LEVEL II

AD_____

REPORT NO. T 8/80

(12)

SCREENING FOR PHYSICAL CAPACITY IN THE US ARMY: AN ANALYSIS OF MEASURES PREDICTIVE OF STRENGTH AND STAMINA

AD A097457

US ARMY RESEARCH INSTITUTE

OF

ENVIRONMENTAL MEDICINE

Natick, Massachusetts

10 JUNE 1980

DTIC
ELECTE
APR 0 8 1981
S D
F

DTIC FILE COPY

Approved for public release; distribution unlimited.

UNITED STATES ARMY

MEDICAL RESEARCH & DEVELOPMENT COMMAND

81 4 8 0

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

DISPOSITION INSTRUCTIONS

Destroy this report when no longer needed.
Do not return to the originator.

Approved for public release;
distribution unlimited.

AD _____

(14) USARIEM
~~Report No.~~ T-8/80

(6)

Screening for Physical Capacity in the US Army: An Analysis of Measures Predictive of Strength and Stamina

(15)

Dan S. Sharp, [redacted], James E. Wright, [redacted], James A. Vogel, Ph.D.,
John F. Patton, [redacted], William L. Daniels, Ph.D., Joseph Knapik, M.S.,
and Dennis M. Kowal, Pn.D.

(11) 14 J...

(12) 12...

(16)

Project References: 3E162777A845

Study References: PH-1-78 and Ph-2-79

U. S. Army Research Institute of Environmental Medicine

Natick, MA 01760

FOREWORD

This report presents the detailed research findings in response to a 25 July 1977 tasking from the ODCSPER to OTSG "to develop, for pilot testing, a battery of physical fitness tests suitable for screening new accessions for MOS classification during the AFEES medical examination." In response to this tasking, the Exercise Physiology Division of this Institute carried out two separate research studies. The first, entitled "Evaluation of a physical fitness test battery for Armed Forces Examining and Entrance Stations" was carried out in January through May 1978 at the Training Center, Ft. Jackson, SC. Based on the preliminary findings from the Ft. Jackson study, a follow-up study with revised objectives entitled "Development of MOS fitness standards and an AFEES classification system for MOS assignment qualification" was carried out in September and October 1979 with soldiers of the 24th Infantry Division, at Ft. Stewart, GA. The principle findings from these two studies relative to the development of a physical fitness screening system for the AFEES are presented herein. The report is purposefully detailed and elaborate in order to document the methodology and rationale. It is recommended that the sections titled ABSTRACT, INTRODUCTION, and SUMMARY AND CONCLUSIONS be read first to provide an overview of the project.

Accession For
NTIS GRA&I ☒
DTIC TAB ☐
Unannounced ☐
Justification

By
Distribution/
Availability Codes
Dist | Avail and/or Special
A

TABLE OF CONTENTS

	Page
Foreword	iii
List of Tables	v
List of Figures	viii
Abstract	x
Introduction	1
Background	5
Categorization of Tasks	5
Criteria of Job Performance	7
Stamina	7
Strength	10
Swedish Physical Fitness Screening System	11
Proposed USA Physical Fitness Screening System	14
Aerobic Capacity	14
Step Test	15
Anthropometry	17
Strength Capacity	19
The Role of Gender	22
Guidelines for Setting Standards	24
Strength	24
Stamina	26
Study Design and Methods	28
Fort Jackson	28
Fort Stewart	31
Statistical Design and Methods	37
Accounting for Gender Effects	38
Multicollinearity	43
Cross Validation	45
Utilization of the Model	46
Results and Discussion	49
Sample Characteristics	49
Model of $\dot{V}O_2$ max	53
Training	53
Basis of $\dot{V}O_2$ max	55
Three Predictor Model	56
Two Predictor Model	69
Model of MLC	81
Analysis of Covariance	81
Determination of Predictive Model	84
Final Models for $\dot{V}O_2$ max and MLC	102
Repetitive Lift and Carry Performance	103
Conclusions and Summary	104
References	110

LIST OF TABLES

Table	Legend	Page
1	"Ready for Labor and Defense" (GTO), 1972, Stage 4.	4
2	MOS Clustering Criteria.	6
3	MOS Clusters.	7
4	Relationship between point scale and measure of performance for Swedish Physical Fitness Screening System.	13
5	Linear Regression Coefficients for the estimation of body density from the logarithm of the sum of four skinfolds.	19
6	Sample characteristics of two groups for each sex for Fort Jackson.	50
7	Sample characteristics of two groups for each sex for Fort Stewart.	50
8	Test for homogeneity of distribtuion characteristics between groups using t test.	52
9	Average difference in Fort Jackson post-training and pre-training $\dot{V}O_2$ max for each group and sex, and t test of significance for zero difference.	53
10	Test of Fort Jackson data for parallel and coincidental behavior using t tests, and homogeneity of variance using F test. Comparisons are between sexes in the same gender.	57
11	Test of Fort Jackson data for parallel and coincidental behavior using t tests, and homogeneity of variance using F test. Comparisons are between sexes in the same group.	58
12	Intercorrelation matrix for criterion and predictor measures for each group in the Fort Jackson data.	59
13	Eigenvalues and unbiased standardized regression coefficients for the prediction of $\dot{V}O_2$ max from SEX, $\dot{V}O_2$ AR, and % BF.	63

LIST OF TABLES (Con't.)

Table	Legend	Page
14	Standardized regression coefficients and squared multiple correlation coefficients for two models of $\dot{V}O_2$ max (ml/kg/min).	67
15	Eigenvalues and unbiased standardized regression coefficients for a single combined groups model of $\dot{V}O_2$ max.	69
16	Classification of subjects for $\dot{V}O_2$ max for a cluster standard of 42 ml/kg/min and 95% probability for two and three predictor models.	74
17	Standardized regression coefficients and squared multiple correlation coefficients for two models and a combined groups model of $\dot{V}O_2$ max for two predictors.	81
18	Test of Fort Stewart data for parallel and coincidental behavior using t tests, and homogeneity of variance using F test. Comparisons are between groups for the same gender.	82
19	Test of Fort Stewart data for parallel and coincidental behavior using t test, and homogeneity of variance using F test. Comparisons are between sexes in the same group.	82
20	Intercorrelation matrix for criterion and predictor variables for each group in the Fort Stewart data.	86
21	Eigenvalues and unbiased standardized regression coefficients for the prediction of ML132 from LBM, UT, LEG, TR, HG, UP132, UP38, and SEX.	90
22	Eigenvalues and unbiased standardized regression coefficients for the prediction of ML132 from LBM, UT, HG, UP38, and SEX.	93
23	Eigenvalues and unbiased standardized regression coefficients for the prediction of ML132 from LBM, UP38, and SEX.	97
24	Standardized regression coefficients and squared multiple correlation coefficients for two models of ML132.	100

LIST OF TABLES (Con't.)

Table	Legend	Page
25	Eigenvalues and standardized regression coefficients for a single combined groups model of ML132.	102
26	Raw score scaled coefficients, standard error of the estimate (SEE), and sample size for combined groups data for the prediction of ML132 in kg and relative $\dot{V}O_2$ max in ml/kg/min. (males = 1, females = 2 for SEX).	103
27	Regression analysis for the prediction of lift and carry performance at two loads for each gender separately from ML132 and $\dot{V}O_2$ AR predictors.	104

LIST OF FIGURES

Figure		Legend	Page
1	-	Static Strength Device.	30
2	-	Handgrip Device.	33
3	-	132 cm Upright Pull Device.	34
4	-	38 cm Upright Pull Device.	34
5	-	Idealized Parallel and Coincidence Effects in Two Groups: a) No parallel or coincidence effects. b) Parallel but non-coincidence effects. c) Parallel and coincidence effects.	40
6	-	Effect of non-homogeneity of variance due to group differences.	42
7	-	Group 1 model of relative $\dot{V}O_2$ max. Variation of three standardized regression coefficients with bias.	61
8	-	Group 2 model of relative $\dot{V}O_2$ max. Variation of three standardized regression coefficients with bias.	62
9	-	Group 2 predictor data in three predictor Group 1 model for relative $\dot{V}O_2$ max. Variation of prediction standard deviation with bias.	64
10	-	Group 1 predictor data in three predictor Group 2 model for relative $\dot{V}O_2$ max. Variation of prediction standard deviation with bias.	65
11	-	Combined groups three predictor model for relative $\dot{V}O_2$ max. Variation of standardized regression coefficients with bias.	68
12	-	Measured versus predicted relative $\dot{V}O_2$ max for the three predictor model.	71
13	-	Measured versus predicted relative $\dot{V}O_2$ max for the two predictor model.	72
14	-	Group 1 model of relative $\dot{V}O_2$ max. Variation of two standardized regression coefficients with bias.	76
15	-	Group 2 model of relative $\dot{V}O_2$ max. Variation of two standardized regression coefficients with bias.	77
16	-	Group 2 predictor data in two predictor Group 1 model for relative $\dot{V}O_2$ max. Variation of prediction standard deviation with bias.	78

LIST OF FIGURES (Con't.)

Figure		Legend	Page
17	-	Group 1 predictor data in two predictor Group 2 model for relative MOC. Variation of prediction standard deviation with bias.	79
18	-	Combined groups two predictor model for relative MOC. Variation of standardized regression coefficients with bias.	80
19	-	Group 1 model of MLC. Variation of eight standardized regression coefficients with bias.	88
20	-	Group 2 model of MLC. Variation of eight standardized regression coefficients with bias.	89
21	-	Group 1 model of MLC. Variation of five standardized regression coefficients with bias.	91
22	-	Group 2 model of MLC. Variation of five standardized regression coefficients with bias.	92
23	-	Group 1 model of MLC. Variation of three standardized regression coefficients with bias.	95
24	-	Group 2 model of MLC. Variation of three standardized regression coefficients with bias.	96
25	-	Group 2 predictor data in three predictor Group 1 model of MLC. Variation of prediction standard deviation with bias.	98
26	-	Group 1 predictor data in three predictor Group 2 model of MLC. Variation of prediction standard deviation with bias.	99
27	-	Combined groups three predictor model for MLC. Variation of standardized regression coefficients with bias.	101

ABSTRACT

Two models to predict aerobic and strength capacities have been developed. Prediction of these capacities has been predicated on demonstrated relationships between them and simple measures of anthropometry and performance.

The relative maximal oxygen consumption ($\dot{V}O_2$ max) was chosen as the criterion measure for aerobic capacity. This choice reflects well understood physiological principles relating $\dot{V}O_2$ max and the aerobic requirements of real world tasks. The safe maximal lifting capacity to a 132 cm platform was chosen as the strength capacity criterion. This choice reflects a simplification of strength-demanding performance requirements in the U.S Army. The simplification is justified by the demonstration that in excess of 90% of Army tasks having non-trivial strength requirements have lifting and/or repetitive lift and carrying solely as the strength demanding task.

The use of the criterion measures to set physical capacity standards and describe enlistee population characteristics is constrained by a number of weaknesses in the sample populations used to construct the models. Fortunately, however, these limitations need not detract from practical utilization of the system. The criterion measures represent simulators of real world performance requirements, and thereby need not be considered as the ultimate criteria by which to set the screening standards. Rather, manpower needs, injury rates, etc., can be used in a dynamic mode to vary standards periodically, and thereby assure that the best personnel are placed in the more physically demanding jobs.

INTRODUCTION

In May, 1976, the General Accounting Office (GAO) issued recommendations to the military services to develop physical and operational fitness standards for job specialties. These standards should reflect the operational performance requirements in strength and stamina for job specialties requiring these factors for effective performance. They should be job specific, and there should be no differentiation in standards between males and females.

The U.S. Army decided to pursue these recommendations along two basic lines. One line would deal with the development of training programs and testing standards that reflected the physical fitness requirements of specific job specialties. The second line would deal with the development of fitness screening procedures to be administered to new accessions at the time of enlistment. This line would test and screen enlistees as to their suitability to meet the fitness requirements of the job speciality for which they were being recruited. Inherent in both of these lines is the determination of the actual physical demands for the job specialties. This report deals with the second line - development of enlistee testing and screening procedures.

Testing and screening for physical fitness at the time of induction is not a new concept. In 1969 Sweden instituted a comprehensive screening process which included fitness testing[1]. This system is based on a model initially suggested by the work of Tornvall[2] and later validated by Nordesjo and Schele.[3] The Soviet system of fitness testing and screening employs an entirely different approach. It is based on a formalized system of training and performance evaluation in a program called the GTO.[4] This acronym translated means "Ready for Labor and Defense". The current version of this program was introduced in 1972 as a formal means to train for and measure physical fitness

skills. At the age of 10 years the Soviet citizen is introduced to the system through the school system. Initially, the child is expected to perform in seven events ranging from sprinting to swimming. The Soviet citizen progresses through five stages as he/she ages. Records of performance are maintained throughout an individual's adolescence, and at the time of induction are used as a means of assessing fitness and suitability for military occupations.

One advantage the Soviet system offers over the Swedish is the use of performance on tasks and events that have high face validity. World War II was a test of the principles embodied in the GTO. Events such as cross country running, skiing, shooting, grenade throwing and combat sports suddenly became quite relevant to the newly inducted Soviet soldier at the battlefront.[5] Table 1 details the ten events required of citizens from ages 19 to 28.[5] Ostensibly this program represents an effort by the Soviet Government to enhance physical fitness and physical preparedness for Soviet society as a whole. A major benefit of such a program is that it provides a convenient mechanism by which to match individual performance capacity to occupational physical performance requirements at the time of induction into the military system. The military inductee presents to the screening process with a longitudinal history of performance capability. The value of this type of information in better matching the individual to military occupation cannot be underestimated.

The effectiveness of such a system of screening is enhanced in Soviet society where rigid social mechanisms already exist to administer and maintain such a complex program. Western societies, however, lack such mechanisms. Accordingly, the Swedish model based on cross sectional testing of physiological capacities at the time of induction that in turn correlate significantly with criterion tasks having high face validity suggests itself to be the most fruitful approach for the U. S. Army to pursue in meeting its fitness screening goals.

The purpose of this report is to present the methodology by which to implement a screening process for physical capacity at the time of enlistment. Factors addressed in this presentation include determination of physical job requirements, development of a scheme to quantify physical capacity, deriving a model to predict physical capacity and the methodology by which to administer the screening process and utilize the screening information.

The latter factor particularly involves a number of issues possessing administrative and utilization dilemmas. These include the setting of standards for both job requirements and screening procedures, guarding against gender bias, and balancing manpower needs with adherence to the screening system.

The effective utilization of a device to better match an individual and his/her capabilities to the physical demands of their job cannot always be measured directly, or demonstrate acceptable short term results. The benefits of such a system are long term and reflect themselves in greater productivity and efficiency, decreased injury rates, etc. Often only the short term costs and risks of implementing a system with benefits difficult to identify and/or quantify seem to inhibit implementation of such programs or even prohibit discussion of the principles behind the issues. Such a course could cost us where our military personnel may be required to confront an adversary who has taken into account individual suitability to physical task demands. Similarly, in this day of limited resources and fiscal restraint, methods to enhance efficiency and productivity may be the only recourse in effectively maintaining a reliable and capable military establishment.

