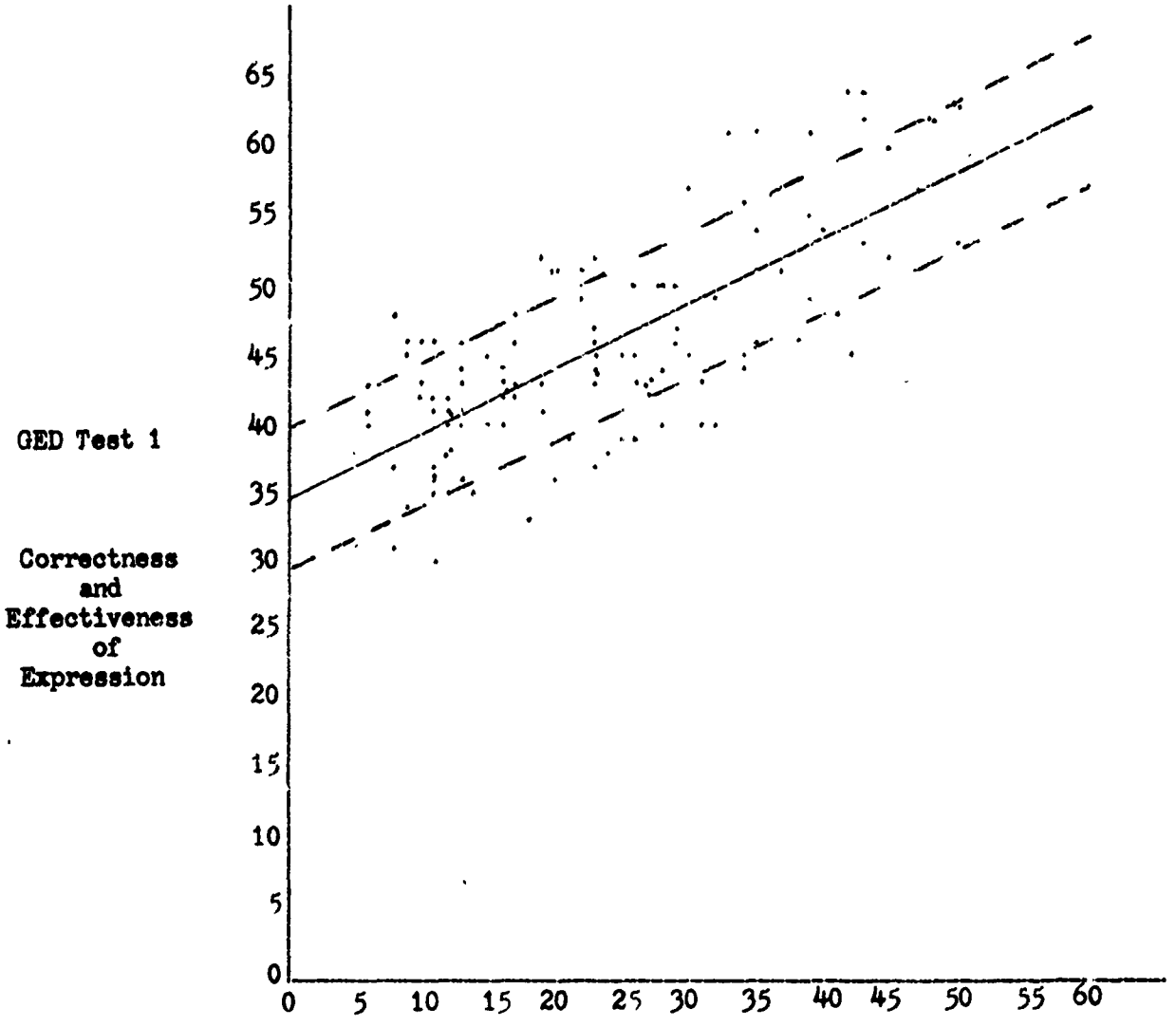
ABLE III Percentile Ranks & Stanines

Raw Score	VOCABULARY n=243		SPELLING n=247		READING n=246		COMPUTATION n=247		PROB. SOLVING n=242		TOTAL MATH n=242	
	%ile	STA	%ile	STA	%ile	STA	%ile	STA	%ile	STA	%ile	STA
60 or higher	99	9			99	9					85	7
59	99	9			99	9					83	7
58	99	9			98	9					81	7
57	98	9			97	9					79	7
56	98	9			97	9					77	7
55	97	9			96	9					75	6
54	96	9			95	8					73	6
53	95	8			93	8					70	6
52	93	8			92	8					68	6
51	92	8			90	8					65	6
50	90	8	99	9	88	7					63	6
49	87	7	99	9	85	7					60	6
48	85	7	99	9	82	7					57	5
47	82	7	99	9	79	7					54	5
46	78	7	99	9	76	6					52	5
45	75	6	99	9	72	6					49	5
44	71	6	98	9	68	6					46	5
43	67	6	98	9	64	6					43	5
42	62	6	97	9	59	5	99	9	99	9	40	5
41	57	5	97	9	55	5	99	9	99	9	38	4
40	53	5	96	9	50	5	99	9	99	9	35	4
39	49	5	95	8	45	5	98	9	98	9	33	4
38	44	5	94	8	41	5	98	9	98	9	30	4
37	39	4	92	8	36	4	97	9	97	9	28	4
36	35	4	91	8	32	4	96	9	95	8	25	4
35	31	4	89	8	28	4	95	8	94	8	23	4
34	27	4	87	7	24	4	94	8	92	8	21	3
33	23	4	85	7	21	3	92	8	90	8	19	3
32	19	3	83	7	18	3	90	8	87	8	17	3
31	17	3	80	7	15	3	88	7	83	7	15	3
30	14	3	77	7	12	3	86	7	79	7	14	3
29	11	3	74	6	10	2	83	7	75	6	12	3
28	9	2	70	6	8	2	80	7	70	6	11	3
27	7	2	67	6	6	2	76	6	65	6	10	2
26	6	2	64	6	5	2	72	6	59	5	8	2
25	5	2	60	6	4	2	68	6	53	5	7	2
24	4	2	56	5	3	1	64	6	48	5	6	2
23	3	1	52	5	2	1	59	5	42	5	5	2
22	2	1	48	5	2	1	55	5	36	4	5	2
21	2	1	44	5	1	1	50	5	31	4	4	2
20	1	1	40	5	1	1	45	5	26	4	3	1
19	1	1	36	4	1	1	40	5	21	3	3	1
18	1	1	33	4	1	1	36	4	17	3	3	1
17	1	1	26	4	1	1	28	4	11	3	2	1
15	1	1	23	4	1	1	24	4	8	2	1	1

TABLE 4

ABLE III Percentile Ranks & Stanines(Cont.)

Raw Score	VOCABULARY n=243		SPELLING n=247		READING n=246		COMPUTATION n=247		PROB SOLVING n=242		TOTAL MATH n=242	
	%ile	STA	%ile	STA	%ile	STA	%ile	STA	%ile	STA	%ile	STA
14	1	1	20	3	1	1	20	3	6	2	1	1
13	1	1	17	3	1	1	17	3	5	2	1	1
12	1	1	15	3	1	1	14	3	3	1	1	1
11	1	1	13	3	1	1	12	3	2	1	1	1
10	1	1	11	3	1	1	9	2	2	1	1	1
9	1	1	9	2	1	1	8	2	1	1	1	1
8	1	1	8	2	1	1	6	2	1	1	1	1
7	1	1	6	2	1	1	5	2	1	1	1	1
6	1	1	5	2	1	1	4	2	1	1	1	1
5	1	1	4	2	1	1	3	1	1	1	1	1
4	1	1	1	1	1	1	2	1	1	1	1	1
3	1	1	1	1	1	1	2	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1
	n=243		n=247		n=246		n=247		n=242		n=242	
	x=39.3		x=22.4		x=40		x=21		x=24.4		x=45.4	
	s=8.5		s=10.1		s=8.6		s=8.4		s=6.8		s=14.1	



ABLE III - Test 2

Spelling

N = 112

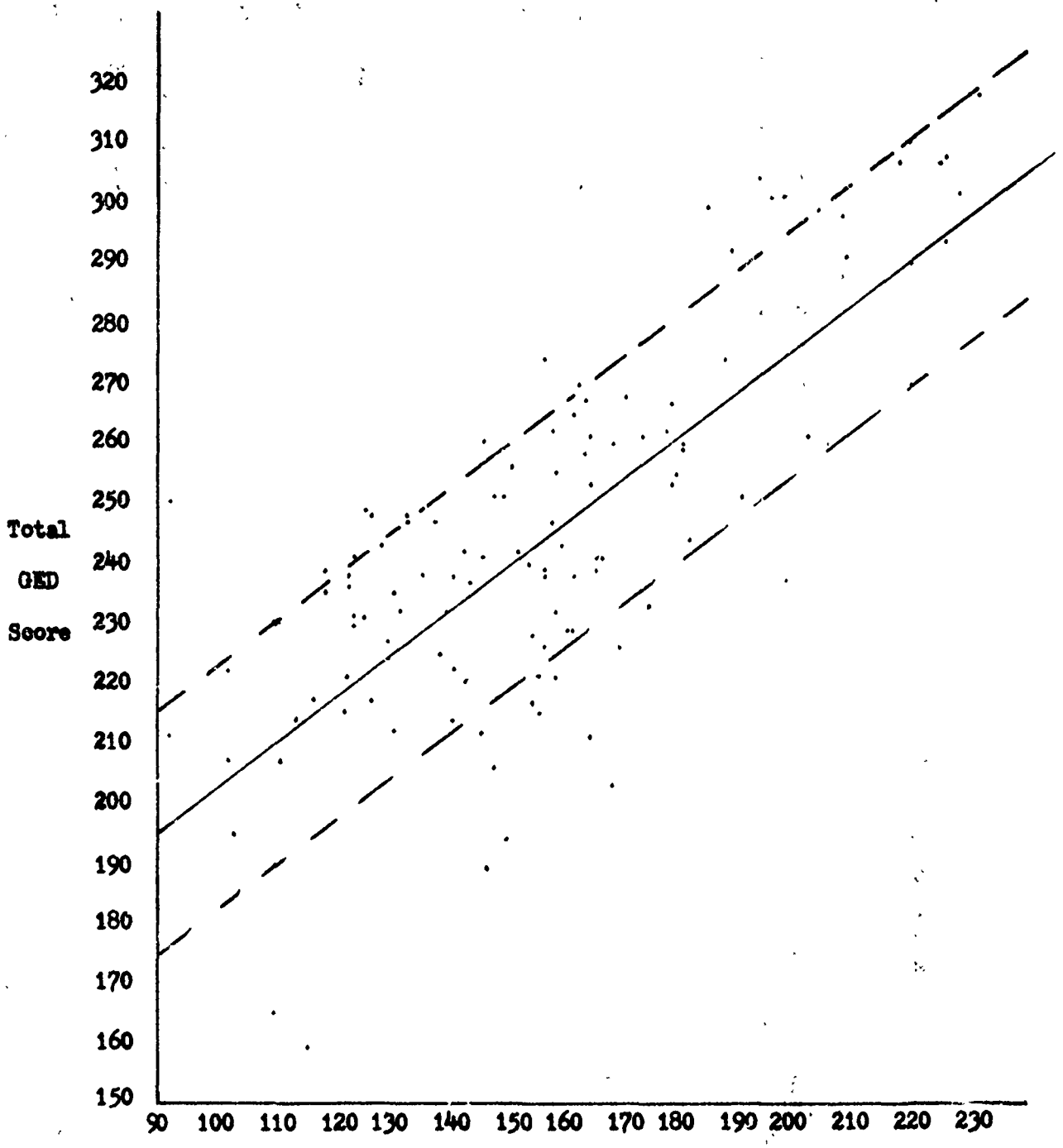
r = 0.710

r^2 = 0.505

SEe = 5.203

FIGURE 1

(468)