Table 1

"Ready for Labor and Defense" (GTO), 1972, Stage 4[a]

Academic requirements

1. To have knowledge of "Physical Culture and Sport in the USSR".
2. To know and practice the rules for personal and public hygiene.
3. To know the basic rules of civil defense and wear a gas mask for one hour.
4. To be able to explain the importance of and to perform a set of morning exercises.

Physical Exercises: qualifying standards

		MALE		FEMALE	
	Event	Silver	Gold	Silver	Gold
1.	Run 100m (sec)	14.0	13.0	16.0	15.2
2.	Run 500m (min:sec)	-	-	2:00	1:45
	or 1000m (min:sec)	3:20	3:10	4:30	4:10
	or 3000m (min:sec)	11:00	10:30	-	-
3.	High jump (cm)	130	145	110	120
	or long jump (cm)	460	500	350	380
4.	Hurl hand grenade of 500 gm (m)	-	-	23	27
	of 700 gm (m)	40	47	-	-
	or putt shot of 4 kg (m)	-	-	6.5	7.5
	of 7.257 kg (m)	7.5	9.0	-	-
5.	Ski 3 km (min)	-	-	19	17
	or 5 km (min)	25	24	35	33
	or 10 km (min)	54	50	-	-
	In snow free regions:				
	Run cross country 3 km (min)	-	-	19	17
	6 km (min)	36	33	-	-
	or cycle cross country 10 km (min)	-	-	28	25
	20 km (min)	46	43	-	-
6.	Swim 100 m (min:sec)	2:05	1:05	2:20	2:00
7.	Pull ups:				
	one's own weight up to 70 kg	9	13	-	-
	one's own weight over 70 kg	7	11	-	-
	or lift weights above one's head (as a percentage of own weight)				
	own weight up to 70 kg	55	75	-	-
	own weight over 70 kg	65	85	-	-
	or push ups	-	-	12	14
	or sit ups	-	-	40	50

8.	Fire a small bore rifle 25 m (pts)	37	43	37	43
	or at 50 m (pts)	34	40	-	-
	Fire a heavy weapon at 100 m (pts)	70	75	-	-
9.	Orienteering with test of knowledge	25 km	30 km	25 km	30 km
10.	Obtain a sports ranking (level)	-	II	-	II

Note: for the Gold Badge one must attain not less than 7 qualifying standards at Gold Badge level plus two at Silver Badge level.

[a] From Ref 5.

BACKGROUND

Categorization of Tasks

Before any screening or testing procedure can be developed it is imperative that the actual physical work demands of the job specialty be determined. Currently, the U.S. Army has in excess of 350 military occupational specialties (MOS). The U.S. Army Infantry School at Fort Benning, Georgia was tasked to generate a MOS Physical Task List. This list is a compilation of physical tasks performed by personnel within each MOS. Information was provided by service schools, and represented a brief operational description of specific task demands. These descriptions were derived by instructors and military personnel with combat experience, and represent experienced opinion rather than observed practice. For example, for the MOS designated 13B (artillery) one of the task descriptions is, "With projectiles weighing from 16 to 90 kg and a 5-ton cargo truck, lift and carry a maximum of 45 kg 20 meters 100 times per day." Upon completion of the task list, MOS's with similar physical demands were clustered together based on two components of physical capacity; i.e., muscular strength and stamina. This grouping was accomplished solely on the basis of inspection of the task description. Table 2 illustrates the classification criteria utilized in the sorting of the MOS's into clusters.

Table 2

MOS Clustering Criteria

Category	Strength (kg of weight lifted)	Stamina (Calories/minute)
Low	< 30	< 7.5
Medium	30-40	7.5-11.25
High	> 40	> 11.25

Weight lifting requirements for the three categories were selected primarily on the basis of standards established by the Training and Doctrine Command (TRADOC) for MOS's already determined by them to be low demand, and by natural breaks in the weights of objects lifted in the more demanding MOS's. This classification was predicated upon the demands of a single lift or lift and carry task. Extended durations of activity (repetitive lifting), unusual postural or other factors increasing or decreasing task demands can alter the classification scheme significantly. Delineation of two components of physical capacity represent an attempt to simplify physical job requirements. Stamina or aerobic capacity classification criteria were derived from estimated energy costs of the most demanding repetitive lifting, pushing, pulling, supporting and/or carrying tasks within the MOS's. The few data available in the literature on energy cost classification scales for industrial populations[6-8] were of limited value in establishing these criteria due to major differences in the demands of military versus civilian jobs, and in the physical characteristics and training backgrounds of the work force itself. Even the low, or baseline, requirements of the Army would be classified as heavy to very heavy exercise according to several accepted classification schemes[6-8].

Table 3 indicates the relative strength and stamina demands of five finalized clusters. The total number of MOS's and the percentage of enlisted personnel within each cluster are also given. Certain combinations of strength and stamina requirements were not evinced, thereby leaving a total of only five clusters.

Table 3

MOS Clusters

Cluster	Fitness Requirement Strength	Stamina	Total MOS's	% Enlisted Personnel
Alpha	high	high	10	19
Bravo	high	medium	39	13
Charlie	high	low	63	21
Delta	medium	low	53	21
Echo	low	low	184	26

Criteria of Job Performance

At this point the first major problem is presented - that of establishing valid criteria of job performance. The problem has at least been initially addressed by the formation of two separate components of physical performance -strength versus stamina. It is well established that an individual's ability to maintain a repetitive task such as running may be unrelated to that individual's ability to do impulse work such as a single lift of 100 kg. Separation of these functional abilities is also supported by the relatively distinct physiological and biochemical mechanisms associated with each type of work performance.

Stamina

Stamina performance requirements can be objectively determined by actually measuring the calories expended (or oxygen consumed) in performing the

task. In order to qualify a task as being predominantly a stamina task it must meet certain conditions. The primary requirement is that it must be a repetitive task capable of being sustained for at least ten to fifteen minutes. Secondarily, it must not require relatively large amounts of sustained "strength".

Because the actual cost of aerobic tasks can be measured, it is relatively simple to derive standards by which to judge an individual's capacity to perform the task. For example, a task such as unloading 15 kg cartons from a truck at a frequency of one carton every 20 seconds may call for an average oxygen consumption of 30 ml O_2/kg/min. If this rate of performance were to be sustained for a relatively long period of time (e.g., 2 hours) it would not be unreasonable to expect an individual to be performing this task at no more than 60% of his/her maximal oxygen consumption ($\dot{V}O_2$ max). Therefore, individuals with $\dot{V}O_2$ max's of at least 50 ml/kg/min would be judged capable of performing this task well under their capacity. Tasks that are shorter in duration but call for the same rate of energy expenditure may be performed at a high percentage of $\dot{V}O_2$ max. In this example, if the length of the task was for thirty minutes, then it would be reasonable to expect individuals to work as high as 75% of their $\dot{V}O_2$ max, and an individual with a $\dot{V}O_2$ max as low as 40 ml/kg/min would be acceptable.

Inherent in this approach of describing the criterion of stamina performance in terms of a ratio of actual task cost to $\dot{V}O_2$ max is a simplification. It involves the mode of activity by which the $\dot{V}O_2$ max is determined. The value of the $\dot{V}O_2$ max depends on the activity by which it is measured. For example, Hermansen and Saltin[9] showed that in the same subjects $\dot{V}O_2$ max's measured by uphill treadmill running were on the average 7% higher than those measured by the cycle ergometer. Astrand and Saltin[10]

showed that $\dot{V}O_2$ max's obtained by supine cycling were 15% less than those obtained by sitting cycling, and that $\dot{V}O_2$ max's for only arm cycling were 70% those of sitting cycling.

This would suggest that the $\dot{V}O_2$ max should be determined using the activity described by the specific task. Also suggested is that an individual specifically trained for one type of activity such as cycling would manifest a relatively higher $\dot{V}O_2$ max (i.e., have a selective advantage) compared to someone else who may be trained in another activity such as rowing, when tested in his/her trained mode. This latter case may be true to some extent in highly trained athletes; however, the subjects of Astrand and Saltin's study[10] suggest otherwise. The rank order of five subjects across six activities remained constant with the exception of one adjacent interchange in two of the activities.

This would suggest that it makes little practical difference in the mode that the $\dot{V}O_2$ max is determined. What would be required, however, is an adjustment in the percentage of $\dot{V}O_2$ max that a task may be required to be performed. For example, a simple lifting task of moving 15 kg cartons from floor to table at a rate of 10 repetitions a minute may cost 25 ml/kg/min. If this task were to be sustained for many hours, it would be reasonable to ask someone to work at no more than 50% of their $\dot{V}O_2$ max. However, an individual with a $\dot{V}O_2$ max of 50 ml/kg/min determined by uphill treadmill running would actually be performing this task at a high percentage of $\dot{V}O_2$ max. The $\dot{V}O_2$ max associated with the actual task (e.g., measured by increasing rate of repetitions) may actually be on the order of 40 ml/kg/min, and the subject actually working at 63% of his/her capacity.

Given these limitations the establishment of a valid criterion for job performance involving aerobic demands can be formulated using the concept of percentage of $\dot{V}O_2$ max.

Strength

Establishment of a valid criterion for job performance involving strength is not so simple. There is no simple common demoninator to express strength capacity as there is in endurance capacity using $\dot{V}O_2$ max. The actual cost of strength oriented tasks cannot be non-invasively determined. Also, because tasks requiring near maximal or high force development involve such factors as muscle mass, recruitment of additional muscle fibers, and enhanced sympathetic tone, performance of the task is affected by factors such as previous strength training and experience, motivation, and concentration.

The strength aspects of fitness are also very specific for the task considered. For example, a task may require high force generation by the legs, but involve the upper torso minimally. Other tasks may have the opposite characteristics.

Fortunately, inspection of the MOS task list revealed that in excess of 90% of those tasks having non-trivial strength requirements were characterized by being a single lift performance or repetitive lift and carry performance. Therefore, a single maximal dynamic lift could be used as the criterion variable that reflects task strength performance in the Army. However, lifting tasks that require repetitive lifting obviously require the ability to generate enough force to move objects many times. If an individual's maximal single lifting capacity is 45 kg, but the task requires repetitively lifting 40 kg, it is doubtful that individual will be able to sustain the lifting task. It would be reasonable, then, to require an individual's maximal lifting capacity to exceed by a certain percentage the requirements of the task. However, to determine how one's repetitive lifting performance relates to one's single maximal lifting performance remains to be done. If these two measures of performance are

fairly well correlated then it would be reasonable to set these strength standards in terms of the maximal lifting capacity (MLC) after accounting for the percentage of MLC it would be reasonable to perform the job task. Setting the percentage of MLC depends on a number of factors including duration of the task, efficiency, and injury risk.

Practically, then, a valid criterion of strength performance is suggested by the observatic that over 90% of the strength tasks require only lifting. Prediction of individual maximal lifting capacity would address this second component of work fitnes

It should be kept in mind that both these components of physical capacity represent an attempt to simplify and quantify physical job requirements. Thus, $\dot{V}O_2$ max and MLC, while possessing high face validity as measures of two aspects of physical capacity are not measures of real job performance in the context of the Army. Because of this limitation, it is necessary to accept the validity of these two criteria as estimates of true physical performance requirements on an experienced opinion and subjective basis.

Swedish Physical Fitness Screening System

The present development of a methodology to screen for physical fitness at the time of enlistment has been derived from methods and techniques formulated by the Swedes. Fitness testing is based on measurement of two components labeled "Muscular Power" and "Physical Working Capacity."[3]

Physical Work Capacity is measured using the method of Tornvall[2]. This test is based on the estimation of the subject's maximal exercise rate in six minutes using the cycle ergometer. It is calculated using Eq (1).

$$\text{Log } W_{\text{max, 6 min}} = \frac{\log t - \log 6}{4959} + \log N \qquad (1)$$

$W_{max, 6\ min}$ is the estimated maximal work performance for 6 minutes, t is the maximum performance time, and N is the actual work load used in performing the test.

Nordesjo and Schele[3] were able to show that for 84 males there was a correlation of -0.71 between $W_{max, 6\ min}$ and time to complete a 2.8 km cross-country course with a 22 kg pack using a monetary reward as incentive. Thus, about 50% of the variation in performance times was accounted for by $W_{max, 6\ min}$ performance in this sample.

Lifting capability was employed as a criterion performance in evaluating isometric strength measures as predictors. Subjects were required to lift an ammunition box weighing 20 kg and measuring 20x25x40 cm from the ground to a platform 103 cm high. The box was lifted then lowered 100 times as quickly as possible. Correlations of time with isometric strength measures were significant but moderate - being on the order of -0.25 to -0.45.

The third criterion measure that was tested was carrying capacity. The subject was required to carry as far as possible a 17 kg case in each hand equipped with a canvas carrying strap attached slightly off-center. No gloves were allowed. The subjects walked around a 400 m track until they could no longer hold a case. The criterion measure was the time to exhaustion. Again, correlations with strength measures were significant, but moderate - varying between 0.25 and 0.47.

The Swedes have demonstrated that the relative simple measures of physical work capacity ($W_{max, 6\ min}$) and of isometric strength significantly correlate with criterion measures they consider relevant. Accordingly, they have incorporated four tests of capacity distributed between the two categories of fitness previously mentioned with a nine point scale for each category. An

individual's point scale position is determined by his level of performance on the tests. Table 4 illustrates the point scale pairings with levels of performance in the muscle strength tests and the physical work capacity test.[11] Muscle strength performance level is determined by a weighted sum[12] of the force measurement performances on the three tests of handgrip, arm, and leg isometric strength.

Table 4

Relationship between point scale and measure of performance for Swedish Physical Fitness Screening System

Point Scale	Muscle Power[a] (Kilopond)	Physical Work Capacity (Kpm/min)
9	250- >	1651- >
8	240-249	1551-1650
7	230-239	1451-1550
6	215-229	1351-1450
5	200-214	1251-1350
4	175-199	1151-1250
3	133-174	1051-1150
2	100-134	901-1050
1	< - 99	801- 900
0		< - 800

[a] Muscle Power = 1.7 x (handgrip) + 1.3 x (knee extensor) + 0.8 x (elbow flexor) (From Ref 12)

Establishment of standards of test performance related to actual job specialty task demands was accomplished initially using an "experienced opinion" approach. Selected job tasks were studied and performance demands of the task were "translated" into levels of performance on the tests for two categories of fitness. In practice these standards of test battery performances for specific job specialties vary with demand and resources. In this way, "the levels of requirement could then be adjusted to fit the actual resources, or in other words,

they could be evened out so that they corresponded in quantity and quality to the performance of the current population."[1]

PROPOSED USA PHYSICAL FITNESS SCREENING SYSTEM

The system proposed for the U.S. Army follows the basic principles utilized by the Swedish military personnel selection system. The system is to screen accessions at the time of enlistment for their suitability to perform the physical work demands of their expected MOS. Screening is to be based on two aspects of fitness - stamina and strength.

Aerobic Capacity

From the previous discussion a measurement of stamina capacity is suggested by estimation of the maximum oxygen consumption ($\dot{V}O_2$ max). In essence this is what is indirectly being measured by the Swedish physical work capacity test, $W_{max, 6\ min}$. Nordesjo[13] demonstrated on a sample of 27 men that the correlation between $W_{max\ 6,\ min}$ and $\dot{V}O_2$ max in l/min was 0.88. Thus, in this sample 77% of the variation in performances on $W_{max, 6\ min}$ is accounted for by the $\dot{V}O_2$ max. Tornvall[2] similarly demonstrated a high correlation of 0.94 between $W_{max, 6\ min}$ and $\dot{V}O_2$ max on nine subjects. Unfortunately, use of the cycle ergometer to predict $\dot{V}O_2$ max, while highly efficacious, is impractical under the U.S. system of induction screening. This is due to the larger numbers processed (60,000 in Sweden versus 534,000 in USA, per year), small amount of time allocated for screening (one day for USA, two days for Sweden), fiscal restraints in capital outlay and maintenance, and maintaining a technically competent staff to administer and maintain a relatively "complex" screening system.

Development of a test to screen for endurance capacity must be constrained by the aforementioned factors. The test procedure must be

technically simple to administer and short in duration. Finally, it must be inexpensive and durable.

With these constraints in mind it was decided to inspect two factors in developing a prediction system for $\dot{V}O_2$ max. These two factors were anthropometric measures that correlate significantly and strongly with $\dot{V}O_2$ max, and simple performance measures possessing the same attributes.

Step Test

The first practical procedure to predict $\dot{V}O_2$ max using a relatively simple submaximal performance test is that of Astrand and Ryhming[14]. They developed a nomogram to predict $\dot{V}O_2$ max based on heart rate response to a submaximal work load on either the cycle ergometer or the step test. The basis for this nomogram is the demonstrated linear relationship between oxygen consumption ($\dot{V}O_2$) and heart rate. It is the use of the step test that meets the constraints aforementioned. The Astrand-Ryhming nomogram is expressed in equation form[15] by Eqs (2) and (3) for men and women respectively.

$$\dot{V}O_2 \text{ max} = \frac{195-61}{P-61} \quad \dot{V}O_2 \tag{2}$$

$$\dot{V}O_2 \text{ max} = \frac{198-72}{P-72} \quad \dot{V}O_2 \tag{3}$$

P is the steady-state pulse rate at the submaximal oxygen consumption, $\dot{V}O_2$. The terms 195 and 198 for men and women respectively represent the maximum heart rate during maximal aerobic exercise. The terms 61 and 72 for men and women respectively represent the "resting" heart rate. Probably a better term for "resting" would be basal since it would be inappropriate to determine this term by resting pulses. Resting pulses are easily affected by factors other than level of energy expenditure. This includes among other factors the level of anxiety as mediated by catecholamine release.

If one is willing to accept these constants for basal and maximum heart rates in the population considered here (enlistees) then one can predict $\dot{V}O_2$ max by measuring the pulse rate on a step test associated with a given oxygen consumption ($\dot{V}O_2$). First, however, it is necessary to have some estimate of $\dot{V}O_2$. In a laboratory setting one could actually measure $\dot{V}O_2$ at the same time the pulse rate was being measured. Practically, however, an estimate of $\dot{V}O_2$ must be made which accounts for three major factors affecting it. These factors are the size of the subject, the step height, and the stepping frequency.

It is obvious that a subject's energy expenditure for a stepping test would depend on his/her size The entire body mass is being raised vertically in such a task. A 100 kg male would be doing proportionately more mechanical work than a 50 kg female stepping the same height. Accordingly, the heavier individual would be required to expend more energy to raise the greater body mass the set step height. This factor is compensated for by expressing the energy cost of the stepping task on a per kilogram body weight basis. Margaria, et al.[16] demonstrated that when the energy cost (i.e., $\dot{V}O_2$) of stepping at a given height and frequency was expressed as ml O_2/kg/min the variation in energy cost due to size was effectively taken into account.

The effect of step height and stepping frequency on the value of $\dot{V}O_2$ is again intuitively obvious on a purely mechanical basis. Margaria, et al.[16] presents a simple nomogram to determine $\dot{V}O_2$ on a ml O_2/kg/min basis for a given step height and stepping frequency. No sex difference is suggested in Margaria, et al.'s article, therefore none is presumed. One can predict $\dot{V}O_2$ max using either Eqs. (2) or (3) with this estimated value of $\dot{V}O_2$ and the measured pulse rate. $\dot{V}O_2$ max either on a l/min or ml/kg/min basis is predicted by expressing $\dot{V}O_2$ in the appropriate units.

An additional correction to Eqs (2) and (3) is required when considering a population of subjects with a relative large age range. Astrand[17] demonstrated that an overestimation of $\dot{V}O_2$ max was inherent in the use of these two expressions for older people. Accordingly, a correction factor for age was introduced. These are given by Eqs (4) and (5) for males and females respectively.[15]

$$R_m = \frac{100}{100 + 1.37\,(\text{Age}) - 33.2} \tag{4}$$

$$R_f = \frac{100}{100 + 1.14\,(\text{Age}) - 23.0} \tag{5}$$

The correction factor, R_i, would be multiplied by the predicted $\dot{V}O_2$ max calculated directly from either Eq (2) or (3) to achieve a more accurate estimate of the true $\dot{V}O_2$ max.

Anthropometry

The second factor to be considered in developing a prediction scheme for $\dot{V}O_2$ max is anthropometric measures. The work of Buskirk and Taylor[18] illustrates the association between $\dot{V}O_2$ max and anthropometric measures. On a sample of 54 males they showed that the correlation between $\dot{V}O_2$ max on a l/min basis and fat-free weight was 0.85. Fat-free weight was estimated by immersion densitometry. They also demonstrated a correlation of 0.63 between $\dot{V}O_2$ max and body weight. Thus, in this sample 72% of the variation in $\dot{V}O_2$ max can be accounted for by fat-free weight, or lean body mass. Forty percent of the variation would be accounted for by considering just body weight. It would appear that the use of lean body mass in developing a predictive relationship for $\dot{V}O_2$ max would be efficacious.

Immersion densitometry, however, does not lend itself to rapid screening of large numbers of people. Accordingly, a "direct" measure of lean body mass as

offered by immersion densitometry cannot be considered. Methods of estimating lean body mass, however, are available. Measurements of skinfold thickness have been shown to correlate strongly with amount of body fat[19-22]. Haisman[19] reports a correlation of 0.76 between body fat content measured by densitometry and that estimated by the combination of four skinfolds. The estimation procedure of Durnin and Womersley[20] offers a simple straight-forward method for determining body fat. Body density is estimated by the expression of Eq. (6).

$$\rho = c - m \log(\text{sum of 4 skinfolds}) \tag{6}$$

The four skinfolds are the biceps, triceps, subscapular, and supra-iliac measured in millimeters.

The coefficients c and m vary with age range and sex. Table 5 details values of the coefficients for sex and age ranges. The percentage of fat is then estimated by Eq. (7).

$$\% \text{ BF} = \left(\frac{4.95}{\rho} - 4.50 \right) \times 100 \tag{7}$$

Lean body mass is calculated with Eq. (8)

$$\text{LBM} = \text{Wt}(100 - \% \text{ BF})/100 \tag{8}$$

Wt is the subject's body weight.

Table 5

Linear regression coefficients for the estimation of body density from the logarithm of the sum of four skinfolds.[a]

$$\rho = c - m \log(\text{sum of four skinfolds})$$

	Age (years, For Males		
	17-19	20-29	30-39
c	1.1620	1.1631	1.1422
m	0.0630	0.0632	0.0544

	Age (years) for Females		
	16-19	20-29	30-39
c	1.1549	1.1599	1.1423
m	0.0678	0.0717	0.0632

[a]From Ref 20

Measurements of step test performance and adiposity provide indirect estimates of aerobic capacity. Each factor is relatively simple to determine and measures operationally distinct aspects of aerobic capacity.

Strength Capacity

Development of a screening procedure for muscle strength capacity procedes along the same general principles enumerated above for aerobic capacity. As previously stated, the strength requirements for U.S. Army MOS's can, to a large extent, be approximated by a capacity to lift objects from the ground to a platform, and by lift and carrying capacity. The work of Poulsen[23–25] is particularly applicable to this situation. Poulsen[23] measured the maximum lifting capacity of 21 males and 25 females. The lifting task was to raise a wooden box 30x35x26 cm with handles to a standing position using a

straight back, straight arms, and flexed hip and leg technique. Performance on this maximal lifting task was then correlated with body weight and isometric back extensor strength. Correlation of maximum dead lift capacity with body weight and isometric back strength were 0.06 and 0.72 respectively for men and 0.28 and 0.78 respectively for women. The correlations were not significant at a type I error probability of 0.05 for the body weight correlation; however, the small number of subjects mitigates against detecting a correlation less than 0.4 at this level of significance.

Poulsen[23] also tested a theoretical model for predicting maximum dead lift capacity. The model stated mathematically is given by Eq. (9).

$$M_{max} = 1.4\ F - \tfrac{1}{2}\ Wt \tag{9}$$

M_{max} is the predicted maximum weight lifted, F is the isometric back strength, and Wt is the body weight. This model represents the theoretical effect of isometric back strength performance and body weight on lifting capacity. Correlations between actual and predicted maximum lifting capacity were 0.76 and 0.73 for males and females respectively.

The most significant conclusion drawn by Poulsen[23] from this investigation was "that the maximum weight a person can lift can neither be fixed as a standard load, nor defined as a load related to the person's body weight." It would appear that performance measures offer the best predictive capability from this study.

Further support for an isometric strength test extends from the work of Rasch and Pierson[26]. They studied the relationship between body size, isotonic weight lifting performance, and isometric strength performance on 27 males. The correlation between the sum of maximum weights lifted in the two hands press, two hands curl, supine bench press, and two hands reverse curl, and the

sum of the two measures of isometric elbow flexor and elbow extensor strength was 0.69. They also report a correlation of 0.45 between body weight and isotonic strength.

These studies would suggest that the role of isometric strength evaluation would be appropriate in developing a model to predict maximum lifting capacity (MLC). Anthropometric measures would appear to play less of a role, but it would not be inappropriate to evaluate the extent of these measures in an enlistee population in accounting for variation in MLC. It is also apparent that the isometric strength test should mimic the actual lifting task as closly as possible. Therefore, the actual lifting task needs to be more rigidly defined.

The Swedes employed a lifting task as one of their criterion measures from ground level to a platform height of 103 cm. Inspection of the MOS task list descriptions revealed that the most common lifting task involved lifting into a bed of a cargo truck. The bed height of the standard 5 ton cargo truck is 132 cm. A task described as lifting a load from ground level to a platform height of 132 cm would involve a number of muscle groups. These would include leg extensors, back extensors, arm flexors, and possibly grip strength.

A compounding factor is introduced by specifying the lift height to be constant for the criterion task. The effect of body size would be suspected to be much more important. The criterion tasks of Poulson[23] and Rasch and Pierson[26] were designed to mininize body size effects. It is readily apparent, however, that larger, taller individuals would have a distinct advantage over smaller individuals in lifting to a set height. The appropriate design for the criterion task must reflect the overall purpose of the investigation. The laboratory investigation appropriately studies physiological mechanisms and thereby tries to compensate for perturbing effects of body size and habitus. The purpose of this

study, however, is to develop a methodology by which to predict performance in a real world task environment. A single lift to a set platform height best mimics the actual task demands in the real world. This also may enhance the importance of anthropometric measures in deriving a predictive scheme for the criterion task.

The addition of repetition to a lifting task adds additional factors in performance capabilities. Jorgensen and Poulsen[24] address these issues in setting tolerance limits for repetitive lifting. They demonstrated "that in repetitive submaximal lifting both the capacity of the oxygen transport system and the muscle strength in the back act as limiting factors." Probably the most practical consideration they showed was that "nothing is gained by increasing the weight of the burden above 50% of the maximum" lifting capacity. The work output per unit time does not increase. There are also increased injury risks and back pain associated with lifting tasks approaching the capacity of the individual[27,28]. This suggests then, that categorization of any repetitive lifting task must account for both strength and endurance aspects, and that an individual capacity in both aspects of fitness must be taken into account for proper screening for a job task requiring repetitive lifting.

The Role of Gender

The role of gender in developing a model to predict performance capacity in a criterion task or variable remains to be examined. Gender itself has no role in setting the standard of performance for the criterion variable. Standards are to be set by the requirements of the job tasks as mediated through the criterion variables or tasks. However, the role of gender in performance on the predictive tasks and variables must be taken into account. This is true for measures reflecting both aerobic and strength capacity. Astrand and Rhyming's[14]

nomograms for predicting $\dot{V}O_2$ max separate sex. This is due to the fact that at a given percentage of $\dot{V}O_2$ max a female's heart rate will be on the average ten beats per minute higher than a male's. Drinkwater[29] states that "in most instances a given workload will be a greater strain on the female cardio-respiratory system than on a male." One explanation for this gender difference is that women must compensate for a smaller oxygen carrying capacity due to smaller blood volumes and lower hemoglobin levels by increasing cardiac output. Increasing heart rate is one means by which cardiac output is increased. Compensation by increasing stroke volume to increase cardiac output is relatively less effective due to the smaller heart volume in females. It is thereby suggested that women's $\dot{V}O_2$ max is largely limited due to hemoglobin level and relative heart size. It is readily apparent then, that gender should be considered in any predictive test incorporating heart rate as a variable.

Similar characteristics are seen when isometric back strength is correlated with maximum dead lift capacity[23]. Poulsen[23] showed that on the average men lifted 8-10 kg more than women at identical levels of maximum isometric back strength capacity. Again, consideration of gender is suggested in development of a predictive test using isometric strength measures.

The same characteristic is also apparent in determination of percent body fat from skin fold measurements. Purportedly, the distribution of subcutaneous fat in females is greater than that of males. Durnin and Womersley[20], however, dispute this contention. Their data using immersion densitometry techniques indicates a higher proportion of body fat situated subcutaneously in males relative to females. They also site the work of Forbes and Anirhakimi[30] using a ^{40}K dilution technique to estimate body fat as support for their conclusion. Regardless of the direction of difference in proportion of fat distribution

between males and females, an operational difference effect must be considered in correlating skinfold measures with a criterion variable.

Guidelines for Setting Standards

It is not the purpose of this presentation to actually set stamina and strength standards for occupational assignment qualification. However, the methodology by which standards can be set lends itself to this presentation.

Strength

The basis for setting strength standards has already been alluded to previously in the context of Jorgensen and Poulsen's[24] work. They have demonstrated that in a repetititve lift and carry task, exceeding a load of 50% of MLC will not increase work output per unit time. This observation is relevant, however, only in the context of job task demands approaching the limits of physiological capabilities for strength and endurance for a sizable proportion of the population. If, for example, the task demand is only to lift a load of 50 kg four times a day it would be inappropriate to allow only individuals with MLC's of 100 kg or greater to perform such tasks. The proportion of the population with this high MLC is not very high, and one would be left with a dearth of manpower in a MOS with this type of task demand. Setting the percentage of MLC higher would qualify more personnel for the MOS, but at the cost of increased injury incidence.