N = 110

r = 0.761

r^2 = 0.579

3Ee = 20.10

FIGURE 2

469

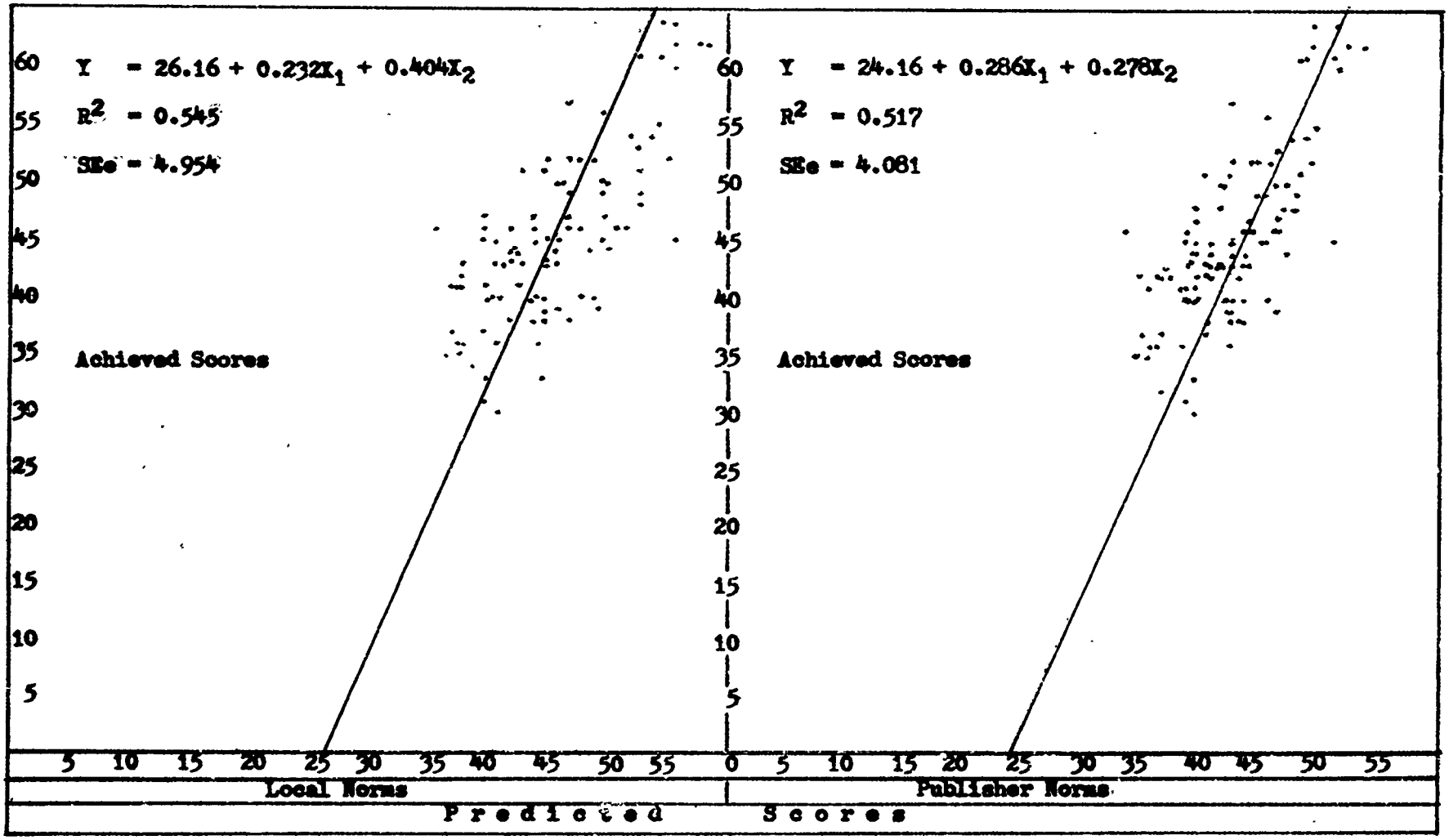


FIGURE 3 I-16

GED 1

Correctness and Effectiveness of Expression

JENKINS, Ronald C., Department of Defense, Meade, Fort George, Maryland.

LONGITUDINAL STUDY OF FOREIGN LANGUAGE SKILLS (Wed P.M.)

A data base to store and manipulate demographic information and language proficiency test results has been established to aid a longitudinal study of growth, maintenance, and attrition of language skills among language-trained members of the military services. The study will include several thousand personnel trained in a number of languages who will be tested systematically using tests designed to measure proficiency as defined in the government-wide standard descriptions. Test results well be studied in the light of the demographic data to determine the effects of assignment on language growth or attrition, length of time required to achieve the various proficiency levels under different circumstances, effect of motivational factors on language learning, and the overall state of foreign language preparedness in the military services.

LONGITUDINAL STUDY OF FOREIGN LANGUAGE SKILLS

Ronald C. Jenkins
U.S. Department of Defense

The training of military personnel in second languages has demanded a considerable amount of Armed Forces time and money since the experiences of World War II graphically demonstrated the need to have soldiers who could speak and comprehend the languages of the enemy as well as the languages of the allies. The need for qualified linguists is even more pressing today when communications capabilities are enhanced by satellites and technological advances have driven the state of the art of warfare to frightening heights. The problem of training foreign language personnel is sometimes further complicated by the less than solid lines of alliances among some families of nations; witness the recent affair in Iran. A common comment heard in circles that must deal with that crisis is, "We have no Farsi linguists."

The language training task is immense. Currently, the U.S. military establishment trains more than 4,000 personnel in more than 35 languages yearly. The training costs run into millions of dollars.

Yet, with the investment of these huge amounts of manpower, time, and money, the training system is without a systematic evaluation and feedback system. The reasons for this void are many and varied but they are not the topic of discussion here. It goes without saying, however, that such a system is needed and it is the topic of this presentation to outline the structure and objectives of an evaluation system that has been designed to alleviate the aforementioned shortcoming.

The training of personnel in second languages is not the only problem faced by the Armed Forces in attempting to meet the challenge that requires foreign language proficient personnel. The challenge is beset by a myraid of problems, some of which are language related and some of which are not. For example, many linguists are trained in language schools to a certain level of proficiency and then assigned to a unit or location where the language skill is allowed to atrophy because there is no need for the language skill at that location at that time and no maintenance programs are available. Also, because not enough of the personnel enlisting in the services today are achieving high enough scores on the language aptitude tests to qualify for language training, the qualifying scores

are being lowered to assure that enough personnel can be identified to fill the billets. These and other problems are further complicating an already complex task. For ease of reference, this phenomenon is usually refered to as the "language problem."

Within the Department of Defense, and especially including the four service branches, a group has been established to cope with the "language problem." The group is known as the Language Work-Management Group and has taken significant steps forward in attempting to solve some of the issues that reduce the effectiveness of the language training program. It is about the activities of one of the subgroups of the Work-Management Group that I wish to discuss this afternoon.

The Work-Management Group has assessed the job needs in an analysis of the tasks to be performed and defined the level of proficiency needed to perform each task. To gauge the effectiveness of any remedial actions that might be recommended to help solve the "language problem," the Work-Management Group created an Evaluation Subgroup. As chairman of the subgroup, I have designed an evaluation program which I believe will prove to be an effective tool for providing information to both managers of the language training program and managers of the product of the training, the product being the linguists themselves.

The program consists of proficiency tests in several languages, a questionnaire, feedback vehicles and a data base. I will discuss each of the items in some detail.

First, a word about the tests. These language proficiency tests are a departure from the normal discrete item or translation tests that one normally associates with a large scale testing program. The tests are contextual and designed to render scores that equate to the standardized language level descriptions that have been adopted for government-wide use. The standardized descriptions recognize five levels of proficiency -- levels 1, 2, 3, 4 and 5. Basically, language level 1 describes someone with the proficiency to ask and answer questions in the foreign language on topics very familiar to him or her with frequent errors in pronunciation and grammar. In short, the individual can order a simple meal, ask for shelter or lodging, ask and give simple directions, and tell time. At this level the linguist is able to satisfy only the most minimal work tasks such as filing or sorting items in the foreign language. At level 2, the linguist is able to satisfy routine social demands and limited work requirements. With extensive use of the dictionary, the linguist can get the sense of routine business or technical articles in his or her field of competence. Level 3 describes an individual with minimum professional proficiency, someone who can speak the language with sufficient structural accuracy and vocabulary to

participate effectively in most formal and informal conversations on practical and professional topics. The linguist at this level is able to read standard newspaper items addressed to the general reader as well as reports and technical articles in his or her special field. Level 4 describes a full professional and level 5 describes someone with a proficiency equivalent to that of an educated native. The military testing program is directed primarily toward language level 2 and entry level 3 because we know that the vast majority of the military linguists fall into these two categories. We know this because we have other contextual language proficiency tests aimed at high level 2 and entry level 3 which are given to military linguists in certain circumstances and we find that very few have a high enough proficiency in the language to "move the needle" on the measuring scale. These tests, which have served well for several years and which correlate highly with supervisor ratings, served as models for the our military tests.

Currently, we have four tests -- in Russian, Chinese, Czech and Korean -- that are being validated under contract and we have tests in Arabic, German, Hebrew and Vietnamese in various stages of development. We plan to construct tests in at least ten languages initially, and eventually in all languages of immediate interest to the military language training program. In addition, we will develop parallel tests to serve as alternate tests in an effort to guard the integrity of the tests.

With luck, we will begin testing and gathering data in 1981.

At the time the tests are administered, a questionnaire will be given to collect selected training and demographic data that will provide insight into language skill growth or attrition. The demographic data will relate to the linguist's current assignment, most recent language training, rank, what type of language work is being performed, and whether the linguist uses the language other than at work. The standard demographic data -- age, sex, educational level, basic language training information, primary language identification, etc. -- will also be solicited. These data will be discussed in more detail later. The testing and the questionnaire will be administered to each linguist at regular intervals throughout the linguist's career, beginning at graduation from basic language training. The test will provide the common metric for growth and attrition and the demographic data will allow judgements about the impact of training, assignments, type of work performed, maintenance programs, etc. on language skill growth or attrition.

The linguist and the linguist's managers will also benefit from the testing. Two feedback vehicles have been designed to provide useful information to both the linguist and his or her managers. The individual linguist will be provided a feedback form which will contain the following information: name of the

individual being tested, social security number, service (Army, Navy, Air Force or Marine Corps), the unit the individual was assigned to at the time of testing, the language being tested, the date of the testing, the percent score achieved on the test, the percent score achieved on the most recent previous test, any measured loss or gain in proficiency, the individual's ranking within his or her unit and within the service, and the language level achieved.

Given this information, the individual linguist will be able to realistically appraise his or her own standing vis-a-vis the general military linguist population (a norm referenced standing) and vis-a-vis real language proficiency (based on the standardized language level descriptions). This represents a major step in giving the linguist a meaningful look at his or her skills.

The linguists' managers will receive information on the linguists assigned to them. The manager (generally a unit commander) will receive a feedback form containing the following information: the name of the unit, the language being tested, the date of the testing, the unit mean score (a percentage), the unit mean score from the most recent previous testing (again a percentage), any measured loss or gain, the unit's ranking within its command, the units ranking within the service, and a list of the linguists in the unit who were tested. Also included will be the scores achieved by the linguists, their percentile ranking within the unit within the command and within the service. The language level achieved by each linguist will also be reported.

A computer data base for the storage, manipulation, and retrieval of the test scores and demographic data has been devised. In addition to the test scores and information from the questionnaires, other data concerning the individual linguist will be collected and stored in the data base. For example, because the tests we are constructing are tests of language competence and not tests of job proficiency, the military job skill test scores may be useful aids in measuring overall growth or attrition of language skills. And, the Defense Language Aptitude Battery (DLAB) score for each linguist will be placed in the data base so an analysis of aptitude and achievement can be made.

The following table is a list of fields currently identified in the data base. Note that there is a header line which contains the name of the linguist and demographic data that is unlikely to change such as name, sex, DLAB score, date of birth, etc. A trailer line will be entered each time the linguist is tested and will contain test scores and demographic data that is current at the time of testing but which might be subject to change such as rank, duty station, etc.

HEADER LINE		TRAILER LINE	
FIELD	EXAMPLE	FIELD	EXAMPLE
NAME	DOE JOHN	LANG TEST AND DATE	RU7810
SSN	012345678	TYPE OF TEST	DLPT
SEX	M		
EDUCATIONAL LEVEL	14	T S PART 1	30
SERVICE	A	E C PART 2	38
DATE ENTERED SERVICE	5006	S O PART 3	
SERVICE JOB CODE	98G3LVEK8	T R PART 4	
DATE OF BIRTH	500317	E TOTAL	68
DLAB SCORE	103		
B L T DATE FINISHED	6906	DUTY STATION	101ST ARMR FT HOOD
A A R LOCATION	1	DATE ARRIVED	7806
S N N WEEKS OF STUDY	47	WORK IN LANG?	Y
E G G COURSE GRADE	89	TYPE OF WORK	XLATE
		RANK	E7
I L T DATE FINISHED			
N A R LOCATION		R L T DATE FINISH	7703
T N N WEEKS OF STUDY		E A R LOCATION	2
R G G COURSE GRADE		C N N COURSE NAME	GARR REFRSH
		E G G WEEKS	5
A L T DATE FINISHED		N COURSE GRADE	87
D A R LOCATION		T	
V N N WEEKS OF STUDY			
G G COURSE GRADE		VOL FOR THIS LANG?	Y
PRIMARY LANGUAGE	EG	O E T L CHILDHOOD	N
OTHER LANGUAGES	SP GR	T X O A ELEMENTARY	N
		H P N HIGH SCHL	N
		E O T G COLLEGE	N
		R S H MAJOR IN L	N
		U I NOW SPOKEN	N
		R S IN HOME	
		E	

As is readily apparent, some of the fields are coded and much of the information appears in abbreviated form. Nevertheless, the information remains readable and can be printed out as a reference work that the services might find useful for identifying personnel for assignments. The data base is designed as a research tool to aid in a long-range study of language skill growth and attrition so that managers of the language training program can made informed decisions on training policy. For example, if an individual is being trained for a job that has been identified as requiring a level 2 linguist, the training period can be tailored to produce that level of competency within a given period of time. The current policy is to train everyone in the same language program to

the highest level of competence no matter what level of competence is needed on the job. The result is that very few of those trained reach a very high level of competence. (For example, although the statistics are not readily available, probably less than 20 percent of the linguists who are tested using the contextual language proficiency tests I mentioned earlier are able to demonstrate mid-level 2 proficiency.) In addition to only a few achieving a fair degree of proficiency in the foreign language, some of those who do are assigned to jobs where their language skills are not used to the extent that their skills can grow or even be maintained at the level they brought to the job. The problem is further compounded in that many of the language jobs in the military are high level 2 or level 3 jobs and the training program fails to produce enough linguists with that level of competency to fill the jobs.