Establishment of a relationship between frequency of lift and "allowable" percentage of MLC cannot be based on limitations of endurance capacity in this case of infrequent "heavy" lifting. Rather, it would be more efficacious to base the relationship on some a priori estimated, and acceptable, injury or incapacity incidence. For example, an injury rate of 1 person per 1000 people per week may be deemed "acceptable". The relationship between frequency of lift and % MLC

would then be derived such that at the point the injury incident rate equalled 0.1 a certain value of % MLC is paired with the correspond. ifting frequency. In this manner guidelines could be established for strength-requirements in job tasks with infrequent, though heavy, lifting. Unfortunately, the data base to derive guidelines on this basis does not exist, and could be difficult to obtain. One is left with the choice of using estimation and experienced opinion in setting these guidelines.

In the case of muscle strength the main purpose of the guidelines is to categorize the MOS task list in the proper cluster level. For example, a MOS job task requiring a single lift of 35 kg would be rated as requiring medium strength and fall in the Delta Cluster according to Table 3. However, the strength requirements for a MOS job task requiring repetitive lifting of 35 kg, five repetitions a minute would need to take into account the repetition factor. Therefore, using as a guideline 50% of MLC for repetitive lifting an individual would need a MLC of 70 kg to qualify for this latter MOS. The MOS requiring only a single 35 kg lift is less strength demanding. It might seem reasonable (after trying to compensate for injury incidence rate) to allow as high as 80% MLC for a guideline for infrequent single lifts. Therefore, an individual with a MLC of 44 kg would qualify for this MOS.

This adjustment procedure was at least qualitatively used in the initial sorting of MOS into clusters. What remains to be done, however, is a transformation of the "Muscle Strength Requirements" listed in Table 2 from a job task lifting requirement to a MLC requirement. This would most efficaciously be done by inspecting representative job tasks at the three strength requirement levels and deriving a MLC requirement after taking into account repetition and injury incidence rate factors. The muscle strength requirement in

terms of MLC could then be set by some scheme (averaging, the highest requirement in a cluster, etc.) for that cluster.

Stamina

A scheme for setting aerobic standards and sorting MOS's with non-trivial aerobic requirements is more readily devised. The work of Bink[31] suggests a method by which to develop these standards. The critical elements in determining endurance characteristics of a job task are the energy cost of the task, the $\dot{V}O_2$ max, and the duration of the task. Bink[31] suggests a model to relate these three characteristics of stamina performance as that given by Eq. (10)

$$\frac{\dot{V}O_2}{\dot{V}O_2\ max} = m \log t + b \qquad (10)$$

$\dot{V}O_2$ is the energy cost of the task expressed as a rate (i.e. l/min or ml/kg/min), t is the time to exhaustion, and m and b are empirically determined constants. This model states that the proportion of $\dot{V}O_2$ max an individual can work is linearly related to the logarithm of the time to exhaustion.

The assumption that performance intensity (i.e., $\dot{V}O_2/\dot{V}O_2$ max) decreases in a linear manner with log t is well established experimentally[2,32-34]. The work of Glesser and Vogel[32] particularly illustrates the relationship using cycling as the task. They tested eight males for endurance time at various submaximal exercise loads ranging from 60% to 100% of their $\dot{V}O_2$ max as measured on the cycle. They were able to demonstrate the utility of a linear relationship between exercise capacity and log t for this mode of exercise, and also suggested that this logarithmic relationship may in fact be mediated by the kinetics of glycogen utilization. One of the practical demonstrations of this study was that an individual on the average could work at 50% of his $\dot{V}O_2$ max for 8 hours. This

would appear to be the upper limit for the "average" fit individual, and thereby it would be inappropriate to actually expect someone to work at 50% of $\dot{V}O_2$ max for eight hours routinely.

The coefficients m and b of Eq. (10) can be ascertained either empirically, as they were in the study of Gleser and Vogel[32], or by assumption, as Bink[31] has done. Two points in the linear $\dot{V}O_2/\dot{V}O_2$ max versus log t relationship will define these constants. Bink[31] made the assumption that one point was determined by the presumption that an individual could work at his/her $\dot{V}O_2$ max for four minutes. A second assumption was that an individual could be expected to work at about 35% of $\dot{V}O_2$ max for eight hours (480 min) per day in a 48 hour work week. These two assumptions alone are sufficient to determine values of the constants. Accordingly, the relationship expressed by Eq. (10) becomes,

$$\frac{\dot{V}O_2}{\dot{V}O_2\ \text{max}} = \frac{\log 632! - \log t}{3.47} \tag{11}$$

The solution is more generally presented by Eq. (12) if only the $\dot{V}O_2$ max:4 min assumption is maintained and variable retained for percentage of $\dot{V}O_2$ max for 480 minutes.

$$\frac{\dot{V}O_2}{\dot{V}O_2\ \text{max}} = \left(\frac{1-p}{\log 120}\right)\left[\left(\frac{1}{1-p}\right)\log 480 - \left(\frac{p}{1-p}\right)\log 4 - \log t\right] \tag{12}$$

p is the proportion of $\dot{V}O_2$ max assumed for 480 minutes (or eight hours).

These guidelines can again be used for two purposes. The first is the appropriate sorting of MOS's into a cluster with the proper level of aerobic requirement. The second is to set the levels of $\dot{V}O_2$ max required for screening for cluster endurance standards.

The critical elements in arriving at a standard for $\dot{V}O_2$ max have already been enumerated above, and are reflected in Eq. (12). The $\dot{V}O_2$ max required for a representative job task can be determined by solving for $\dot{V}O_2$ max in that

equation. The energy cost ($\dot{V}O_2$) of the task can be measured, the duration of the task is specified by the job description, and a reasonable assumption can be made concerning the percentage of $\dot{V}O_2$ max an individual can be expected to perform the task routinely. Again, representative job tasks can be evaluated in this manner in each cluster, and an overall cluster standard for endurance can be ascertained in terms of $\dot{V}O_2$ max by any scheme considered appropriate (i.e., averaging, most demanding, etc.)

STUDY DESIGN AND METHODS

Fort Jackson

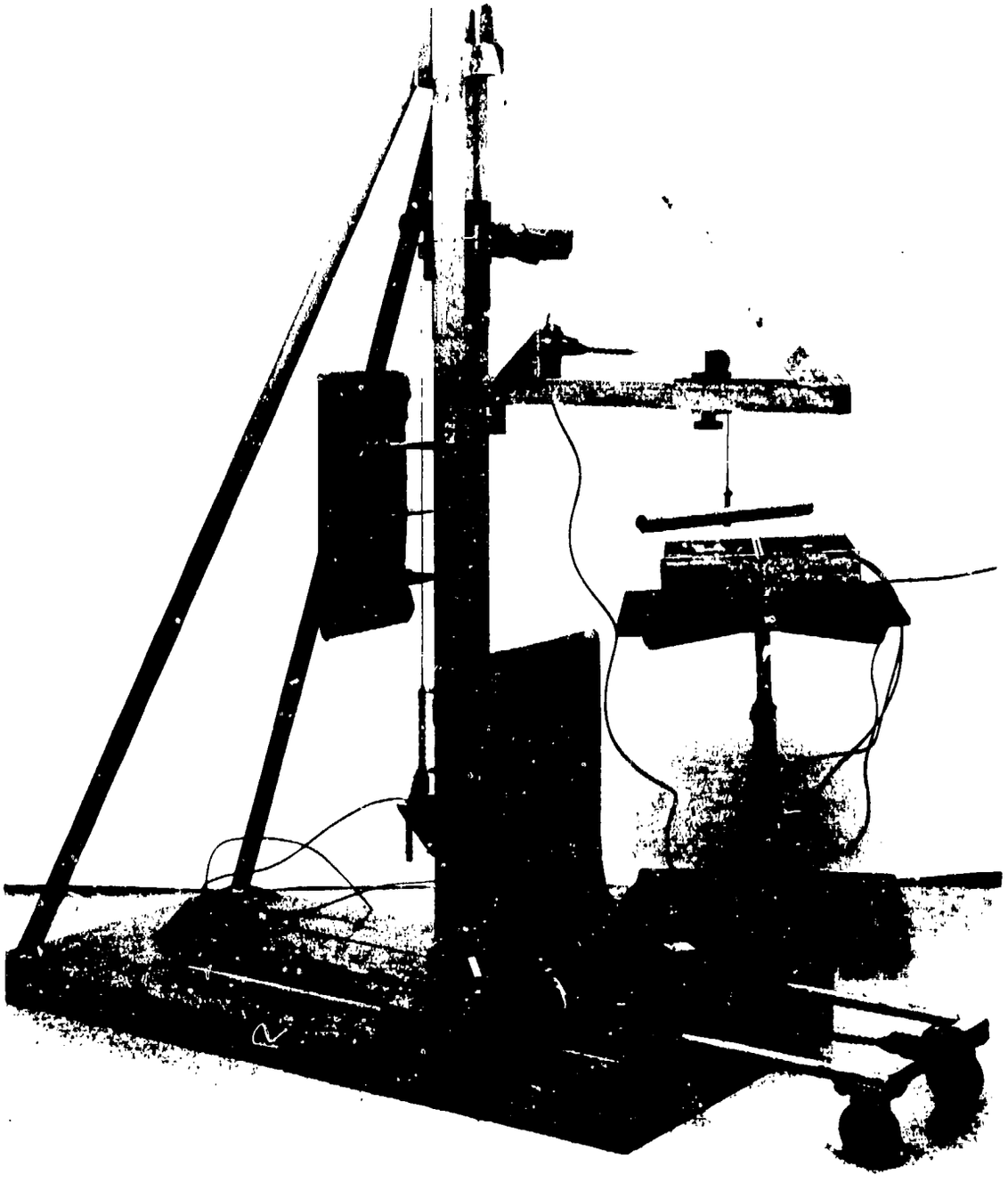
The data to develop predictive models for the endurance and strength criterion variables was collected in two phases. The first phase was in conjunction with a multi-faceted study at Fort Jackson, SC, in the winter and spring of 1978. This study examined recruit population characteristics for a large number of physiological, anthropometric, psychological, and job performance tasks. Information was collected immediately prior to the start of basic training and during the last week of the eight week training period. A total of 948 male and 496 female recruits were initially evaluated. From this sample 100 males and 100 females were selected for $\dot{V}O_2$ max determinations. The selection procedure was not based on any overt randomization scheme, but rather a first-come, first-serve process over a three week period. The age of the 200 subjects for $\dot{V}O_2$ max determination varied from 17 years to 25 years. Eighty-seven males and 57 females were retested at the end of the eight week basic training period. The loss of subjects was due to various reasons such as administrative discharges, medical profiles, etc.

The $\dot{V}O_2$ max was determined using an interrupted uphill running treadmill technique[35,36]. Subjects ran for six minutes at 5-6 mph, 0% grade, as a warmup.

They then rested 5-10 minutes followed by 2-4 additional runs lasting 3-4 minutes. The exercise load was increased by increasing the grade by 2.5%. The $\dot{V}O_2$ max was operationally defined as being successive $\dot{V}O_2$ determinations less than 0.15 l/min in difference at two contiguous exercise loads. Expired air was collected in the last minute of an exercise load via a mouthpiece attached to a Koegel valve into a Douglas bag system. Gas analyses were performed using an AEI S3-A oxygen analyzer and a Beckman LB-2 carbon dioxide analyzer. Volume was measured using a Collins chain-compensated gasometer. The heart rate was electrocardiographically determined using a modified V_5 lead position.

Concommitant with the $\dot{V}O_2$ max determinations, information on four skinfold measurements (biceps, triceps, subscapular, and suprailiac) using the Harpenden skinfold calipers; height; weight; measures of isometric leg extensor, upper torso, and trunk strength; and step-test heart rates were collected. Figure 1 illustrates the device used to measure isometric strength of the leg extensors, the upper torso, and the trunk. A previous technical report details the development, testing, and validation of this device[37].

Figure 1 - Static Strength Device



Step-test heart rates were measured at three levels for each subject. These were 10 cm, 20 cm, and 30 cm for females and 20 cm, 30 cm and 40 cm for males. Subjects remained two minutes at each level. Pulse rate was determined by an electrocardiographic cardiotachometer. The stepping frequency was 25 complete steps per minute. No attempt was made to control for environmental temperature or humidity. The data collected on these recruits prior to basic, among other things, was to be used in the formulation of a predictive model for $\dot{V}O_2$ max. The effect of training was also to be accounted for over the eight week basic training period.

Fort Stewart

At the time of the design and execution of the first study the criterion variable for strength performance had not been formulated. Development and execution of the second phase was based primarily on the need to address this issue of a predictive model for the strength criterion variable. One hundred eighty-three males and 44 females were studied during this second phase study. These personnel were experienced active duty troops assigned to the 24th Infantry Division, Fort Stewart, GA, during the fall of 1979. They cannot be considered representative of the U.S. Army as a whole, or respresentative of inductees in terms of population distribution characteristics for the data collected. These soldiers were studied during two three-week periods in September and October. They were required to return four to five times during a three-week period.

The first session collected data on performance in a two-mile run, number of pushups in two minutes, and number of situps in two minutes. The second session collected data on six measures of isometric strength. Three of the isometric strength measures are those described above. Three additional

measures included handgrip strength and two measures of upright-pull strength. Figures 2, 3 and 4 illustrate these latter devices. The handgrip device was adjusted through a turn-buckle assembly so that the angle at the metacarpal-phalangeal joint of the index finger approximated 110° and the proximal interphalangeal joint angle was 150°. The upright pull devices assess a composite of isometric strength of arm, shoulder, back and leg muscles. They were devised to mimic the maximal lift capacity task. Figure 3 illustrates the subject position for the lower pull. The distance from ground platform to handle was set at 38 cm. The distance for the higher pull was set at 132 cm. The upright pull platforms were placed against a wall and the subject positioned facing away from the wall. The wall was used as a vertical guide to assist the subject in maintaining proper form. The subjects were instructed <u>not</u> to lean back or stand on tip-toes in the 132 cm pull. Subjects were also instructed to use a lifting form similar to the dead lift form discussed below for the 38 cm pull.

Figure 2 - Handgrip Device

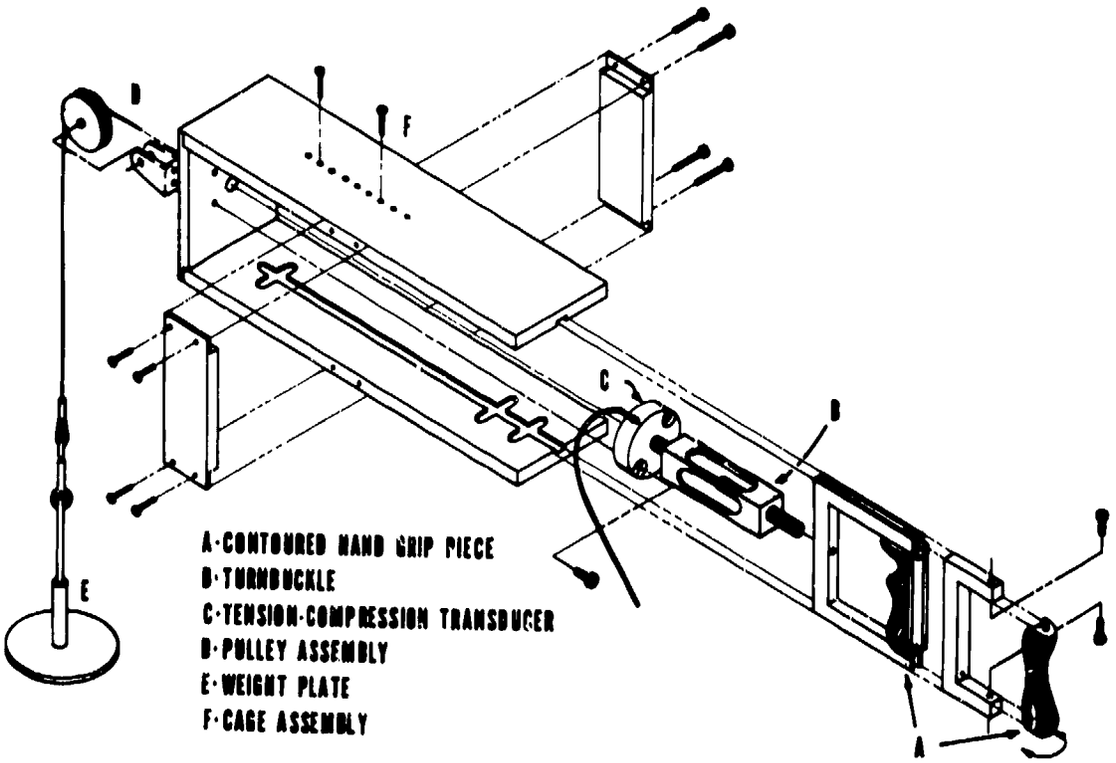


Figure 3 - 132 cm Upright Pull Device

Figure 4

38 cm Upright Pull Device

The third session dealt with anthropometric measures of height, weight, four skinfolds and pulse rate at a single step test level - 30 cm for females and 40 cm for males. Stepping frequency and time at the level were the same as in the Fort Jackson study. Subjects had a two minute warm-up at 20 cm and 30 cm for females and males respectively immediately prior to stepping at the next higher level. Subjects returned for a fourth session to measure performance on the strength criterion variables.

The primary criterion variable measured was the MLC to 132 cm. Weights were placed in a metal rectangular box with handles. This box was constructed according to the dimensions given by Poulsen[23]. The handles were padded with foam rubber and adhesive tape. All subjects began lifting the empty box (15.6 kg). Weight was added to the box in increments ranging from 1.2 to 11 kg depending upon the ease with which the subject lifted the previous weight. Subjects were allowed as much time as they desired between lifts (usually 2 minutes). They reached their MLC usually in 4-10 lifts. Subjects were instructed to use a flexed hip, straight back, and straight arm lifting technique. They were instructed to use one smooth motion in lifting from ground to the platform. No jerking was allowed.

Four guidelines were used in determining when subjects had reached their safe maximums. The first was inability to actually place the weighted box on the platform even when proper lifting technique was being used. The second was the observation of marked hyperextension of the back in an attempt to "angle" the edge of the box onto the lip of the platform. The third was degeneration of a single smooth evenly controlled lift into jerked disrupted segments. The fourth was the deterioration of the straight back form into marked thoracolumbar flexion during the initial part of the lift. Many subjects were physically capable

of placing the weighted box onto the platform at higher weights. However, the MLC criterion was operationally defined with the modifier of needing to be a safe execution of the task. Determination of the safe MLC was made by the subjective judgment of an investigator using the four guidelines enumerated above.

Upon completion of the determination of the MLC all female subjects were tested for maximum dead lift capacity. Inability to stand erect with the weight using proper form was the criterion for establishing performance capacity. Males were not tested since a constraint of 100 kg was placed on the maximum weight allowed to lift, and in a subsample of approximately 40 men, all were capable of dead lifting this weight.

Subjects were allowed to rest for half an hour to two hours. Performance on a lift and carry task was then evaluated. All subjects were required to lift the weighted box described previously (weighing 25 kg), carry it five meters, and lower it beyond a marked line. They were to then turn around and lift the box and carry it back the five meters to the starting line. The number of five meter trips in ten minutes was the measure of performance. The subjects were instructed to make as many trips as possible, as quickly as possible, and always using proper lifting technique. The lift and carry was always demonstrated by one of the investigators, and carrying was always demonstrated using a run. The subjects were then cautioned to pace themselves, but to do the best job they could. Subjects were monitored constantly by one of the investigators for proper lifting technique. No overt encouragement was offered the subjects; however, when subjects appeared to be not trying, they were told, "Do the best job you can do," and, "Try to make one more trip."

At the conclusion of the ten-minute performance bout the subjects were allowed a rest period of 20 to 30 minutes and then returned for an additional ten-

minute performance period. This time the box weight was increased to 43 kg. The performance measure, safety precautions, and instructions where the same as for the 25 kg performance bout. The subjects executed these lift and carry tasks indoors on a concrete surface in regulation boots. Ambient temperature and humidity were not controlled.

STATISTICAL DESIGN AND METHODS

The modeling method most appropriate for the objectives of this project is multiple linear regression. The technique is described in any intermediate statistical text[38,39]. The previous sections have developed a modeling approach based on lawlike relationships between a single criterion variable and a number of independent variables. The suggestion of lawlike relationships is based on intuition and observation. The development and use of a relationship, however, subsumes a system or method by which this very relationship may be derived and verified. The uses of a lawlike relationship encompass three major practical aspects[40]. First, the relationship integrates a variety of different sets of data by describing how one variable varies approximately with another under all the various conditions of observation. Second, the relationship can be used to determine whether additional sets of data obey or disobey the same relationship displayed by previous sets of data. Third, it can be used for prediction, which subsumes the relationship is obeyed by a different set of data.

The lawlike relationships of science often are mistakenly thought of as reflecting cause-and-effect or some fundamental "law of nature"[40]. These lawlike relationships, however, would be more correctly interpreted as primarily describing the functional relationship between variables under a limited range of conditions. The use of statistical methods, particularly regression methods, is not meant to yield "laws of nature." The discussion in previous sections has

stressed both physiological (i.e., lawlike) relationships and operational (i.e., statistical) relationships in developing a reasonable scheme to assess a recruit population's physical work capacity. The use of statistical methods to arrive at a practical means of screening a population does not in itself require <u>any</u> theoretical or lawlike physiological relationship to exist between criterion measure and screening variable. It is entirely possible to develop practical empirically valid screening procedures using statistical methods to "relate" variables where there would appear to be no reasonable causal relationship. An apparent increase in admissions to the obstetrical unit of a hospital with the phase of the moon is illustrative.

It is with these constructs in mind that an empirically based model can be developed for the purpose of screening enlistees for physical performance capacity using a statistical methodology. As an example, the choice of four skinfold measures as an estimation of percent body fat, which in turn is related to lean body mass, which in turn is related to $\dot{V}O_2$ max illustrates the intuitive physiological basis for this choice. It is sufficient to show that a significant statistical relationship between a measured $\dot{V}O_2$ max, (which has physiological meaning in terms of work performance) and some mathematical transformation of four skinfold measurements (which has little direct physiological meaning in terms of work performance) exists, in arriving at a practical model for predicting aerobic capacity.

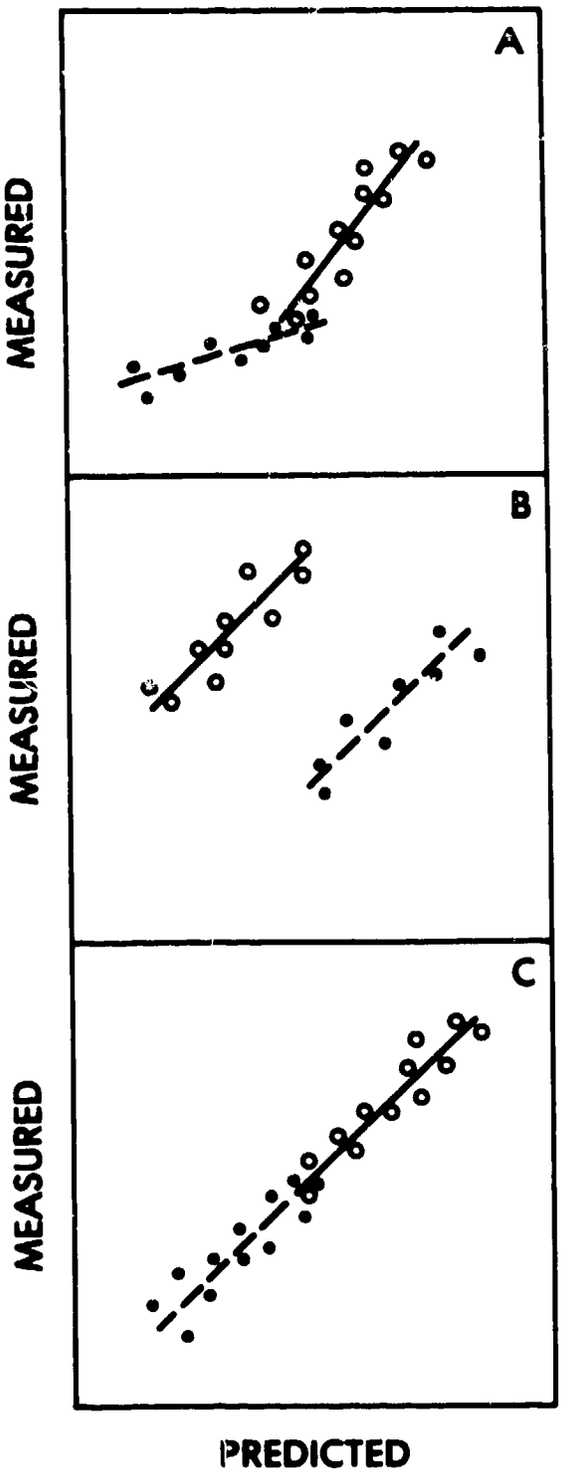
Accounting for Gender Effects

Most analyses of physiological data that develop models of some criterion in terms of apparent constitutent variables tend to derive separate relationships for males and females. The reason for this separation is based on demonstrated differences in physiological measures and mechanisms[29] between the sexes. In a

simple correlational analysis two aspects must be considered in establishing this difference. These aspects can be labeled as the parallel behavior and the coincidental behavior.[38] Analysis of these aspects falls under the technique of analysis of covariance.

An analysis for parallel behavior addresses the issue of differing slopes between two or more groups for which there is a demonstrated relationship (i.e., correlation) between two variables. An analysis for coincidental behavior addresses the issue of relative elevation above the coordinate axis. Figure 5 depicts three possible situations in determining the parallel and coincidental behavior of two groups. Figure 5a indicates no parallel or coincidental relationship between the two groups. Figure 5b depicts parallel behavior but noncoincidental behavior. Figure 5c illustrates both parallel and coincidental behavior. It should be readily apparent that a test which "fails" for parallel behavior mitigates against further testing for coincidental behavior. A pair of groups that passes the test for parallel behavior but fails that for coincidence allows a model to be developed for the criterion variable whereby group membership becomes a constituent or independent variable.

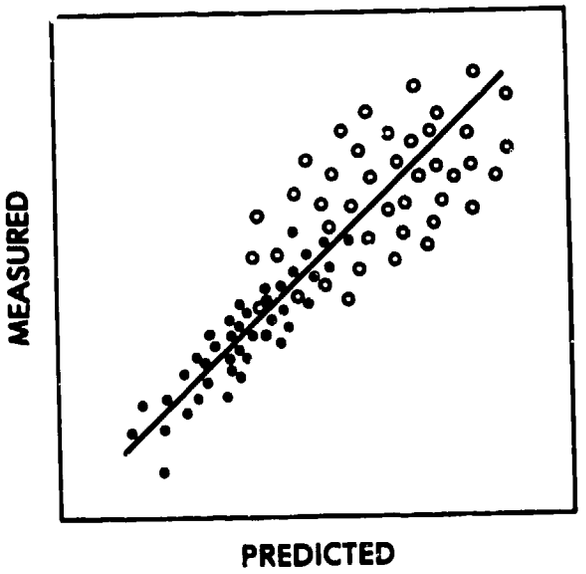
Figure 5 - Idealized Parallel and Coincidence Effects in Two Groups
a) No parallel cr coincidence effects
b) Parallel but non-coincidence effects
c) Parellel and coincidence effects



In the case of gender, if it can be shown that a significant functional relationship between the criterion variable and the independent variable exists, and that the slope relationship between the two sexes is parallel and coincidental, then a model can be developed for the criterion variable for the sexes combined, and gender (i.e., group membership) excluded as an independent variable. In a multiple regression model based on multiple independent variables this would presume parallelism and coincidence for all constitutent variables. In the case where parallel behavior is demonstrated but coincidence is not, then gender would be added as a constituent variable. If the data failed both parallel and coincidental tests then separate models for males and females would be mandatory.

In the case of a model developed with gender as an independent variable another aspect must be considered. That is the comparison of the residual variances of the two sexes when each group is considered separately[39]. Ramifications of this comparison involve the derivation of meaningful confidence limits. If it can be demonstrated that the residual variances are homogeneous, then the confidence limits can be reliabily used. However, if the test for homogeneity of variance fails one may be hard pressed to develop a model with confidence limits that would not be misleading. Figure 6 demonstrates the effect of compiling data from two groups that at least pass the test for parallel behavior, but possess heterogeneity of residual variance. What is suggested by this phenomenon is an inadequate understanding or accounting of the functional relationship, of group membership, or both. If such a model were to be used practically, one might be put in the position of overestimating the population characteristics of one group and underestimating in the other group.

Figure 6 - Effect of non-homogeneity of variance due to group differences



Multicollinearity

One of the purposes in using the technique of multiple regression is the determination of the relative importance of the independent variables in modeling the criterion measure. A problem arises, however, when the constituent variables are highly correlated among themselves. The greater this intercorrelation, the less reliably one can ascertain the relative importance of the partial regression coefficients. This phenomenon is called multicollinearity.

The eigenvalues of the symmetrical correlation matrix of the predictor variables reflect the degree of multicollinearity in a data system. Eigenvalues are a set of numbers reflecting certain characteristics of square matices and are actually derived from the entries in a matrix. It is sufficient to this presentation to discuss the use of these numbers in detecting the characteristics of multicollinearity in an intercorrelation matrix without going into detail about their derivation. If there is no relationship between predictor variables (i.e., they are mutually independent or orthogonal) then all eigenvalues would be 1.0. A high degree of multicollinearity is reflected in the eigenvalues by the first eigenvalue being many times greater in magnitude than the last one, and the last eigenvalue approaching zero.

The issue of multicollinearity could be particularly important in developing a practical model for maximum safe lifting capacity. It might be expected that high performance on any one measure of isometric muscle strength by a subject would be associated with high performance on any other device. This expectation underlies the intuition that in general, strong people are strong all around. However, if one is constrained to decrease the number of predictors in arriving at a usable model of performance, one might be hard pressed to pick the most important constituent variables.

Given that the issue of which predictor variables to be incorporated in a model of the criterion measure can be resolved, one is still plagued by another problem associated with multicollinearity. Estimates of regression coefficients in a given sample may be gross misestimates of the population regression coefficients. Alternately expressed, estimates of the regression coefficients may markedly fluctuate from sample to sample. Thus, one is presented with the possibility that a model derived from a given sample may fail in its job to model the population.