While the evaluation program will not in and of itself solve the "language problem," I feel that it can provide useful information on training courses and length of study required to achieve stated levels of proficiency, effects of various language jobs on language skill growth or attrition, effects of language maintenance or upgrading programs, etc. With this type of information, managers will be better able to make decisions that can result in time and money savings and, more importantly, produce better linguists.

JOHNSON, Carol A. Ph.D., TOKUNAGA, Howard., and HILLER., Dr. J.

VALIDATION OF A JOB ANALYSIS QUESTIONNAIRE THROUGH INTENSIVE OBSERVATION (Thu A.M.)

There are a variety of methods that can be used to perform job task analyses. The most frequently used method is to gather data by administering detailed questionnaires. However, the accuracy of questionnaire estimates of how time is spent on various job activities is open to question.

This study examined the validity of a job task questionnaire designed for officers and NCOs in Infantry Companies and Artillery Batteries by comparing questionnaire data with data that were collected by observing how time was actually spent.

There were two principal findings:

1. The questionnaire data were highly correlated with the observational data;

2. The absolute time-spent estimates reported in the questionnaires were inflated. It was concluded that the questionnaire data are useful for determining the relative amount of time spent on job tasks, and can be used to calculate fairly accurate estimates of the actual amount of time spent.

VALIDATION OF A JOB ANALYSIS QUESTIONNAIRE THROUGH INTENSIVE OBSERVATION

Carol A. Johnson, Ph.D. and Howard T. Tokunaga

McFann.Gray & Associates, Inc.
2100 Garden Road, Suite J
Monterey, CA 93940

Jack Hiller, Ph.D.

Army Research Institute for the Behavioral
and Social Sciences
Presidio of Monterey Field Unit
P.O. Box 5787
Presidio of Monterey, CA 93940

INTRODUCTION

There are many different methods that can be used to obtain the information required for a job analysis, each of which has advantages and disadvantages. Probably the most frequently used method is to gather data by employing questionnaires. A major advantage of utilizing questionnaires is that large samples of job holders can be surveyed and the data analyzed at a relatively low cost.

However, the accuracy or validity of questionnaire estimates of how time is spent on various work activities is open to question. This is particularly true for extensive task inventories. McCormick (1979, p. 133) stated that data regarding validity of task inventory information ". . . are difficult to come by, and there are very limited instances in which data have actually been obtained."

One of the issues is, of course, what to use as the criterion of actual job behavior. Burns (1957) compared worker's diaries with questionnaire responses. He found that individuals tend to under-estimate personal time and over-estimate time spent on activities perceived as important. One drawback of this study, however, is that both data collection methods were based on the subjective perceptions of the employees.

Another study compared questionnaire data with random sampling of job task activities (Klemmer & Snyder, 1972). They report that there was a great deal of variance in the accuracy with which various activities are reported. For example, time spent in face-to-face conversation was under-estimated while time spent reading and writing was over-estimated.

Observations were used by Hartley, Brecht, Pagerey, Weeks, Chapanis, & Hoecker (1977) to assess the accuracy of three types of self-report

information: identification of which tasks had been performed, the relative amount of time spent in those activities, and the amount of time spent in any one activity. A behavior sampling scheme in which behavior was sampled every 30 seconds for one day per subject was used as the criterion. They found that concomitant with an increase in quantitative information was a decrease in accurate time information. Employees ". . . were able to identify fairly well those tasks they had engaged in, they were less able to judge the relative amounts of time spent in those activities, and could not say with satisfactory precision how much time they had spent performing any one activity." These studies indicate that it is preferable to use objective methods to validate self-reports, that the accuracy of self-report methods may vary by job activity category, and that less quantitative information is more likely to be accurate. This latter statement is consistent with McCormick's (1979) conclusion that relative time spent scales are better than an absolute time scale.

The present study was conducted within the context of a three year project. This project is designed to provide management systems and associated training materials to companies and batteries in order to increase time available for combat training. In order to accomplish this, it was necessary to have a clear understanding of how time was actually being spent. While a job task inventory was the basic data collection method, this was supplemented with observations of actual job behavior in order to validate the data collected by the inventory.

The individuals who completed the questionnaire were not necessarily the same ones observed. While this was done due to project constraints, the comparison gives even stronger evidence regarding the extent to which the questionnaire data actually described the jobs covered by the research.

METHOD

Incumbents Included in Study

All of the individuals in this study were in key duty positions in Infantry companies and Artillery batteries within one Army Division. Table 1 gives the number of individuals in each position who were observed and who completed the inventory. Platoon SGTs are not included because of insufficient data.

Table 1

Number of Incumbents in Each Duty Position

	Questionnaire	Observations
Company/Battery Commander	13	12
Executive Officer	11	9
Platoon Leader/Assistant Executive Officer	16	11
First Sergeant	11	10
Squad Leader/Section Chief	47	14
TOTAL	98	56

Questionnaire

One task inventory was developed for use by all levels of personnel. A total of 571 task items were included. The task statements were developed through a comprehensive review of Army publications and interviews with relevant personnel. The inventory asked for the following information:

- If the task is a part of present job.
- How often each task is performed in a typical month.
- How long it takes to perform each task once.
- How much help or assistance is needed to learn each task.
- How much of each task could be done by a civilian.

A list of tasks, by duty position, which a minimum of 25% of the sample indicated they performed was compiled. This list totalled more than the time available. While this would appear to indicate that the absolute time estimates were inflated, it should also be noted that the tasks themselves were not always totally independent. Some were subsets of others. For example, "evaluate unit morale and welfare" is not independent from "analyze feedback from subordinates".

Observations

An observation form was developed and pilot tested. It consisted of a matrix of job content categories and function categories. The form is shown in Appendix A. All observations were conducted by trained members of the research staff. The observation periods averaged 4.2 hours. Each observation period was divided into ten minute recording segments. There were a total of 64 observation periods for Company/Battery Commanders, 28 for Executive Officers, 44 for Platoon Leaders/Assistant Executive Officers, 64 for First Sergeants, and 26 for Squad Leaders/Section Chiefs. At the end of each ten minute period, the dominant behavior was entered into the appropriate cell. The total amount of time spent in each content area could thus be determined.

Recoding Inventory Data Into Observation Categories

In order to compare the information from the two methods of data collection, the items in the task questionnaire were categorized into the broad content areas which matched the observation form. The instructions given to the researchers are reproduced in Appendix B.

Four researchers independently categorized the tasks into the observation content areas. The mode of the judgments was used to determine the content category into which a task was placed. If there were no agreement, or if the judgments resulted in a bi-modal distribution, the task was not included in the analysis.

A total of 69 tasks from the inventory could not be categorized. The exclusion of these tasks resulted in an average of 6.6% of each duty position's hours being uncategorized and, therefore, not used in the analyses. Agreement among the judges was determined by the percentage of the judges agreeing on the model choice for each task. Excluding the 69 unused tasks, the average agreement of the judges was 80%. Including these tasks resulted in an average agreement of 76%.

For each duty position, the total percentage of time spent in each content category was computed for both the questionnaire and the observational data. These totals are presented at the bottom of Table 2. Due to rounding, the values for the questionnaire data may not equal 100%. The original observational data included two categories, Personal Activities and an overall Miscellaneous category, which were excluded from the analysis. This was because the task inventory didn't contain parallel tasks. For that reason, the percentages for the observational data do not total 100%.

In order to validate the questionnaire data using observational data as the criterion, Pearson product-moment correlation coefficients were then computed for each duty position.

RESULTS AND DISCUSSION

Table 2, on the following page, shows the results of the analysis, by duty position. Overall, the correlations between the questionnaire data and the observations were significant, except for the Squad Leader/Section Chief position. This, in fact, provides discriminant validity for the questionnaire, since it was designed to concentrate on management functions and Squad Leaders, as would be expected, were observed to spend relatively little time on management tasks.

It is commonly agreed that relative time spent scales are "better" than absolute time scales (McCormick, 1979). The questionnaire used in this study, however, used an absolute time spent scale. This was done in the hope of providing more definitive information on the actual amount of time spent on tasks than could be extracted from rank-order or relative time spent scales. For example, on a set of tasks which respondents scale as "least amount of time spent", it may be that 20% of the actual time or 1% of the actual time is spent on that set of tasks. The purposes of this project required more precise information, if it could possibly be obtained, specifically to guide the selection of tasks that merited attention for development of managment innovations and/or training material. The high validity coefficients for the management level positions indicate that this measurement approach was successful.

There was, as previously mentioned, a systematic inflation of the absolute amount of time spent in the inventory responses. However, the large amount of variance accounted for, (e.g., 74% of the variance in the observation categories for the First Sergeant was accounted for by the

Table 2

Percent of Time in Each Activity Area as Determined by Inventory (I) and Observation (O), Correlation Coefficient, and Percent of Variance Accounted for (r^2)

	CO/BC		XO		1SG		PL/AXO		SQ LEADER/ SEC. CHIEF	
	% time I	% time O	% time I	% time O	% time I	% time O	% time I	% time O	% time I	% time O
INDIVIDUAL TRAINING	7.15	14.77	3.71	10.35	2.07	5.70	7.49	16.52	.35	24.41
COLLECTIVE TRAINING	3.67	7.11	1.83	3.78	3.04	1.18	1.60	4.53	0.0	2.20
MANDATORY TRAINING	6.82	1.35	1.34	2.14	1.72	.50	.50	5.38	0.0	1.19
MISC. TRAINING	17.25	14.22	7.86	8.88	3.74	4.09	11.52	10.29	*7.09*	*8.81*
LOGISTICS	1.00	.67	*.85*	*1.48*	*.13*	*.87*	.93	3.79	0.0	.34
MAINTENANCE	11.34	3.37	19.79	12.83	.30	2.17	12.82	12.65	22.79	12.37
SUPPLY	8.96	3.19	15.47	3.62	.95	1.98	8.09	7.56	.87	5.59
DETAILS/SUPPORT	2.29	3.92	2.38	3.78	1.51	3.04	.97	1.23	3.88	5.25
HOUSEKEEPING	3.24	.49	.85	1.81	4.65	6.51	.47	.76	20.47	2.71
PERSONNEL MGMT.	19.35	16.12	18.94	11.68	27.97	39.53	19.41	12.37	21.12	6.10
UNIT ADMIN. MISC.	18.92	22.86	26.98	26.32	47.93	27.94	36.22	11.80	23.43	21.69
TOTAL	99.99	88.07	100	86.67	100.01	93.51	100.02	86.88	100	90.66
	$r = .79^{**}$ $r^2 = .62$		$r = .83^{**}$ $r^2 = .69$		$r = .86^{**}$ $r^2 = .74$		$r = .63^{*}$ $r^2 = .40$		$r = .29$ $r^2 = .08$	

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

inventory responses) indicate that the data collected by the inventory was reliably related to the observation criterion. This bias may be easily corrected by dividing each of the inventory task items (or the categories) by the total amount of time reported on the inventory. The proportional time estimates calculated this way may then be used directly, or absolute time estimates may be calculated by multiplying the proportions by the actual amount of time available within a given time span.

The job task inventory used in this study required respondents to give absolute time estimates rather than more traditional relative time estimates. The data from the inventory was validated against observations of job holders, and thereby, determined to be accurate. Thus, the current, widely held position that relative time spent rather than absolute time spent scales should be used may have been prematurely adopted, especially in light of the few empirical studies published. Based on the sizeable validity coefficients found in this study for an absolute time scale, further research is clearly warranted.

REFERENCES

Burns, T. Management in action. Operational Research Quarterly, 1957, 8, 45-60.

Hartley, C., Brecht, M., Pagery, P., Weeks, G., Chapanis, A. and Hoecker, D. Subjective time estimates of work tasks by office workers. Journal of Occupational Psychology, 1977, 50, 23-36.

Klemmer, E. and Snyder, F. Measurement of the time spent communicating. Journal of Communication, 1972, 22, 142-158.

McCormick, E. Job Analysis: Methods and Applications. N.Y.: American Management Association, 1979.

APPENDIX A

		BDE/Division Support	Training				Personal Activities	Misc.
			Collective Skills	Individual Skills	Mandatory Trng Topics	Misc.		
Planning/Preparing/Developing								
Providing Information								
Requesting Information								
Receiving Requested Info								
Receiving Orders/Direction, etc.								
Performing								
Receiving Training								
Conducting Training								
Directing	Ordering							
	Staffing							
	Delegating							
	Organizing							
	Misc.							
Observing Performance								
Inspecting								
Reviewing (reports, etc.)								
Counseling								
Traveling								
Down Time								

JOH-1-8

		Unit Administration					
		Logistics	Supply	Mainten-ance	Unit (BN/CO) House-keeping	Personnel Manage-ment	Misc.
Planning/Preparing/Developing							
Providing Information							
Requesting Information							
Receiving Requested Info.							
Receiving Orders/Direction, etc.							
Performing							
Receiving Training							
Conducting Training							
Directing	Ordering						
	Staffing						
	Delegating						
	Organizing						
	Misc.						
Observing Performance							
Inspecting							
Reviewing (reports, etc.)							
Counseling							
Traveling							
Down Time							

APPENDIX B

PROCEDURES FOR CATEGORIZING TASK QUESTIONNAIRE ITEMS INTO OBSERVATIONAL DATA CONTENT AREAS

The objective of this exercise is to recategorize the items on the task questionnaire into the content areas used in the observations conducted during the first year. By making the content areas of both methodologies the same, direct comparisons can be made between the observed time versus the self-reported time spent in each content area. In other words, we will be able to compare objective (observations) and subjective (responses to questionnaire items) estimates of time spent in work related activities.

The purpose of this paper is to introduce the content areas to those who will be doing the categorization. Each of the observational content areas will be defined, described, and include examples extracted from the observer's notes. The goal of these guidelines is to show how the observers perceived each content area, using their own judgment as well as stated rules in categorizing the observed officer's activities.

It is not the purpose of this paper to specifically define how the task questionnaire items are to be categorized. In order to insure the validity of these content areas, it is imperative that each rater should do their categorizing independently, using their own criteria. But perhaps by showing the observer's perspective, it may be easier to categorize those tasks which to not fall cleanly into any one category.

Each content area will now be presented, starting with a definition of that area; this definition is identical to the one used by the observers. This will be followed either by a description of the content area, examples of the observer's notes which belonged to this area, or both.

JOHNSON, James H., and NOVAK, Kathy., Psych Systems, Baltimore, Maryland.

HUMAN-ENGINEERING A COMPUTERIZED TESTING STATION (Thu P.M.)

A fundamental problem for developers of computerized testing systems is human-engineering the testing station hardware and software. The hardware must be appropriate for the minimal ability levels that are likely to be encountered during broad scale testing. Complicated keyboards may need to be replaced with test specific response panels. Furthermore, displays need to be of a size and intensity such that testing will not become tedious. The software for item presentation and response collection must make provisions for the fact that most examinees will have had previous experience only with paper-and-pencil tests and will be naive to the computer medium. Simple instructions, trial blocks, consistent presentation patterns, and error correction routines must be provided. This paper details considerations, such as these, that designers of tests for the computer medium must take into account to develop valid computerized approaches to testing.

HUMAN ENGINEERING:

A COMPUTERIZED TESTING STATION

James H. Johnson & Kathy Novak

In 1975 Johnson and Williams reported on the development and implementation of a large scale on-line computer based testing system for patients at the VA Hospital in Salt Lake City. This was the first major effort to make use of an on-line computer terminal for routine psychological testing. Evaluation results with this system showed that it was cost effective, and, still well liked by test takers (Klinger, Johnson & Williams, 1976; Klinger, Miller, Johnson & Williams, 1977). These results were found despite the fact that efforts to human engineer the terminal system were only very rudimentary (Cole, Johnson & Williams, 1975).

Concurrent with the development of this operational psychological testing system were many other developments important to on-line computer based assessment. Among the most important of these was the introduction of adaptive testing methodologies. Adaptive testing strategies are characterized by large item pools that are only partially administered to each particular test taker depending on initial item responses and the individual's level of functioning. Because the adaptive approach concentrates on examining primarily in the area of ability appropriate to the individual being tested, it offers the possibility of narrow band assessment with only a small number of items being presented to any one test taker.

The positive empirical findings by Johnson and his colleagues in Utah and the obvious theoretical advantages of adaptive strategies over other assessment techniques leads us to expect that an ever-growing percentage of future testing packages will be developed for the on-line medium. As a result, it appears that we will be entering a new age of research in psychological assessment. In this era several new topics that were previously only tangentially considered will emerge as relevant concerns for test constructors. The present

paper is concerned with one of these new assessment research areas, human engineering the computerized testing station. Our purpose in presenting this paper is to review the nature of the research problem, to suggest some approaches for minimizing the problem, and to make some proposals for future research in the area. In essence, we are proposing that efforts need to be taken to bring about "man-computer symbiosis" (Licklider, 1960) in the adaptive testing area, and that these efforts will have great importance to the success of this new testing approach.

HUMAN ENGINEERING TESTING STATIONS

AS A DESCENDANT OF RESPONSE BIAS

Some years ago, Edwards, (1957) and Jackson and Messick (1961) argued that most of the variance in objective personality testing could be accounted for by response sets. Even though Block (1965) showed quite clearly that this was not the case, relevant concerns remain in the testing area about minimizing the amount of invalid test variance due to the response process. Figure 1 shows a schematic of the computerized test response process. It is apparent from this schematic that variance due solely to styles of response is only a small part of the invalid variance possible in a testing situation. (Please refer to Figure 1). Errors due to reading and understanding on the input side, and due to slips between the thinking process and the physical response movement on the output side are also likely to lead to undesirable test variance.

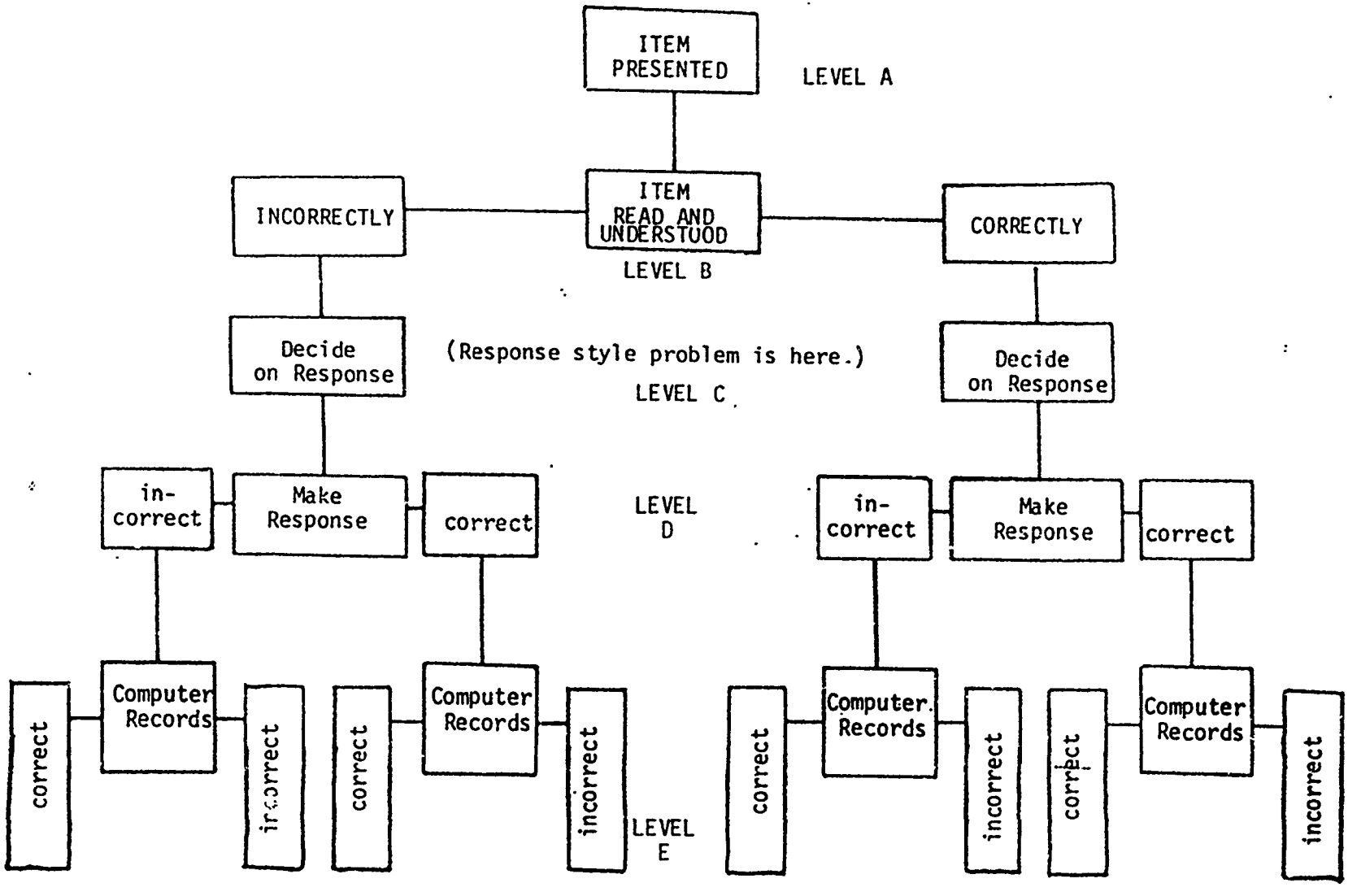
Because of this possibility of errors involving the person-test system interface, it is important to consider ways to human engineer computerized testing stations in order to reduce unwanted variance. One can see that while this problem is a variation on that first considered by Edwards and then by Jackson and Messick, it is potentially of much greater significance to developers of assessment approaches for the on-line computer medium. Furthermore, it is especially important in adaptive testing due to the fact each individual response is directly related to all following item presentations.

EXAMPLES OF HOW TO HUMAN ENGINEER

A TESTING STATION

From Figure 1 it is clear that human engineering a testing station must be completed at both the input and the output level. Therefore, these problem areas will be treated separately.

FIGURE 1


ITEM PRESENTED
LEVEL A
INCORRECTLY
ITEM READ AND UNDERSTOOD
CORRECTLY
LEVEL B
Decide on Response
(Response style problem is here.)
LEVEL C
Decide on Response
in-correct
Make Response
correct
LEVEL D
in-correct
Make Response
correct
correct
Computer Records
incorrect
correct
Computer Records
incorrect
LEVEL E
correct
Computer Records
incorrect
correct
Computer Records
incorrect

INPUT PROBLEMS

The first possibility of error results from misunderstanding instructions related to taking the test. Greist and Klein (1980) have dealt with this problem by giving very explicit instructions and then testing for understanding. Johnson and Williams (1980) found that in actual practice this approach slowed test takers and often led to irritation on their part. They preferred simple instructions with "on-call" personnel to give help if needed. Probably some compromise is required here. Moderately detailed instructions with simple probing for misunderstanding coupled with "on-call" staff help may be the best solution to assure complete understanding. This problem of whether or not to have highly detailed computerized instructions for the naive user has been considered in detail by a number of researchers (e.g., Kennedy, 1975; Schneiderman, 1980). The concensus of previous research results seems to be that instructions should go from more detailed to less detailed as the user becomes more familiar with usage. Rouse (1977) suggests that the best solution is a dynamic system where the user chooses the amount of instruction he/she requires. Furthermore, he suggests that a simple game or "trial run" is probably the best way to develop such a dynamic system. What we are presently proposing is consistant with this thinking.

The problem involved in assuring correct reading and understanding of items are more difficult. First, test developers need to refer to the human factors literature to assure that illumination, character size, and item length are optimal for error reduction (e.g., Schneiderman, 1980). Second, they should probably check proposed items with alternates for readability and comprehension before inclusion in a test. Finally, they should use some kind of screening index such as the Q1 (Johnson, Williams Klinger, & Giannetti, 1977) to predetermine whether a test taker is capable of understanding the item pool.

Careful attention to these aspects of computer based testing can help to reduce many of the errors possible at Level B of the testing process shown in Figure 1.

OUTPUT PROBLEMS

Perhaps it is relevant to begin this section with the ancient proverb attributed to Palladus, "There is many a slip twixt the cup and the lip." As is illustrated at Level D of Figure 1, it is possible for the respondent to read an item correctly, make a

correct decision about how to respond to the item, and still give an incorrect response. For example, classic human factors research on keyboard usage has shown that even experienced keyboard users have a finger related error rate ranging from .00414 to .01171 (Hirsch, 1976).

Paying attention to some of this research can help to overcome these problems. From Hirsch's data we know that the chances of making an error on choices "C", "D", or "E" are approximately 50% greater than they are on choice "A". Therefore, it seem obvious that there is a need to use specifically designed key pads for response collection. These key pads should be based on the human factors literature and should be designed to minimize error rate as well as error rate variance across choices. If we are going to have uncontrollable error, then it should at least be random across response choices.