The problem of multicollinearity can be compensated to some extent by a number of mathematical techniques. The technique utilized in this project is termed ridge regression[41-42]. Ridge regression attempts to arrive at a better estimate of the population regression coefficients by introducing bias into the statistical procedure in deriving the coefficients. The effect of introducing bias is to decrease the variance of the coefficient estimates at the expense of increasing the standard error of the estimate. The biasing procedure is effected by adding to the diagonal of the correlation matrix a small positive constant. Formally then, the expression for the vector of standard regression coefficients is given by Eq. (13) in the case of straight multiple regression.

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{13}$$

$\hat{\beta}$ is the vector of standard regression coefficient estimates, $X'X$ is the correlation matrix of independent variables, and $X'Y$ is the correlation vector of each independent variable with the criterion variable. Ridge regression introduces bias into the correlation matrix by adding to it the expression kI.

$$\hat{\beta}^* = (X'X + kI)^{-1} X'Y \tag{14}$$

kI is the scaler multiplication of the identity matrix by a small positive number k.

The difficulty in using this technique is determining the value of k to be used. Unfortunately there is no universally accepted procedure to determine the optimal value of k. In practice a plot of the vector of standard coefficients versus the bias k allows one to see what effect biasing has on the coefficients. "Stable" coefficients may show only a gradual change in being driven to zero as k approaches infinity. Unstable coefficients may be driven to zero much more rapidly compared to other coefficients. Finally, some coefficients may initially change markedly in magnitude, sign, or both, and then "stabilize" at some value of k. The choice of the bias parameter is subjective. However, it appears that results are not affected significantly by an inexact choice of k[43].

Cross Validation

The most important issue in developing a model for a criterion measure is the validity of that model when applied to a population where only the predictor variables characterize that population. Validation is an issue that must continually be addressed in a project of this type. Population characteristics change over time, and thereby so may the relationship between criterion measure and predictor measure. Developing a model using a relative small subset of the population presents the issue of whether that subset is truly representative of the population. This issue is particular important in the context of a screening program where conclusions and decisions may be made affecting both individuals and manpower distribution.

The issue of validation can be initially addressed by separating the subjects from which the model is being developed into two subsets. Effectively what is done is to develop two models based on these two subsets and compare both the form of the models and the performance of the models using as data the contrasting subset. If it can be demonstrated that the two models are similar, then the two subsets can be combined to formulate a combined model.

In the context of ridge regression, cross validation offers the additional benefit of better selection of the bias coefficient k[44]. The standard deviation (S_p) of the residuals calculated by using as data a separate set of data than that used in developing the model can be plotted against the bias coefficient used in the model. If a minimum is demonstrated in this plot of S_p versus k then this suggests the degree of bias in the modelling data that should be used. This "confirmatory" bias can be contrasted with that bias more subjectively determined by the inspection of the $\hat{\beta}^*$ vs k plot. The process of cross validation can be effected by just switching the two subsets, and using as model data that used previously as validation data and visa versa.

Utilization of the Model

Once a model has been developed it remains to be determined exactly how that model is to be used. The model so developed can be used as a "point prediction" (i.e., a "best guess") of the criterion measure, or it can be used in a probabilistic manner[45]. The use of the model in the latter manner can be restated by the question, "What is the (approximate) probability that an individual with this combination of predictor scores will get a criterion score above a specified value?"[45] In this situation it might be better to formulate the inquiry as, "How much higher must a recruit score above the cluster standard on the predictor model test so that one can be at least 75% (or 85%, or 95%) sure that the standard is being met?"

Determination of that minimal predicted score rests on three factors: the actual measured standard, the resolution of the predictive model as manifest by the standard error of the estimate, and the probability which one is willing to accept in knowing the accuracy of the screening process. This latter factor might better be illustrated by an example.

A cluster standard for endurance capacity might be set at a minimum $\dot{V}O_2$ max of 40 ml/kg/min. However, it would be expected that for those inductees scoring 40 ml/kg/min on the predictive test, half would in reality have true $\dot{V}O_2$ max's less than 40 ml/kg/min and the other half a greater $\dot{V}O_2$ max. Setting the predictive score cutoff at the cluster standard in effect sets the probability at 50%.

The predictive score cutoff can either be raised or lowered with respect to the cluster standard depending on the purpose of the standard. A conservative approach would dictate that one wants to be at least 99% sure that an individual truly meets the cluster standard. Setting the probability at 99% and with a given standard error of the estimate may result in only those individuals with predicted $\dot{V}O_2$ max's of 50 ml/kg/min or greater meeting the cluster standard. The advantages of such a conservative approach is practically assuring that personnel in the MOS's requiring high aerobic capacity truly can meet the physical demands of the job. However, such a high assurance is achieved by reducing the available manpower for those MOSs and thereby risking certain MOS's being under manned (and in turn, possibly increasing injury rates). One may wish to operate at a lower level of probability thereby increasing the available manpower, but at the risk of a higher proportion of individuals not being able to meet the work demands of the job.

Setting the prediction score cutoff to less than the cluster standard would suggest a completely different purpose in screening. This would emphasize manpower availability over quality of manpower. For example, setting the probability at 5% and generating some cutoff score less than the cluster standard would result in assuring that at least 95% of those individuals truly meeting the cluster standard being allowed into the high demand MOS's. However, the cost

of such a "liberal" screening standard is the inclusion of a sizable number of inductees into high demand MOS's that cannot truly meet the cluster standard (and again possibly increasing the injury rate, but by a different mechanism than in the conservative mode).

RESULTS AND DISCUSSION

In keeping with the necessity for validation and the methods discussed in the previous discussion, subjects in the two phases were grouped into two subsets. Males and females were grouped separately. Sorting was effected by the use of a random number table[46]. Thus, a total of four groups were generated for the Fort Jackson data and similarly for the Fort Stewart data. Different sections of the table were used for each sex and each phase. Before group selection was done, however, the Fort Jackson data were subjected to preliminary inspections and sorting.

In order to account for the effect of training in enhancing endurance capacity it was necessary to limit the sample size to just those individuals completing measurements of $\dot{V}O_2$ max on both pre-training and post-training phases. Additional subjects were eliminated if they missed more than one week of physical training, and if during either phase the determination of $\dot{V}O_2$ max did not meet the <0.15 l.min difference for a 2.5% increase in grade, or the grade was increased by less than 2.5% at the confirmatory work load. This selection decreased the number of subjects to 47 males and 48 females.

Sample Characteristics

Table 6 depicts the sample characteristics of the two groups for each sex for the Fort Jackson pre-training data. The slightly smaller numbers reflect additional deletion of subjects with incomplete data. Table 7 depicts the sample characteristics of the Fort Stewart data for each group for each sex.

Table 6

Sample characteristics of two groups for each sex for Fort Jackson pre-training data - mean ± standard deviation

Variable	Females group: 1	Females group: 2	Males group: 1	Males group: 2
n (number of subjects)	20	24	22	20
$\dot{V}O_2$ max (l/min, measured)	2.13 ± 0.284	2.10 ± 0.279	3.57 ± 0.329	3.56 ± 0.474
$\dot{V}O_2$AR (l/min, predicted step test)	2.12 ± 0.403	2.07 ± 0.331	3.20 ± 0.487	3.33 ± 0.703
LBM (lean body mass, kg)	41.1 ± 4.70	41.5 ± 4.23	59.8 ± 5.88	58.0 ± 7.53
Weight (kg)	56.7 ± 7.10	57.3 ± 6.11	73.4 ± 11.4	68.2 ± 10.2
Age (years)	19.6 ± 1.79	19.1 ± 1.32	19.0 ± 1.46	19.1 ± 2.00

Table 7

Sample characteristics of two groups for each sex for Fort Stewart pre-training data - mean ± standard deviation

Variable	Females group: 1	Females group: 2	Males group: 1	Males group: 2
n (number of subjects)	19	22	91	90
ML132 (safe MLC in kg to 132 cm)	32.7 ± 5.46	32.4 ± 5.65	57.1 ± 10.9	57.6 ± 9.37
LBM (lean body mass in kg)	44.2 ± 5.17	46.2 ± 5.43	61.9 ± 6.57	62.3 ± 6.19
AGE (years)	22.0 ± 3.27	22.4 ± 2.79	21.0 ± 2.20	21.1 ± 2.39
Isometric measure in kg: LEG (leg extensors)	96.9 ± 19.8	102 ± 33.0	161 ± 49.7	173 ± 40.9
TR (truck extensors)	53.0 ± 10.9	53.0 ± 12.1	80.8 ± 15.5	79.5 ± 17.0
UT (upper torso)	60.9 ± 16.8	60.7 ± 9.93	108 ± 16.4	108 ± 15.5
HG (handgrip)	35.3 ± 7.55	34.8 ± 5.95	54.6 ± 7.73	54.7 ± 9.05
UP38 (upright pull at 38 cm)	84.0 ± 18.6	89.0 ± 18.0	139 ± 21.4	140 ± 26.2
UP132 (upright pull at 132 cm)	39.5 ± 9.45	40.3 ± 10.7	60.6 ± 14.0	59.6 ± 14.8

In order to verify that the groups possessed similar distribution characteristics within each sex, a t was calculated for unequal variances[38]. The purpose of this t is to test for overall similarity between the two groups. Table 8 depicts these values of t for both phases of data. A small value of t supports similarity between groups while a large value suggests a significant difference in the sample characteristic. The use of multiple t-tests to compare multiple characteristics between two groups is subject to an enhanced type I error. This can be compensated for by setting the probability of accepting a falsely positive difference very low. If p is set at 0.01 then a value of t greater than 2.71 and 2.59 for 40 and 200 degrees of freedom respectively would meet this confidence limit criteria. None of the t values meet even the 0.05 level of confidence, and in fact 22 of the 28 comparisons don't even meet the 0.5 level. This strongly supports homogeneity of characteristics between groups.

Table 8

Test for homogeneity of distribution characteristics between groups using t test.

Fort Jackson t values

Variable	females	males
degrees of freedom	42	40
$\dot{V}O_2$ max	0.35	0.08
$\dot{V}O_2AR$	0.45	0.36
LBM	0.30	0.87
weight	0.30	1.55*
Age	1.07*	0.19

Fort Stewart t values

degrees of freedom	39	179
ML132	0.17	0.33
LBM	1.20*	0.42
Age	0.42	0.29
LEG	0.63	1.71**
TR	0.00	0.54
UT	0.05	0.04
HG	0.24	0.15
UP38	0.87	0.29
UP132	0.24	0.42

*significant at 0.5

**significant at 0.1

Model of $\dot{V}O_2$ max

Training

The issue to be dealt with first is the development of the predictive model for $\dot{V}O_2$ max. One of the aspects to be considered in developing the model is the effect of training in altering the $\dot{V}O_2$ max. The first consideration in accounting for a training effect is to document both the existence and then the degree of the effect. It is expected that the training program would result in an increase in $\dot{V}O_2$ max. A simple t test on the difference $\dot{V}O_2\ max_2 - \dot{V}O_2\ max_1$, where the subscripts refer to pre-training (1) and post-training (2), indicates existence of a significent increase. Table 9 illustrates the average difference in l/min, the standard deviation of the difference and the t value for the four groups. A one tailed t test was used to determine level of significance.

Table 9

Average difference in Fort Jackson post-training and pre-training $\dot{V}O_2$ max for each group and sex, and t test of significance for zero difference.

		females		males	
Variable	group:	1	2	1	2
n (number of subjects)		24	24	24	24
mean difference (l/min)		0.168	0.246	0.155	0.049
standard deviation		0.127	0.121	0.252	0.196
t value		6.47**	9.97**	3.01*	1.19

*significant at 0.025

**significant at 0.001

All groups displayed highly significant increases in $\dot{V}O_2$ max on an absolute l/min basis except the group 2 males. The most significant increases are displayed by the females. Group 1 females displayed an average increase of 0.17 l/min with 22 of 24 subjects showing a positive difference for $\dot{V}O_2\ max_2 - \dot{V}O_2\ max_1$. Group 2 females displayed an average increase of 0.25 l/min with 23 of 24 subjects showing a positive difference. Group 1 males average a 0.16 l/min increase; however, only 17 of 24 subjects displayed an increase. Group 2 males only averaged a 0.05 l/min increase with only 14 of 23 subjects indicating a positive difference.

This information suggests that females achieved greater positive training benefits as demonstrated by an increase in their $\dot{V}O_2$ max. However, because the females on the average have initial $\dot{V}O_2$ max's 60% of the males it could reasonably be suggested that they as a group have more to gain This data also suggests that the aerobic fitness level of the average female inductee is markedly less than that of male inductees even when accounting for a natural gender difference.

Although three of the four groups displayed significant increases in aerobic capacity, the magnitude of these increases on a l/min basis is not large. Fourteen of the 48 females did not display an increase greater than 0.15 l/min which is the criterion for determining $\dot{V}O_2$ max at two contiguous work loads. For the males 27 of 47 subjects did not exceed the 0.15 L/min criterion. This suggests that accounting for a training effect in developing a predictive model for $\dot{V}O_2$ max may not be very reliable or practical. The general effect of an eight week training period on increasing the $\dot{V}O_2$ max is so small that it would be impractical to incorporate this effect in the predictive model. The number of people that could reliably be determined to meet the standard who otherwise

would not without some accounting of a training effect would be relatively small considering the resolution of the model. With these limitations in mind it was decided to develop a predictive model for $\dot{V}O_2$ max based only on pre-training data.

Basis of $\dot{V}O_2$ max

In developing a model for $\dot{V}O_2$ max an aspect to be considered is, on what basis should the $\dot{V}O_2$ max be determined. An individual's $\dot{V}O_2$ max can be expressed on an absolute basis (i.e., liters of O_2/minute) or a relative basis (milliliters of O_2/kilogram body weight/minute). The choice depends to some extent on the situation to which the determination is to be applied. In physical work tasks with high aerobic requirements that involve primarily translocation of body mass, the $\dot{V}O_2$ max on a relative basis best accounts for an individual's work capacity. However, in tasks requiring repetitive translocation of sizable mass external to the body mass, the $\dot{V}O_2$ max expressed on an absolute basis best accounts for an individual's work capacity.

This latter observation is to some extent incomplete. In a task such as repetitively lifting an absolute mass, the size of the individual is an obvious mitigating factor in determining performance. A large person has a high $\dot{V}O_2$ max on an absolute basis by virtue of his/her size to a large extent. Similarly, a large person uses a smaller proportion of his/her strength capacity in performing the task by virtue of his/her larger working muscle mass. It would seem apparent, then, that basing the endurance standard on an absolute basis would be required for those large number of tasks in the military requiring repetitive translocation of sizable external mass. A "different" standard would appear to be necessary for those tasks involving primarily body mass translocation, and based on a relative measure of $\dot{V}O_2$ max. This is unnecessary, however.

Those tasks requiring repetitive translocation of external mass are both aerobically and strength demanding. Accounting for the strength demands of the task by requiring a given level of strength capacity will encompass the effect body size has as a determinate in effective performance. However, meeting the strength standard for a task by virtue of the sizable effect of body size does not preclude meeting the endurance requirements. It would seem apparent that a large individual who met the strength standard by virtue of his/her size may be less capable of adequately performing the task when contrasted with another individual who both meets the same strength standard and has a relative $\dot{V}O_2$ max 10 ml/kg/min higher. With these conditions and suggestions in mind, it was decided to develop a predictive model of $\dot{V}O_2$ max on a relative basis.