Moreover, system designers need to take account of the fact that errors will always be made, and, therefore, should develop simple error correction procedures. A "backup key" on the keyboard is one good solution. If a respondent recognized that he/she may have made an error, then pushing the "backup" key will allow trying the item again.

There should also be a procedure whereby errors in recording by the computer can be made apparent to the test taker for final correction. These errors at Level E of Figure 1 are not frequent, but are still sure to occur as a result of hardware and software failure.

What can be done about these errors? One acceptable approach is for the computer to randomly review answers with the respondent at the completion of testing. If the respondent does not agree with this review, then this fact could be reported to invalidate results.

PROBLEMS RELATED TO BOTH INPUT AND OUTPUT

The most obvious problems related to both input and output are those of attitude and anxiety (Schneiderman, 1980). Those with negative attitudes towards computers have been shown to perform more slowly and to make more errors (Walther & O'Neil, 1974). Anxiety also has been related to poorer performance (Denny, 1966). Thus, adaptive testing constructors need to work carefully to reduce poor attitudes and high anxiety.

Schneiderman (1980) has further noted that the issues of control and response time are related to attitude and anxiety. He cites evidence that users feel

worse about interactive computer systems when they lack control over them. Messages or instructions that are patronizing, threatening or condeming lead users to feel a loss of control. Therefore, careful attention should be given to developing an interactive dialogue that allows the user to feel as though he/she is in charge of the situation. Wide variance in the amount of time taken by the computer to respond to the user as well as extended response latencies have also been related to poorer attitudes, high anxiety, and consequent poorer performance (Miller, 1977). Thus, every effort should be made by adaptive systems developers to use approaches that lead to fast and consistent response latencies. This is an especially difficult problem to overcome because adaptive testing requires calculations between each item presentation and the calculations become more difficult after more items have been presented, thus, taking more time. One solution to this problem is to use a fixed latency that is as long as the most difficult item selection computation will take. Of course, this solution is at odds with the need to have generally brief latencies, but Schneiderman suggests that this is the best possible answer to the problem.

FUTURE RESEARCH

It is now beginning to be clear that the problems associated with the person-computerized testing station interface are numerous. Furthermore, it is evident that we know only a small amount about how to solve these problems. A great deal of research is going to be required before we have a good understanding of the problems and solutions related to the interface between person and computer.

To begin with it will be important to know the normal error rate at each of the levels in Figure 1. Knowing this will allow us to test newly developed strategies for improvement in error rate. The development and testing of various new strategies for error reduction probably constitutes the largest research need in this area. For example, one such problem deserves immediate attention. Is a one key or two key (the second a RETURN to allow reviewing results) approach better for data entry and error detection? As another example, it might be that different personality states are related to positive effects from different instruction sets. Use of a computerized testing station could begin with a brief screening test which would determine whether one or another set of instructions would have better results. By following the diagrams presented in Figure 1, it is easy to see that the research problems are almost endless.

SUMMARY

In this paper we have attempted to give an overview of some of the issues related to human engineering a computerized testing station. It is our contention that most of these problems have been overlooked in the present and past literature pertinent to computerized testing stations. Additional work in this area will surely enhance outcomes for computerized adaptive testing systems.

REFERENCES

Block, J. The Challenge of Response Sets, New York: Appleton-Century-Crofts, 1965.

Cole, E.B., Johnson, J.H., and Williams, T.A. Design considerations for an on-line computer system for automated psychiatric assessment. Behavior Research Methods and Instrumentation, 1975, 7, 199-200.

Denny, J.P. Effects of anxiety and intelligence on concept formation. Journal of Experimental Psychology. 1966, 72, 596.

Edwards, A. The Social Desirability in Personality Assessment Research. New York: Holt, 1957.

Greist, J.H. and Klien, M.H. Computer programs for patients, clinicians, and researchers in psychiatry. In J.B. Sidowski, J.H. Johnson, and T.A. Williams (Eds.) Technology in Mental Health Care Delivery Systems, Norwood, N.J.: Ablex, 1980.

Hirsch, R.S. Human Factors in Man Computer Interfaces. San Jose, California: IBM Human Factors Center, 1976.

Jackson, D. and Messick, S. Acquiescence and desirability determinants in the MMPI. Educational and Psychological Measurement, 1961, 22, 771-790.

Johnson, J.H. and Williams, T.A. Using on-line computer technology to improve service response and decision-making effectiveness in a mental health admitting system, In J.B. Sidowski, J.H. Johnson, and T.A. Williams, (Eds.) Technology in Mental Health Care Delivery Systems. Norwood, N.J.: Ablex, 1980.

Johnson, J.H. and Williams, T.A. The use of on-line computer technology in a mental health admitting system. American Psychologist, 1975, 30, 388-390.

Johnson, J.H. and Williams, T.A., Klinger, D.E. and Giannetti, R.A. Interventional relevance and retrofit programming: Concepts for the improvement of clinician acceptance of computer generated assessment reports. Behavior Research Methods and Instrumentation, 1977, 9, 123-132.

Kennedy, T.C. Some behavioral factors affecting the training of naive users of an interactive computer system. International Journal of Man-Machine Studies, 1975, 1, 817-834.

Klinger, D.E., Johnson, J.H. and Williams, T.A. Strategies in the evaluation of an on-line computer-assisted unit for intake assessment of mental health patients. Behavior Research Methods and Instrumentation, 1976, 8, 95-100.

Klinger, D.E., Miller, D.A., Johnson, J.H. and Williams, T.A. Process evaluation of an on-line computer-assisted unit for intake assessment of mental health patients. Behavior Research Methods and Instrumentation, 1977, 9, 110-116.

Licklider, J.C. Man-Computer symbiosis. IEEE Transactions on Human Factors in Electronics, 1960, HFE1, 4-11.

Miller, L.H. A study in man-machine interaction. Proceedings of the National Computer Conference, 1977, 46, 409-421.

Rouse, W.B. Human-computer interaction in multitask situations. IEE Transactions on Systems, Man and Cybernetics, 1977, 5, 384-392.

Schneiderman, B. Software Psychology, Cambridge, Massachusetts: Winthrop, 1980.

Walther, G.H. and O'Neil, H.F. On-line user-computer interface: The effects of interface flexibility, terminal type, and experience on performance. Proceedings of the National Computer Conference, 1974, 43, 379-384.

DEVELOPMENT OF A MARINE SAFETY PERSONNEL TRAINING PROGRAM FOR THE UNITED STATES COAST GUARD

D. T. Jones, P.E.
U. S. Coast Guard
Office of Research and Development
Washington, D. C. 20593

ABSTRACT

For years the Coast Guard has been able to obtain qualified mariner's to help in its marine safety missions through its Licensed Officers in the Merchant Marine (LOMM) Program (previously known as the 219 Program). This program, which allows qualified merchant mariners direct entry into the Coast Guard as a commissioned officer, has recently been attracting fewer and fewer merchant mariners. This is probably because of higher pay in the merchant marine as well as the current lack of interest in working for the Coast Guard. With a declining number of experienced merchant mariners entering the Coast Guard, the number of trained experienced personnel able to handle the complex problems of the increasingly technological marine environment are decreasing. In order to meet the needs of a well trained, educated Coast Guard, capable of dealing with the increased technology of today's merchant marine, a well defined training program must be established.

There is currently not a well organized, well defined training program that acccurately meets the current and projected needs, goals, methods and procedures of the Coast Guard, and provides the education and training necessary for Coast Guard personnel (Officer and Enlisted) to meet the demands of today's and tommorrow's Marine Safety Office.

The objective of this work, currently getting started, is to develop a managable, effective, and comprehensive training program to provide trained marine safety personnel from entry levels through upper management (District Staff). The work will begin with a task analysis to determine the skill and knowledge requirements in all current and planned marine safety (MS) task areas (including special projects under development by the office of Research and Development such as the Marine Safety Information System (MSIS)). From this and a review of all current and planned Coast Guard funded training programs in the Marine Safety field, the information necessary to develop a comprehensive training program will be developed.

This paper outlines the nature and scope of the problem and the approach that will be taken to develop the training program necessary to insure that the U.S. Coast Guard will have trained knowledgeable personnel in the Marine Safety area.

The opinions expessed in this paper are those of the author, who is solely responsible for the accuracy of the contents, and does not necessarily reflect the views of either the U.S. Department of Transportation or the U.S. Coast Guard.

INTRODUCTION

Since the creation of the Revenue Marine in 1790 to the present, the U.S. Coast Guard has had several major functional areas of responsibility. Three of these, Commerical Vessel Safety (CVS), Marine Environmental Protection (MEP), and Port Safety and Security (PSS), have been recently combined under the auspices of marine safety. Marine Safety Offices (MSO) across the country currently handle the majority of the Coast Guard's compliance effort in these three major functional areas.

While Congress provided the Coast Guard with certain specific powers and constraints to enforce marine-related laws and regulations, several dissimilar approaches to law enforcement have evolved as a result of the variance between various statutory authorities. For example two methods of achieving CVS program objectives are: withholding a Certificate of Inspection from certain classes of vessels that do not comply with the minimal safety standards prescribed by laws and regulations; and withholding a License or Merchant Mariner's Document from any person who does not comply with the requirements of appropriate Federal laws and regulations.

The MEP Program, on the other hand, is not restricted to any certain category of clientele since its requirements apply to the general public. Unlike the CVS Program, there are neither licenses or documents issused, nor are potential pollution facilities (except certain vessels and certain bulk liquid facilities) inspected or certificated. Origionally, law enforcement in the MEP Program was one of "crime and punishment" since no concrete pollution guidelines existed. However, since the program's inception, the emphasis has shifted from the punishment aspect to the prevention aspect, because, in the final analysis, prevention of pollution incidents is the only true way to protect our environment.

The PSS Program focuses upon a more limited clientele involving both port facilities and merchant shipping. It is not, however, nearly so restricted to finite safety prevention measures as the CVS Program, since the scope of PSS activities includes foreign flag vessels, U.S. merchant ships, and certain waterfront facilities.

For years the Coast Guard has been able to obtain qualified mariner's to help in these missions through its Licensed Officers in the Merchant Marine (LOMM) Program (previously known as the 219 Program). This program, which allows qualified merchant mariners to enter the Coast Guard, has recently been attraching fewer and fewer merchant mariners. This is probably because of higher pay in the merchant marine as well as the current lack of interest in working for the Coast Guard. With the declining number of experienced merchant mariners entering the Coast Guard, the number of trained experienced personnel able to handle the complex problems of the increasingly technological marine environment are decreasing.

Today all personnel assigned to CVS, MEP, and PSS program-funded billets may be refered to as "marine safety" personnel and it is the Commandant's goal that they be well trained in most aspects of marine safety activities. However, it is recognized that not all personnel will have the same level of experience in "marine safety", regardless of rank or grade, because the term encompasses not only functions such as marine inspection, pollution investigation, etc., but any other operational function at a marine safety unit.

(504)

CURRENT TRAINING

On-the-job-training, which deals with improving the qualifications of individual marine safety personnel, through a program of hands-on experience, is the responsibility of the HQ program manager, who provides policy and direction to district commanders and commanding officers of field units. On-the--job-training coupled with formal CG residence Courses and programmmed specialized training (ie., short courses at industry facilities), presently comprise that phase of marine safety field qualification training. Because the responsibilities for qualifying marine safety marine safety field personnel are diffused among program managers, a support manager, and unit commanding officers, this aspect of training is divided into three parts, termed 'Level I', 'Level II', and 'Level III'. In general, Level I is Mandatory training, Level II is desirable training, and Level III is optional training.

Level I training is designed for marine safety personnel begining their first tour in marine safety duty. The training consists of an optimal period of 2 to 3 months familiarization at the unit prior to attendance at formal resident courses, then a period of resident training, and finally completion of a departmental on-the-job qualification program.

There are currently several formal residence courses available for Level I training. The primary officer formal training course is the Marine Safety Basic Introduction Course (MSBIC). The MSBIC is 12 weeks in length and is mandatory for all first-tour commissioned officers assigned to MSO's, MIO's, COTP units, and selected HQ and district billets. The Marine Environment and Systems Petty Officer Course (MESPOC), provides basic indocrination to petty officers of all ratings who are on their inital tour of duty in marine safety. The MESPOC is five weeks in length and deals primarly with PSS and MEP Program functions. Some CVS Program activity is also presented to provide the basic knowledge upon which to conduct on-the-job-training as assistant marine inspectors. This course is for all Petty Officers routinely assigned to marine safety duties.

Level II training consists of training which supplements Level I training. While Level I training provides the basic knowledge and skills for all marine safety personnel to perform marine safety functions, Level II provides the higher level of knowledge and/or skill for designated personnel to perform the more specialized and technical duties involved with marine safety. Level II training, therefore includes formal specialized resident training that is necessary from the program manager's viewpoint to improve performance. There are presently approximately 50 Level II courses being conducted by various colleges, universities, private industries, and government agencies. Each course is designed to meet a particular professional performance need in marine safety. As a general rule, Level II training is intended for persons who have completed their mandatory Level I training.

Level III training consists of optional training or education which the participating officer or petty officer believes to be of general benefit to him in the performance of his marine safety duties. This is analogous to off-duty training. Examples of Level III training include: (1) Off-duty community college courses in diesel engines, English composition, welding, etc. (2) Participation in Propeller Club seminars, Civil Defense workshops, etc., and (3) Correspondence courses in fields related to marine safety. This level of training is left to the participant. When funding is available through existing Coast Guard or other agency programs, command support is indicated; otherwise the command is not required to participate in this level of training.

5-05

Notwithstanding the comprehensive training plan described above, several problems exist in the training area which must be addressed:

- o A significant percentage of Level II training is not specifically designed to meet Coast Guard marine safety requirements.
- o Field units often lack adequate training aids for on-the-job training.
- o Standards for personnel qualification (of CG Personnel) are often vague, e.g. in the inspection area.
- o Trainees have difficulty securing space in the marine safety school.
- o There is no formal on-the-job qualification program for enlisted personnel.
- o Personnel are often not rotated from division co division in their MSO during their initial tour.

Although steps have been taken toward correcting these problems, a more comprehensive response to the training problem must be developed. There still remains a clear need for:

1) identification of skill and knowledge requirements necessary to perform each the tasks performed by field personnel and to determine what qualifications are necessary,
2) review of the Coast Guard's existing training programs to identify gaps, and
3) development of recommendations for effective and comprehensive training and suggested training objectives.

PERSONNEL STRUCTURE

The general structure of the MSO calls for 75 different types of billets which require the services of commissioned officer. (See Figure (1) for a typical Marine Safety Office (MSO) organization). An additional thirteen MSO billets call for enlisted personnel. It should be noted that enlisted personnel may fill some billets designed for commissioned officers, in the special circumstance that the enlisted person has the knowledge and experience to do so.

It is evident that the MSO personnel organization already provides a strong flexibility in expertise to cover small and large operations. It has an inherent capability for expansion or contraction as the need arises. This flexibility is highly dependent on the training and resourcefulness of personnel at all levels in the MSO structure. Any expanded future training scheme should be designed to promote flexibility because the MSO responsibilities are getting larger by dictate; viz the additon of tank ship inspection and the offshore drilling ship/platform operation responsibilities.

Common sense suggests that the very existence of an MSO should enhance and integrated operation which can spin off into an integrated CVS, MES, PSS training approach. The questions, of course, is how effective can an MSO be without a strong degree of specialization. (It is noted that the GAO maintains that chere is not enough specialization, at least for CVS).

Notwithstanding that specialization is implicit by billet definition and the related OJT and experience therein, any expanded training scheme should at least maintain a common curriculum approach for CVS, MEP and PSS at all levels.

There is a large diversity of disciplines and technologies that must be captured and maintained by MSO personnel. Their duties call for increasing understanding of advancing technologies in order to perform their duties effectively. To accomplish this, specialization at all "hand-on" operating levels is a very powerful argument.

(5-06)

Specialization is an inherent outcome of our technological existence and for MSO is primarily created by previous experience and OJT. As the majority of MSO personnel "rotate" into the service from their normal Coast Guard disciplines, they bring with them a natural specialization to the MSO. It would seem very inappropriate not to use such specializations to the fullest extent possible.

TRAINING PROGRAM

The objective of this work is to provide information on skills and knowledge requirements in all current and planned marine safety (MS) task areas (including special projects under development by the office of Research and Development such as the Marine Safety Information System (MSIS)). Information will be obtained on current and projected marine safety training courses and on-the-job-training. From the information obtained a managable, effective, and comprehensive training program will then be developed to provide trained marine safety personnel from entry levels through upper management (District Staff).

To accomplish these objectives, the following tasks must be performed:

1) A work systems analysis of all marine safety functions. The analysis will cover all tasks and levels in all marine safety functions, ie., Inspection Department, Port Operation Department, Investigation Department. Document and License Department, and Administrative Staff. In addition to existing tasks, the analysis will identify any new tasks that may be expected to result from changes in equipment, organization, laws, regulations, rules and procedures. The analysis will specify the skills and knowledge neded for each task identified.
2) A review of existing marine safety training, both formal and on-the-job training. This review will identify those skill and knowledge areas required for marine safety functions and the level to which they are taught.
3) A review of existing training methods and techniques. This review will provide information on the training methods and techniques applicable to marine safety training.
4) Recommendations for training methods and techniques that will provide managable, effective, and comprehensive training for each training requirement identified.

In order to achieve all the objectives outlined above, a systems approach must be used. Such an approach views the work organization systematically. It must take into consideration all aspects of the work organization (system). These include: the organizational goals (desired productivity/ outputs): what work must be done to achieve these goals; and what attributes workers must have to perform the work. The method used for the analysis must identify the worker, the work organization and the work performed.

a) The worker component must identify the worker's capabilities, experience, education, and training. The worker component must relate to things, data and people, performance standards, specific training, and the nature of instructions required to carry out particular tasks.
b) The work component must identify what is actually being done (or planned to be done).
c) The work organization component must identify the organization's purpose, goals, resources, constraints, and objectives, and the way it structures itself to achieve them.

(507)

These three components of the work system are interdependent. Any change in one component without regard for the others will throw the system out of balance. Figure (2) shows the planned approach to solve this problem.

Work Systems Analysis

This analysis will review all Officer and Enlisted Billets dealing with marine safety functions up to and including District Commander(m). It will delineate differences that may result from differences in the types of geographic areas covered by the MSO -ie., Great Lakes, river, ocean.

The approach taken in performing this analysis is one that will meet the following criteria:

1) Provide a job analysis involving a careful study of jobs within the organization, to define specific job content. The analysis must provide an orderly, systematic collection of data about the job or position. Its purpose is to spell out, in as much detail as possible, what tasks constitute the job, how they are performed, and what behavior (skills, knowledge, and attitudes) the jobholder must have to perform certain specified tasks. The analysis will:
   a) utilize explicit, controlled, standardized language which can be recognized and understood across work fields and yet allows the reader to compare jobs (tasks).
   b) be stated to reveal precisely and concretely:
      1) what gets done (that is, the proceedures, methods, and processes with which the worker is engaged as he performs a task),
      2) what he does (that is, the physical, mental, and interpersonal involvement of the worker as he carries out proceedures and processes).
2) Provide an organizational analysis that examines the system wide components of the organization that affect the training program beyond those factors normally considered in task and person analysis. It will include an examination of the organizational goals; the resources of the organization; and the internal and external constraints present in the environment. The analysis must reflect the natural organization and structure of the agency as a work organization. It must describe the tasks and jobs as they are actually tailored to fit the local conditions.
3) Provide a systems approach that:
   a) Works from a statement of overall purpose, step by step, to determine what must be done to achieve that purpose.
   b) Works from the overall goal(s) of the organization to develop a set of specific objectives.
   c) Works from the objectives to determine what must be done to attain them, and identifies the major areas of work to be done.
   d) Works with these major work areas and identifies and defines specific tasks which need to be done to accomplish the objective of each subsystem.
   e) Defines each task precisely and concretely.
   f) Organizes the tasks in jobs, identifying similar functional levels, performance standards, and skill requirements.

(SOF)

An analysis of McCormick's Position Analysis Questionaire (PAQ), Christal's Job Inventory Approach (JIA), Fine's Functional Job Analysis and Systems Analysis Approach (FJA), the Air Force developed Comprehensive Occupational Data Analysis Program (CODAP), and the Department of Defense's Instructional Systems Development approach (ISD), has shown that of these approaches to task analysis, only Fine's FJA comes the closest to satisfying all of these criteria.

In fact the methodology chosen to analyize the work systems contains all aspects of Fine's FJA approach, yet is modified to provide a "complete systems" approcach for task analysis. This was done in order to provide information from the task analysis for three distinct purposes These are:

a) To provide information for Headquarters use in determining the manning levels of field marine safety units and to provide task data for determining billet descriptions.
b) Develop a qualifications jacket that would be capable of being maintained by each individual and would provide information to the Coast Guard detailer and to the Unit Commanding officer about the work, education, and training background and capabilities of the individual.
c) Develop information similar to what will be generated in this study, on additional training and education requirements for Coast Guard Reservists being called to active duty because of a Declaration of War or a declared National Emergency.

Training Review

The objective of this task is to review existing and planned, formal and on-the-job-training conducted or funded by the Coast Guard, review the training methods and techniques used in each training course identified, and review existing training methods and techniques.

The approach for this task will be to conduct a review of the following:

a) Existing and planned Coast Guard marine safety training,
b) Formal and "on-the-job" marine safety training,
c) Existing training methods and techniques used in training identified above.

This review will identify, for each of existing and planned training, as a minimum, the following items (referring to the skill and knowledge requirements defined by the task analysis).

a) The training requirement currently filled by the training.
b) The training requirement partially filled by the training.
c) The type of training and who conducts it.
d) The location of the training, and the average cost to the Coast Guard for the training.
e) The qualifications of the instructor.
f) Student prerequisites, if any.

Training Intergration/ Recommendations

The objective of this task is to integrate the results of TASKS I and II. This integration will provide the basis for recommendations that will provide managable, effective, and comprehensive training for marine safety personnel (Officer and Enlisted).

Three subtasks are anticipated in order to meet this objective. These subtasks will:

1) A review of the Work Sysytems Analysis and the Training Review are examined as a starting point in the development of recommendations for unit training, OJT, formal training, and training management.
2) Identify discrepancies between existing and optimal training. Where differences between the training currently existing (or planned) and the training requirement developed in the task analysis exist, the training requirements are developed. This definition will include evaluation critera, training objectives, and recommended presentation techniques.
3) Develop recommendations for unit training and marine safety training management.

CONCLUSION

This work is just getting started and it is hoped that the results of this work will provide the Headquarters' Marine Safety Training Council with the information necessary to determine managable, effective, and comprehensive training that will allow maximum utilization of training resources, maximize each individual's personal development, maximize the usefulness of each individual to the overall Coast Guard, and identify the training requirements for personnel working in individual Marine Safety Offices.

REFERENCES

Eschenbrenner, A.J., DeVries. P.B., and Ruck, H.W. "Methods for collecting and analyzing task analysis data", Proceedings, Military Testing Association, Oklahoma City, Oklahoma, 1978.

Fine, S. and Wiley, W., An Introduction to Functional Job Analysis: A Scaling of Selected Tasks from the Social Field. W.E. Upjohn Institute for Employment Research, Kalamazoo, Michigan, 1971.

Fine, S., Holt, A., and Hutchinson, M. Functional Job Analysis: An Annotated Bibliography. W.E. Upjohn Institute for Employment Research, Kalamazoo, Michigan, 1975.

Kershner, A.M., A Report on Job Analysis (ONR report ACR-5). Office of Naval Research, Department of the Navy, Washington, D.C., 1955.

McCormick, E.J., "Job and task analysis" in M.D. Dunnette (Ed.), Handbook of Industrial and Organizational Psychology. Rand McNally College Publishing Company, Chicago, Illinois, 1976.

Mead, D.F., Determining training priorities for job tasks, Proceedings, Military Testing Association, Indianapolis, Indiana, 1975.

Montemerlo, M.D. and Tennyson, M.E., Instructional Systems Development: Conceptional Analysis and Comprehensive Bibliography, (NAVTRAEQPCEN IH-257). Naval Training Equipment Center, Orlando, Florida, 1976.

(510)

Stoehr, L.A. and Paramore, B., Handbook for the Development of Qualifications for Personnel in New Technology Systems, (CG-D-75-76). U.S. Coast Guard, Office of Research and Development, 1976.

U.S. Coast Guard., Marine Safety Manual, (CG-495). U.S. Coast Guard, Washington, D.C., 1979.