Three-Predictor Model

Table 10 depicts the results of the statistical tests for parallel and coincidental behavior. The comparisons are between groups for the same sex. Except for one comparison none of the t values are significant at the 0.05 level thereby indicating that for a given sex the parallel and coincidental behavior is homogeneous between groups.

Table 10

Test of Fort Jackson data for parallel and coincidental behavior using t tests, and homogeneity of variance using F test. Comparisons are between groups for the same gender.

	Females					Males				
Variable with $\dot{V}O_2$ max	n_1	n_2	t_p	t_c	F	n_1	n_2	t_p	t_c	F
% BF (percent body fat)	24	24	1.50	2.37*	1.02	24	23	0.76	1.77	1.62
$\dot{V}O_2$AR (ml/kg/min)	24	24	1.66	0.47	1.09	22	22	1.66	1.62	1.98

*significant at 0.05

Table 11 depicts the t values for tests of parallel and coincidental behavior for a given group between the sexes. It is readily apparent that each sex is similar in its parallel, or slope, behavior, but is markedly non-coincidental. This indicates then, that a single predictive model can be developed but that gender must be used in accounting for the offset in the relationship between the predictor variables (% BF, $\dot{V}O_2$ AR) and the criterion measure ($\dot{V}O_2$ max).

Table 11

Test of Fort Jackson data for parallel and coi 'dental behavior using t tests, and homogeneity of variance using F test. Comparisons are between sexes in the same group.

Variable with $\dot{V}O_2$	Group 1					Group 2				
	n_f	n_m	t_p	t_c	F	n_f	n_m	t_p	t_c	F
% BF	24	24	0.45	4.24**	1.59	24	23	0.94	8.24**	1.00
$\dot{V}O_2$AR (ml/kg/min)	24	22	0.08	6.92**	2.03*	24	22	0.31	11.02**	1.12

*significant at 0.05

**significant at 0.001

Tests of homogeneity of variance are also included in Table 11. These are F values. None of the F values for the Fort Jackson study are significant at the 0.05 level with the exception of the group 1 $\dot{V}O_2$AR F value of 2.03. In general it appears the groups are quite homogeneous with respect to the residual variance for the $\dot{V}O_2$ max data. Confidence limits thereby generated from a model combining both groups and gender for $\dot{V}O_2$ max should not be misleading.

Table 12 depicts the intercorrelation matrix for predictor and criterion variables for each group. All the correlations are significantly different at the 0.01 level. The correlation of sex with $\dot{V}O_2$ max was predicated on using the numerical designators 1 = male and 2 = female. This explains the negative value. The correlation so computed is referred to as a point biserial r. The square of this correlation coefficient has a special meaning. It is the proportion of the total variance of $\dot{V}O_2$ max in the sample population accounted for by simple group (i.e., gender) membership. Sixty two percent of the variance is accounted by gender in group 1, while 85% is accounted for in group 2.

Table 12

Intercorrelation matrix for criterion and predictor measures for each group in the Fort Jackson data.

Group 1 $n = 46$, $n_f = 24$, $n_m = 22$

	SEX	$\dot{V}O_2AR$	% BF	$\dot{V}O_2$ max
SEX	1.000			
$\dot{V}O_2AR$	-0.448	1.000		
% BF	0.684	-0.685	1.000	
$\dot{V}O_2$ max	-0.785	0.643	-0.839	1.000

Group 2 $n = 46$, $n_f = 24$, $n_m = 22$

	SEX	$\dot{V}O_2AR$	% BF	$\dot{V}O_2$ max
SEX	1.000			
$\dot{V}O_2AR$	-0.666	1.000		
% BF	0.861	-0.617	1.000	
$\dot{V}O_2$ max	-0.923	0.680	-0.874	1.000

The subsequent development of a predictive model incorporating gender as a constitutent variable will have its ratio of range to resolution determined to a sizable degree by a simple gender designator. Therefore, development of models with such "high" coefficients of determination should be viewed with this stipulation in mind.

The results of the ridge regression analysis for the two groups are presented in Table 13 and Figures 7 and 8. Contrast of the first eigenvalue with the third in group 1 reveals almost a ten fold difference. This characteristic suggests that multicollinearity may be a factor to be dealt with in group 1 data. Examination of the group 2 eigenvalues show the first to be almost 18 times the third, and thereby suggesting multicollinearity to be significant. Inspection of Figure 7 suggests that the standardized regression coefficients are relative stable. If any bias was warranted it should not exceed k=0.2. Inspection of Figure 8 reveals a higher degree of instability in the standardized regression coefficients for group 2 relative to group 1. It would appear that the gender designator is given too much weight at k=0.0. A range of bias of 0.1 to 0.3 for k is suggested.

Figure 7 - Group 1 model of relative $\dot{V}O_2$ max.

Variation of three standardized regression coefficients with bias

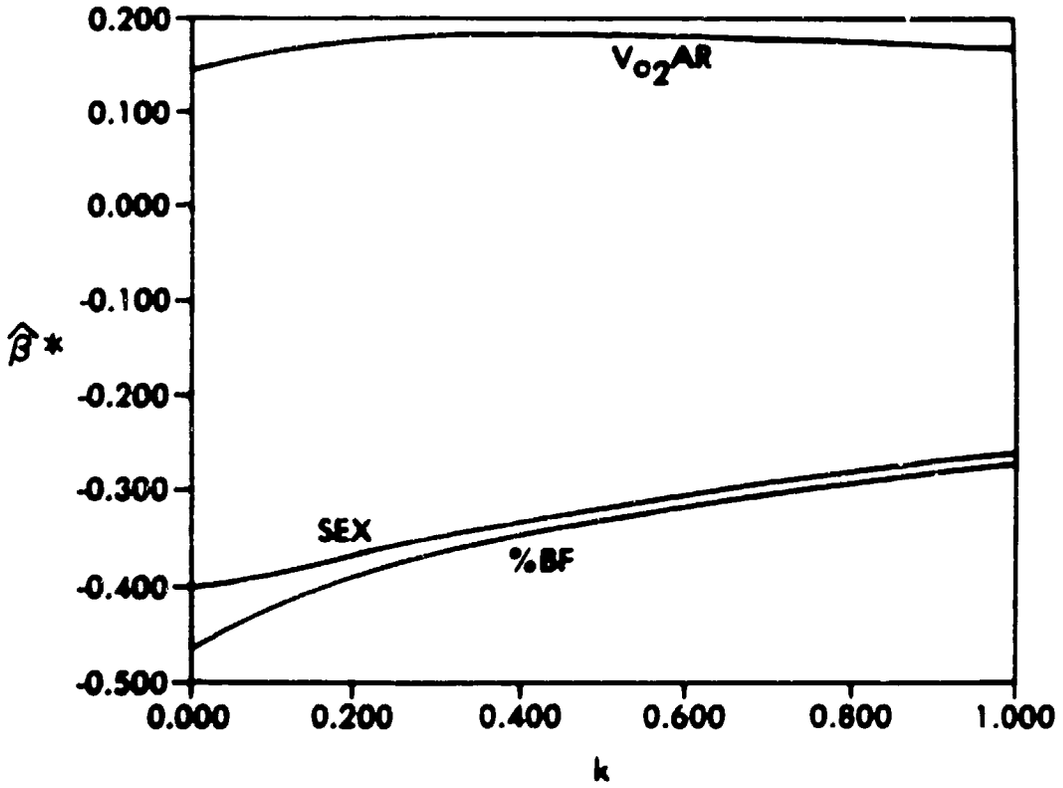


Figure 8 - Group 2 model of relative $\dot{V}O_2$ max.

Variation of three standardized regression coefficients with bias

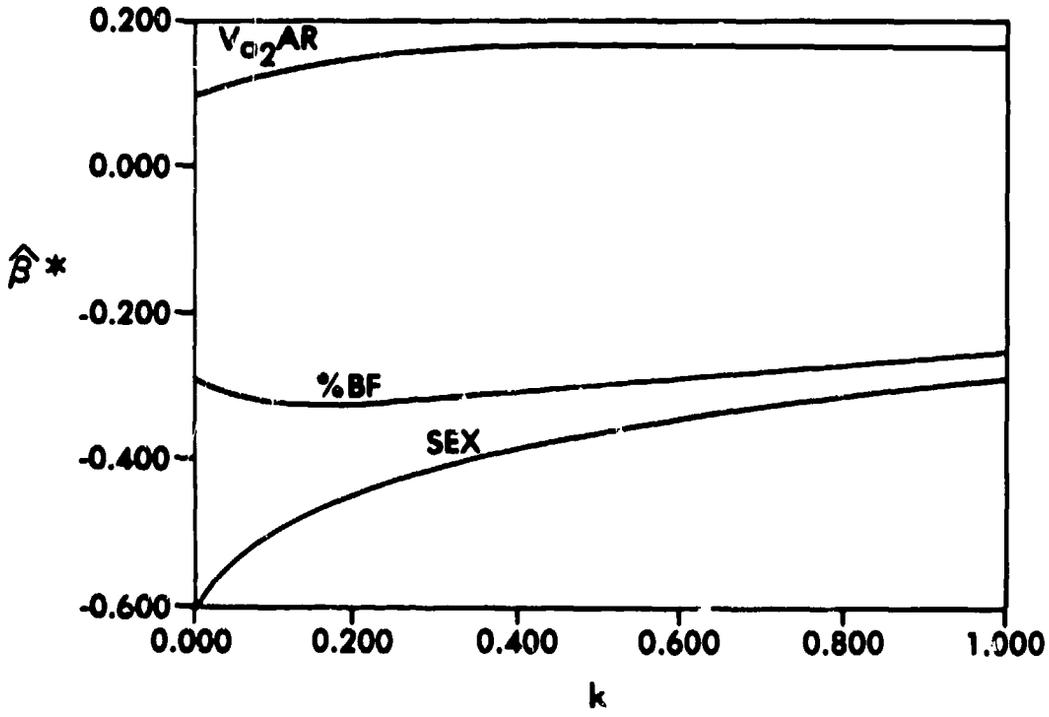


Table 13

Eigenvalues and unbiased standardized regression coefficients for the prediction of $\dot{V}O_2$ max from SEX, $\dot{V}O_2$AR, and % BF.

Group 1 model

variable	β weight	eigenvalue	degree
[illegible]	-0.402	2.217	1
$\dot{V}O_2$AR	0.144	0.552	2
% BF	-0.465	0.230	3

Group 2 model

variable	β weight	eigenvalue	degree
SEX	-0.610	2.435	1
$\dot{V}O_2$AR	0.095	0.429	2
% BF	-0.290	0.136	3

Figures 9 and 10 depict the cross validation procedure. Figure 9 is the standard deviation of the residuals of group 2 data used in the model generated from group 1 data versus the bias coefficient k. No minimum is illustrated thereby supporting that no bias is suggested for the group 1 model. Figure 10 is the S_p vs k plot of group 1 data used in the group 2 model. A minimum is indicated in the range $0.2 < k < 0.3$.

Figure 9 - Group 2 predictor data in three predictor Group 1 model for relative $\dot{V}O_2$ max. Variation of prediction standard deviation with bias.

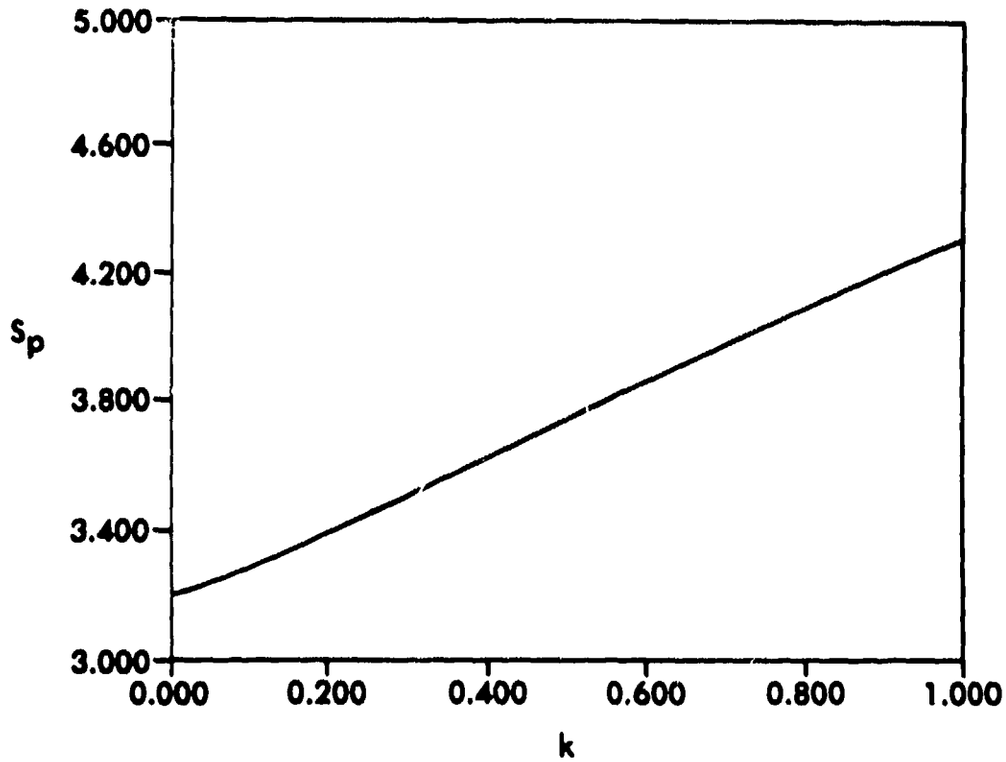
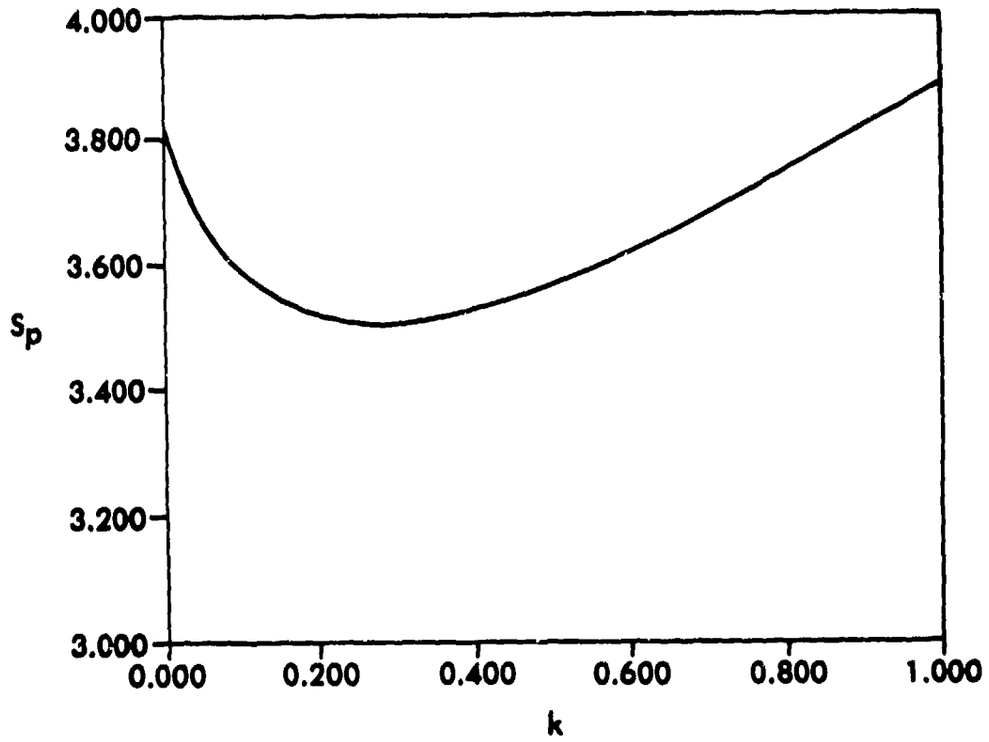


Figure 10 - Group 1 predictor data in three predictor Group 2 model for relative $\dot{V}O_2$ max. Variation of prediction standard deviation with bias.