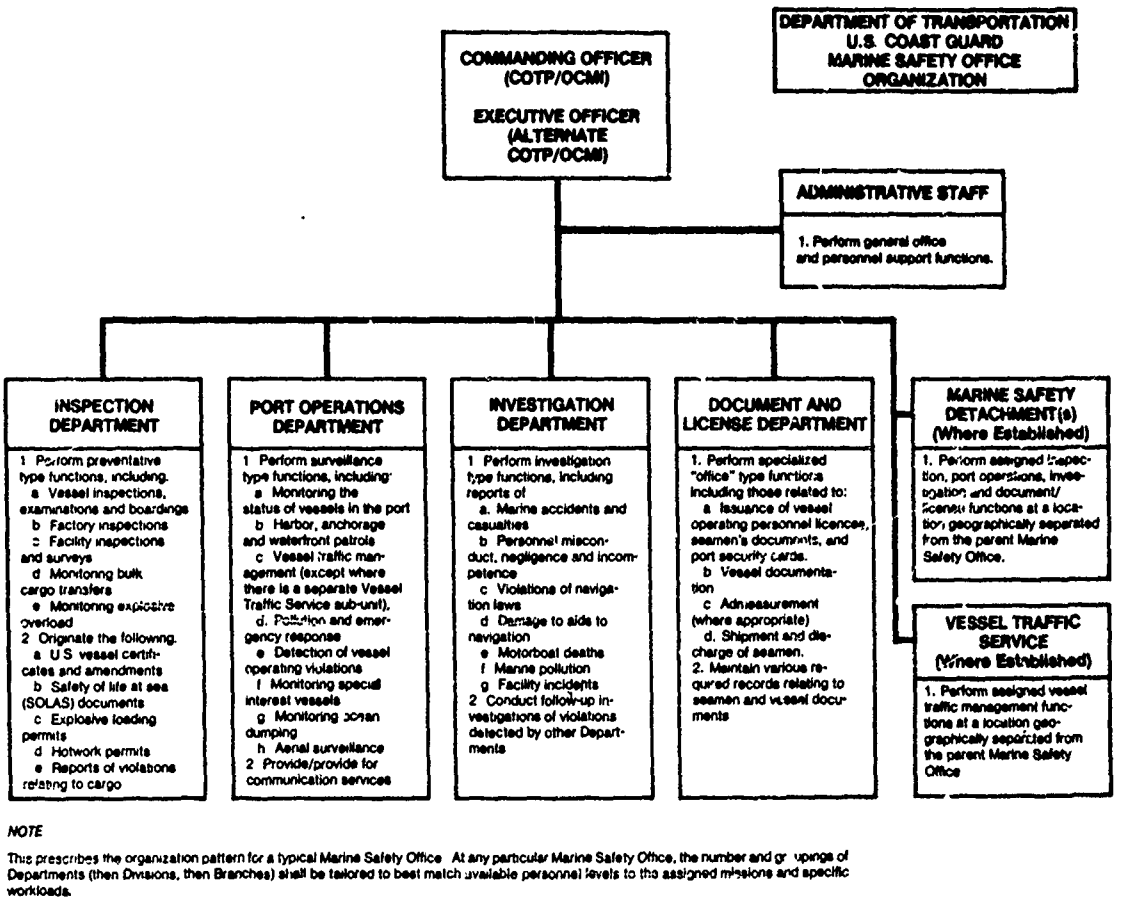


FIGURE 1. Typical Marine Safety Office Organization

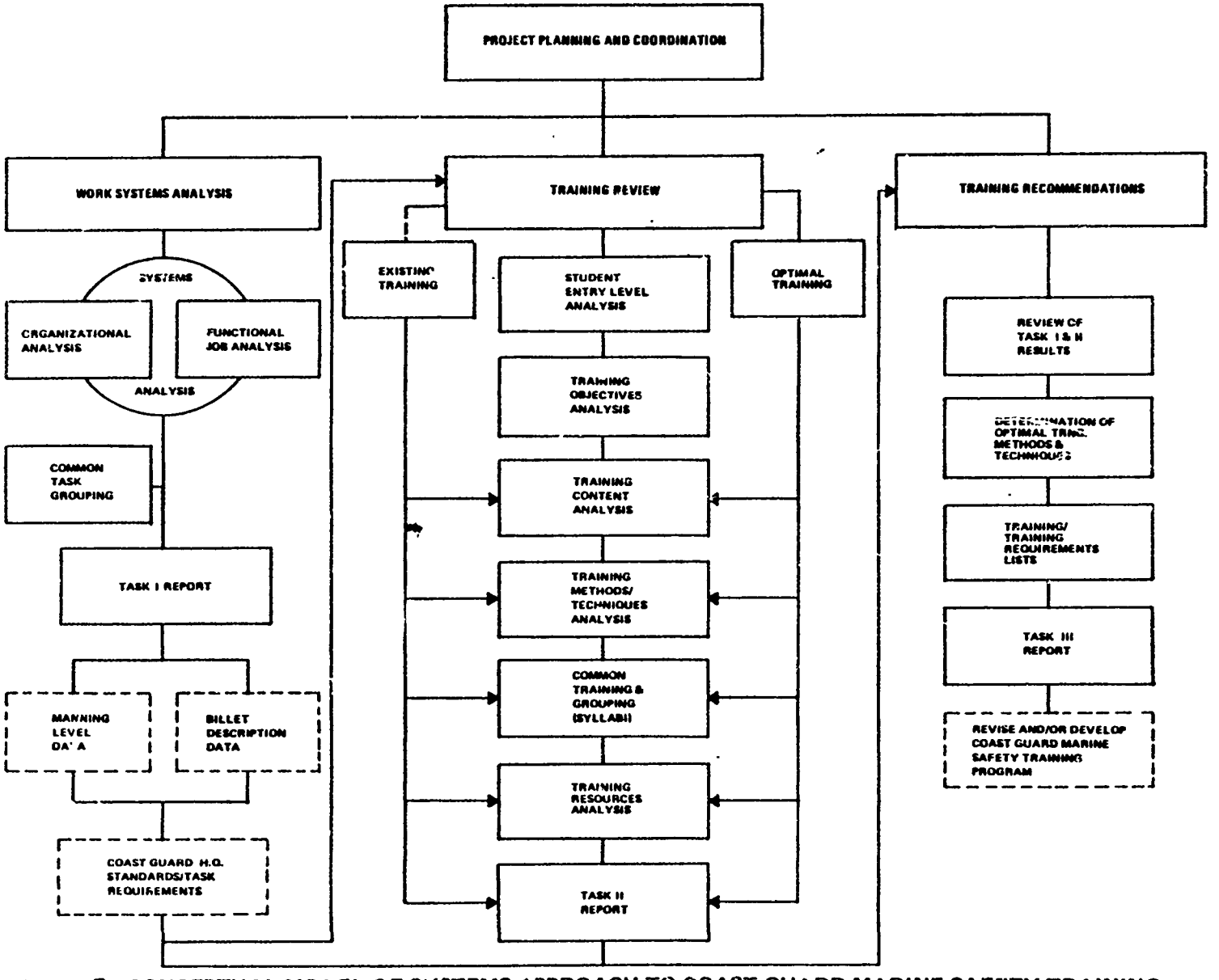


Figure 2. CONCEPTUAL MODEL OF SYSTEMS APPROACH TO COAST GUARD MARINE SAFETY TRAINING

USING ERROR RATES TO SELECT A CUT-SCORE

Karen N. Jones

U. S. Coast Guard Institute

Oklahoma City, Oklahoma

ABSTRACT

This paper presents an easy-to-use graphic method for selecting an optimal cut-score for a selection or classification test or for comparing tests. In this method, which is an adaptation of the receiver-operating-characteristic (ROC) curve, the probabilities of selection errors or incorrect decisions (i.e., the probability of a miss and the probability of a false alarm) for potential cut-scores are plotted. This method enables the user to select a cut-score to achieve a ratio of misses to false alarms as determined by the demands of a specific situation. Examples of the use of this method are presented.

INTRODUCTION

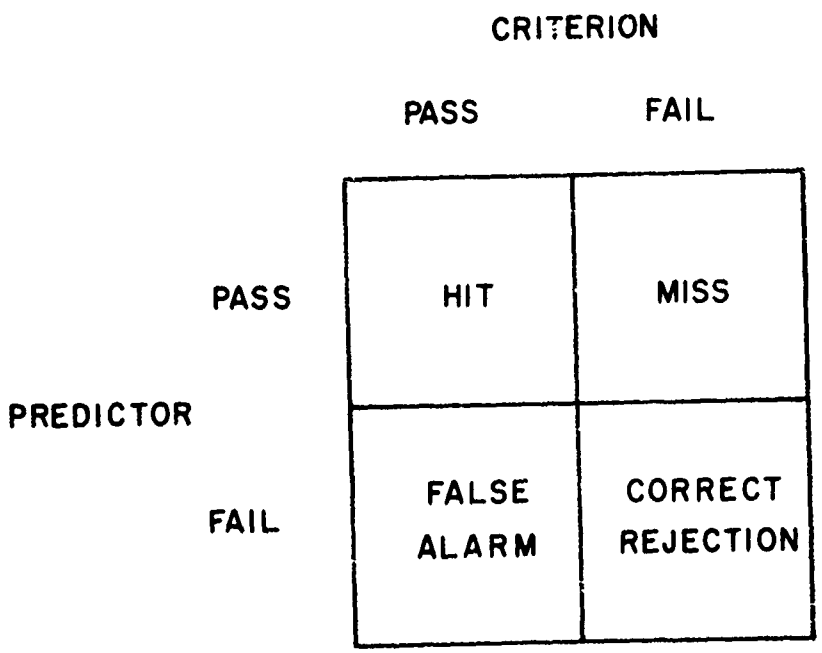
The purpose of this paper is to present a graphic method for selecting optimal cut-scores for predictor tests, i.e., for selection and classification tests. The graph developed using this method illustrates the proportion of errors in personnel decisions which would be made with each potential cut-score on the predictor test. This method uses a measure of the predictive validity of the predictor test and therefore requires pass or fail data on the performance being predicted (i.e., on the criterion). It is designed for use with a selection or classification test that will be used to predict future performance by dividing individuals into two categories: (1) those who pass the predictor test, are accepted, and are expected to succeed on the criterion and (2) those who fail the predictor test, are rejected, and are expected to fail on the criterion.

In the situation where performance on the predictor and performance on the criterion can be divided into pass and fail categories, there are several ways to select an optimal cut-score for the predictor test or to compare predictor tests. For example, the user can select a cut-score and/or a predictor test with the highest correlation or the lowest chi-square between the predictor and the criterion. The method presented in this paper has several advantages over these methods including being easily understood, easy to present, and usable with predictor tests scored by number of correct responses or by number of errors.

This method was developed to compare the validity of several clinical color

vision tests as predictors of performance on the Federal Aviation Administration's aviation signal light gun test (Jones, Steen, & Collins, 1975). It is an adaptation of the receiver-operating-characteristic (ROC) graph or curve which is used in signal detection theory to evaluate measurement sensitivity (Green & Swets, 1966). In the graph developed in this procedure, the user can see the effect which potential cut-scores on the predictor test would have on the personnel decisions made using the test. This enables the user to select a predictor test and/or a cut-score by counterbalancing the two types of errors involved in making personnel decisions--passing someone on the predictor who would fail the criterion (termed a miss) and failing someone on the predictor who would pass the criterion (termed a false alarm). The errors were chosen for graphing in this application rather than using an error and a correct decision as is graphed in the ROC graph, because a graph of the two types of errors appeared to be more sensitive to changes in cut scores.

The graphic definition of errors and correct decisions is presented in the contingency table in Figure 1. As shown in this figure, the data are divided into the following four categories based upon pass or fail on the predictor and pass or fail on the criterion: (1) pass predictor and pass criterion--hit; (2) fail predictor and pass criterion--false alarm; (3) pass predictor and fail criterion--miss; and (4) fail predictor and fail criterion--correct rejection. To make the terms easier to understand in the context of personnel testing, for this application the definition of false alarm and miss have been reversed from the way they are defined in signal detection theory.



p (Hit) = p (Pass predictor | Pass criterion)
p (Miss) = p (Pass predictor | Fail criterion)
p (False Alarm) = p (Fail predictor | Pass criterion)
p (Correct Rejection) = p (Fail predictor | Fail criterion)

Figure 1 - Predictor and criterion contingency table and conditional probability equations for personnel decisions. (Adapted from Jones, et.al, 1975)

514

In Figure 1, the conditional probability equation for each category of personnel decision, e.g., hit, is also presented. In these conditional probability equations, the probability of an occurrence on the predictor is conditional upon the occurrence of the criterion. For example, the equation for a hit is:

$$p(\text{Hit}) = p(\text{Pass predictor} \mid \text{Pass criterion})$$

This equation is read as: "The probability of a hit equals the probability of passing the predictor given that the criterion will be passed." In other words, the probability of a hit for a particular cut-score is equal to the number of examinees passing the predictor and passing the criterion, divided by the number of examinees passing the criterion, i.e.,

$$p(\text{Hit}) = \frac{\text{number of examinees passing the criterion and passing the predictor}}{\text{number of examinees passing the criterion}}$$

The conditional probability equations for the other three categories of personnel decisions are read and calculated in a similar manner.

DESCRIPTION OF METHOD

In the graphs developed using this method (e.g., Figure 2), the proportion of errors and the proportion of false alarms which would have occurred at each cut-score on the predictor test are illustrated. The steps involved in this method are presented below and an example of the product of steps 1, 2, 3, and 4 is presented in Table 1.

Step One: Prepare two frequuency distributions for the predictor scores--one for the examinees who pass the criterion and one for the examinees who fail the criterion.

Step Two: Calculate the number of false alarms and the number of misses which would occur if the cut-score was set at each predictor score.

For a test scored by number of correct responses, the number of false alarms for a specified score is the number of examinees who made scores lower than the specified cut-score and passed the criterion; and the number of misses is the number of examinees who made scores equal to or greater than the cut-score and failed the criterion.

For a test scored by number of errors, the number of false alarms for a specified score is the number of examinees who made more errors than the cut-score and passed the criterion; and the number of misses is the number of examinees who made the same or fewer errors than the cut-score and failed the criterion.

Step Three: Transform the number of false alarms and the number of misses in Step Two into the proportion of false alarms and the proportion of misses.

Step Four: (optional) Transform the proportions in Step Three into z scores.

Step Five: Plot the proportion of false alarms and proportion of misses for each predictor score on double probability paper OR plot the z scores for the

Table 1

Sample problem of steps 1 through 4 for method presented in this paper.

Step 1: Frequency distributions

Score on Predictor	Number who Pass Criterion	Number who Fail Criterion
1	1	20
2	3	11
3	5	7
4	7	2
5	10	0
	12	1

Step 2: Number of errors

Score on Predictor	Number of False Alarms	Number of Misses
1	0	41
2	1	21
3	4	10
4	9	3
5	16	1
6	26	1

JON-2-4

Step 3: Proportion of errors

Score on Predictor	Proportion of False Alarms	Proportion of Misses
1	0.00	1.00
2	0.03	0.51
3	0.10	0.24
4	0.24	0.07
5	0.42	0.02
6	0.68	0.02

Step 4: z scores

Score on Predictor	Number of False Alarms	z score for misses
1	-3.70	+3.70
2	-1.87	+0.02
3	-1.28	-0.70
4	-0.70	-1.47
5	-0.20	-2.06
6	+0.47	-2.06

Note: Data are from test scored by number of correct responses.

proportion of false alarms and misses for each predictor score on regular graph paper.

Step Six: Select an optimal predictor test or cut-score.

The sixth step is often the most difficult since selection of an optimal test or cut-score may incorporate several factors, e.g., cost of training and criticality of success on the criterion, and a decrease in one error rate is accompanied by an increase in the other error rate. For example, as the cut-score is made more stringent, the miss rate decreases and the false alarm rate increases. Therefore, the user must counter-balance the two error rates for a particular situation. In a situation where the cost of failure was high, e.g., people can be killed or injured if the person selected fails at the task during training, the user would want a predictor test with a low miss rate and would be willing to tolerate a higher false alarm rate to achieve a low miss rate. In another situation, where success in training or on-the-job is not so critical or the cost of failure is not excessive, the user might want a lower false alarm rate and would be willing to sacrifice a low miss rate to achieve the lower false alarm rate. There are no rules of thumb for selecting a test or a cut-score based upon the graphs developed in this method. Final interpretation of the data and selection of a test and/or cut-score must be left to the user's discretion.

EXAMPLES OF APPLICATION OF METHOD

Figures 2, 3, and 4 are examples of the application of this method. Figure 2 illustrates the predictive validity of two pseudoisochromatic plate tests--the American Optical Company test (1965 edition) and the Dvorine test--as predictors of performance on the Federal Aviation Administration's signal light gun test. As shown in Figure 2, the predictive validity of the two tests, as represented by the

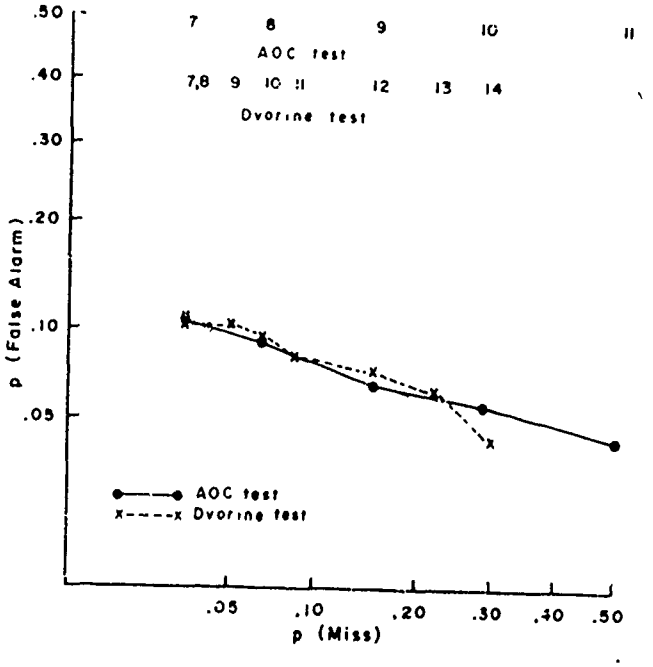


Figure 2-Miss and false alarm rates for the American Optical Company and Dvorine tests as predictors of performance on the aviation signal light gun test. (Adapted from Jones, et.al, 1975)

miss rate associated with a particular false alarm rate or vice versa, was quite similar. In comparing the two tests or selecting an optimal cut-score, evaluation can be made based upon the miss and false alarm rates without reference to the cut-score associated with each miss and false alarm rate. For example, for a miss rate of approximately 0.17, the false alarm rate on both tests was approximately 0.07. This similarity can be seen throughout the graph. An examination of the miss and false alarm rates in relation to the cut-scores, which are printed at the top of the graph, shows that the same cut-score on the two tests often had a different predictive validity or, conversely, the same miss and false alarm rates had a different cut-score. For example, for a miss rate of approximately 0.07 and a false alarm rate of approximately 0.09, a cut-score of eight errors would be selected for the American Optical Company test and a cut-score of ten errors would be selected for the Dvorine test.

Since the false alarm rates for all of the cut scores on the plate tests in Figure 2 were low, the cut score for each test could be selected based upon the miss rate. Unfortunately, this situation does not always exist. An example of a test with less than optimal predictive validity is presented in Figure 3. In this figure, the validity of the Farnsworth 100-Hue test as a predictor of performance on the aviation signal light gun test is represented. As in Figure 2, in Figure 3 the miss and false alarm rates for each potential cut-score are plotted with the cut-score for each point on the graph printed at the top of the graph. As can be seen in this

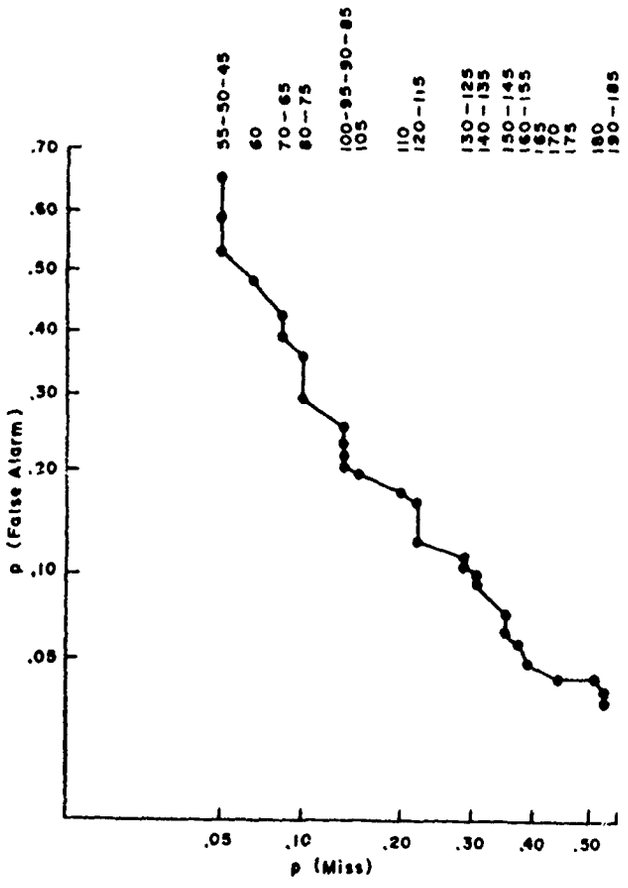


Figure 3-Miss and false alarm rates for the Farnsworth 100-Hue test as a predictor of performance on the aviation signal light gun test. (From Jones,et.al 1975)

(518)

figure, there were no potential cut-scores for which the miss rate and the false alarm rate were as low as was achieved with the plate tests in Figure 2. For example, for a miss rate of 0.05 the false alarm rate was 0.50 or higher as compared to a false alarm rate of 0.10 on the tests in Figure 2. The most optimal cut-score for the test in Figure 3 appeared to be 100-115 where the false alarm rates were approximately 0.12 - 0.20 and the miss rates were approximately 0.15 - 0.22. Although these rates are not as high as ones which might be encountered, since the Federal Aviation Administration had approved other tests which had higher predictive validities, this test was not recommended for use as a predictor of performance on the signal light gun test by Jones, et al. (1975).

In Figure 4, the data for another application of this method are presented. In this figure, the validity of the Armed Forces Vocational Aptitude Battery (ASVAB) as a predictor of success on the U. S. Navy's Basic Test Battery (BTB) for a sample of 271 Coast Guard recruits is presented. Since a composite standard score of 130 is used as the cut-score for enlistment on two of the Coast Guard's selection tests (the Coast Guard Selection Test and the Short Basic Test Battery) and this score equates to a standard score composite of 130 for GCT + ARI + MECH on the BTB, success on the BTB for this comparison was defined as a composite of 130 or higher. The cut-score of 25 on ASVAB which was in effect when this sample of examinees enlisted in the U.S. Coast Guard[1] is not represented in Figure 4 since it was "off-the-scale" at p(Miss) = 1.0 and p(False Alarm) = 0.0. In fact, as can be seen in Figure 4, the miss rate for the potential cut-scores on ASVAB did not drop below 0.50 until a cut-score of 34 was used. This indicated that the cut-score being used with ASVAB was too low. Although recommendations for the selection of a new cut-score for Coast Guard use have not been made based upon these data, it would be feasible to make such a recommendation and the recommendation made should depend on the desired impact of the cut-score. For example, if the desired result was a very low miss rate and a higher false alarm rate could be tolerated, a

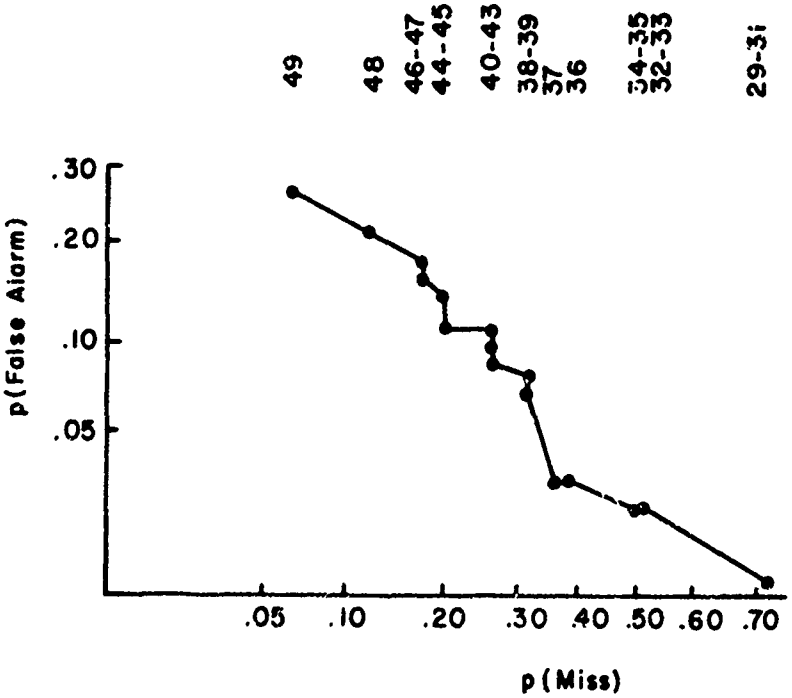


Figure 4 - Miss and false alarm rates for the Armed Services Vocational Aptitude Battery as a predictor of performance on the Basic Test Battery.

(519)

cut-score of 48 or 49 would be recommended. Or if the Coast Guard wanted approximately equal false alarm and miss rates in the range of 0.10 to 0.20 then a cut-score of 44 to 46 would be recommended. As this example indicates, selection of an optimal cut-score depends on the desired impact of the test which in turn could involve a number of factors.

CONCLUSIONS

This method has several advantages for the user. For example, the calculations can be performed manually, the same graph can be used for comparing tests or for selecting cut-scores, and the graph is easy to present and to understand. Also, since the false alarm rate increases as the miss rate decreases and vice versa, it is possible to make generalizations to cut-scores higher or lower than the ones represented on the graph. However, the user should keep in mind that an investigation of the robustness of the method to factors such as sample size and truncation have not been made. Therefore, generalizations beyond the data represented or generalizations with a small or non-representative sample should be made with caution. However, it may be feasible to develop a correction factor for this type of situation.

REFERENCES

Green, D. M. & SWETS, J. A. Signal Detection Theory and Psychophysics. New York: John Wiley and Sons, Inc., 1966

Jones, K. N., Steen, J. A., & Collins, W. E. Predictive validities of several clinical color vision tests for aviation signal light gun performance. Aviation, Space, and Environmental Medicine, 1975, 46, 660-667

FOOTNOTES

1. On March 31, 1980, the Coast Guard suspended use of ASVAB as a selection battery for enlistment.

(270)