Picking an arbitrary value of k = 0.25 for the group 2 model, and no bias for the group 1 model, standardized regression coefficients for the two groups are presented in Table 14. These coefficients, or weights, are remarkably comparable - the largest difference seen in the % BF coefficient. If no bias had been introduced into the group 2 model, weights of -0.610, 0.095, and -0.290 for the gender designator, $\dot{V}O_2$ AR, and % BF respectively would have been suggested by a simple multiple regression. These values are definately not as comparable to the group 1 weights, and one would be less sure as to the validity of a combined model. The decrease in the amount of variance accounted for by the group 2 model in using a bias of k = 0.25 is relatively small. At k = 0.0, R^2 = 0.881, while at k = 0.25, R^2 = 0.867. The gain in using the bias is illustrated by the 95% confidence limit range of the gender designator. At k = 0.0, the range is -0.385 to -0.834. At k = 0.25 the range is -0.337 to -0.528. This is a decrease in range from 0.449 to 0.191. It is a sizable gain for a relatively small trade-off in accuracy.

Also depicted in Table 14 are a number of squared correlation coefficients. The group 1 model accounts for almost 80% of the variance. A new sample of data used in the group 1 model would be expected[45] to have a lower R^2 on the order of 0.763. In fact, when group 2 data is used in the model an R^2 of 0.863 is generated. This strongly supports the group 1 model. A similar set of R^2's are depicted for the group 2 model. The group 1 sample data R^2 is slightly below the expected new sample R^2; however, this difference is not large enough to significantly detract from the group 2 model.

Table 14

Standardized regression coefficients and squared multiple correlation coefficients for two models of $\dot{V}O_2$ max (ml/kg/min).

model group:	1	2		1	2
k	0.0	0.25	estimator R^2	0.798	0.867
β weights:			new sample R^2	0.763	0.844
SEX	-0.402	-0.432	predictor R^2	0.863	0.789
$\dot{V}O_2$AR	0.144	0.153			
% BF	-0.465	-0.326			

The results of the cross validation procedure support combining both group 1 and group 2 data to generate a final model. Because of the ridge regression procedure, the relative magnitude of the β weights for both groups are comparable. The possibility of incorporating the ridge regression procedure is suggested in the combined groups model with k possibly varying between 0.0 and 0.25. The comparable weights presented in Table 14 can be used as a guide in selecting the combined group regression coefficients.

Table 15 and Figure 11 depict the ridge regression characteristics of the combined groups model for $\dot{V}O_2$ max. The first eigenvalue is 10.5 times greater than the last. This is of similar magnitude as group 1. Examination of the ridge plot suggests that the standardized regression coefficients are quite stable. The values of the β weights at k :0.0 are -0.454, 0.141 and -0.417 for the gender designator, $\dot{V}O_2$ AR, and % BF, respectively. These values are quite comparable to those presented in Table 14 for the two groups separately. This suggests that no bias is necessary in formulating a model of relative $\dot{V}O_2$ max for the combined groups data. The squared multiple correlation for this model is 0.839.

Figure 11 - Combined groups three predictor model for relative $\dot{V}O_2$ max.

Variation of standardized regression coefficients with bias .

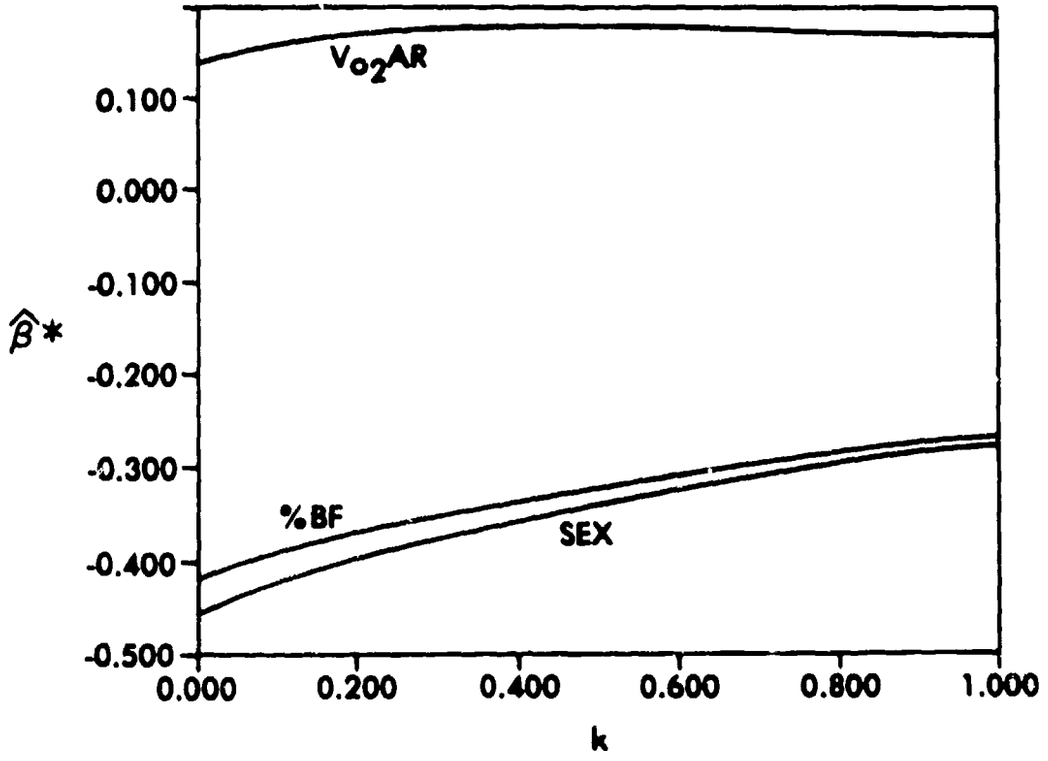


Table 15

Eigenvalues and unbiased standardized regression coefficients for a single combined groups model of $\dot{V}O_2$ max.

$R^2 = 0.839$

variable	β weight	eigenvalue	degree
SEX	-0.454	2.322	1
$\dot{V}O_2AR$	0.141	0.458	2
% BF	-0.417	0.220	3

Two-Predictor Model

The introduction of a practical usable method of screening for physical work capacity is predicated on a number of constraints. These constraints were alluded to briefly in previous sections. The model just developed for relative $\dot{V}O_2$ max includes two measures requiring no minor addition of time and investment of capital in initial procurement, maintenance, and purchase of expendable materials. These two measures are the determination of % BF and of predicted $\dot{V}O_2$ max from heart rate data. Examination of this latter measure in particular reveals a sizable stress on the induction processing system in terms of both time and capital outlay. Some induction centers process in excess of 200 people a day. A single set-up consisting of a variable height platform, a cardiotachometer, a metronome, electrodes, leads, and a timing device could only process 60 individuals in an eight-hour period assuming eight minutes from the start of one subject to the start of another. The initial capital outlay for this system would be $1125.00. The daily capital expenditure just for expendables (e.g., electrodes) would be $63.00. Maintenance of the electronic devices could expect to cost $50.00 per year. Larger induction centers would require at least four systems for males, and possible as many as two systems for

females. A minimum of one staff person to operate two systems would be required. It is readily apparent that introduction of the step test as one of the measures of aerobic work capacity would require a sizable commitment of personnel, initial capital outlay, and operating expenses.

With these costs in mind, and the fiscal and staff constraints placed on the enlistment processing system, it was decided to eliminate the step test as one of the screening devices for aerobic work capacity. Elimination of the step test, however, does involve some risks in trying to develop a model of aerobic capacity. With the step test eliminated, only the gender designator and % BF remain as predictor variables. A model developed on only these two variables ignores the aspect of performance, and thusly training, as a constitutent of aerobic capacity. The model thereby is predicated on the natural difference in aerobic capacity due to gender, and the empirical relationship between body habitus and $\dot{V}O_2$ max. A model so developed could be considered teleologically inadequate. However, the additional resolution offered by a teleologically "correct" model may not be worth the additional cost.

Figure 12 - Measured versus predicted relative $\dot{V}O_2$ max for the three predictor model

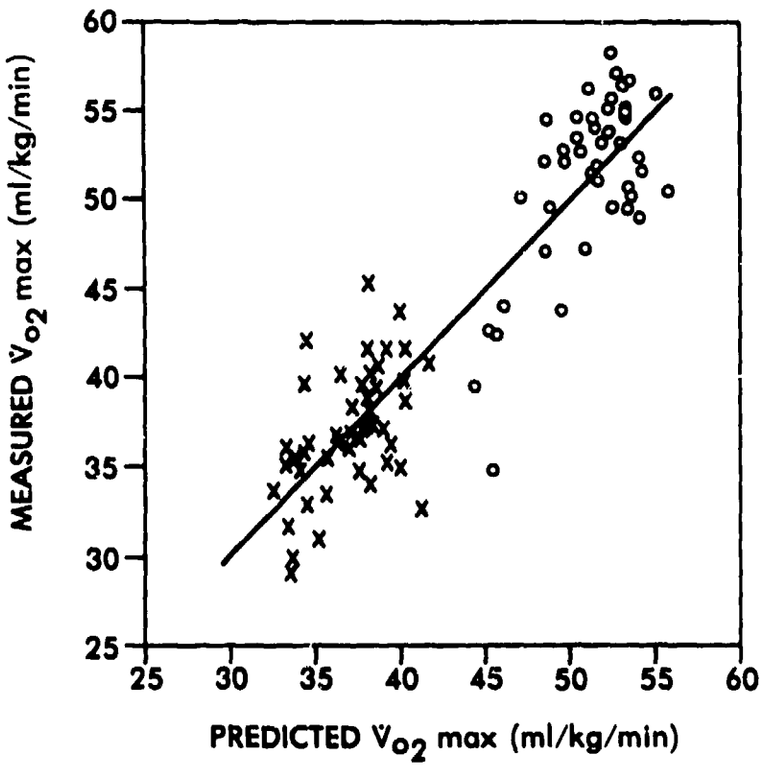
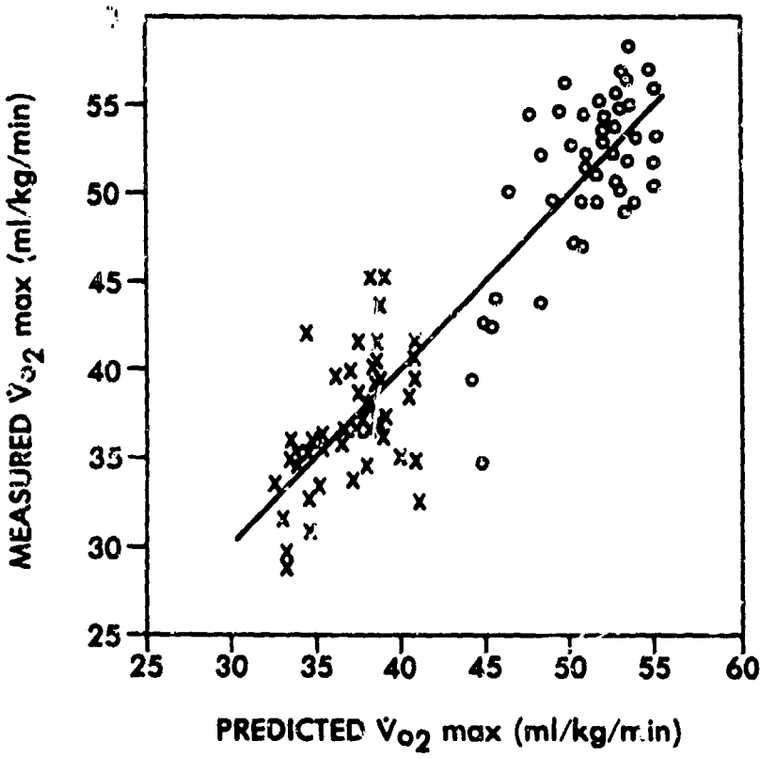


Figure 13 - Measured versus predicted relative $\dot{V}O_2$ max for the two predictor model



The effect of eliminating $\dot{V}O_2AR$ from the model is illustrated in Figures 12 and 13. The increase in R^2 in adding $\dot{V}O_2AR$ as a predictor to a linear model already consisting of the gender designator and % BF is 0.011. This increase is significant at the 0.05 level by an F value equal to 6.24 calculated by the ratio of the change in the sums of squares of the residuals to the mean sums of squares of the residuals of the expanded model on one and 88 degrees of freedom respectively. Although the addition of $\dot{V}O_2AR$ to the predictive model truly enhances resolution, it is difficult to evaluate the practical benefits of this additional resolution. Table 16 also depicts the breakdown of correctly and incorrectly classified subjects in the sample data for an artificial $\dot{V}O_2$ max standard of 42 ml/kg/min and 95% probability. A 95% probability requires an individual to score at least 47.7 ml/kg/min on the predictive model for $\dot{V}O_2$ max using only gender and % BF as predictors, and 47.5 ml/kg/min for the model adding $\dot{V}O_2AR$. The incorrect classification is further broken down into falsely positive (i.e., falsely meeting the standard) and falsely negative (i.e., not meeting the standard when in reality the subject does). With such a small sample of 91 subjects it is difficult to generalize with any degree of certainty about the expected proportions of incorrectly classified personnel in a population exceeding half a million.

Table 16

Classification of subjects for $\dot{V}O_2$ max for a cluster standard of 42 ml/kg/min and 95% probability for two and three predictor models.

Three Predictor

	positive ♀	positive ♂	negative ♀	negative ♂	percent correctly classified
true	0	37	44	2	91.2
false	0	0	4	4	

Two Predictor

	positive ♀	positive ♂	negative ♀	negative ♂	precent correctly classified
true	0	36	44	2	90.1
false	0	0	4	5	

A two predictor model using gender and % BF was developed using the same methodology described previously. Table 17 illustrates the expected coefficients for the two groups and the choice of bias used for the respective group. The magnitude of the β weights are not as comparable as the previous model incorporating $\dot{V}O_2AR$. Use of a bias in group 2 definitely improves the comparability. Figures 14 to 17 depict the relationships between vs k for groups 1 and 2 respectively, and between S_p vs k for model groups 1 and 2 respectively. The final choice of standardized coefficients for combined groups are presented in Table 17 also, and are based on a bias of k = 0.0. Figure 18 depicts the β vs k relationship for the combined data.

Figure 14 - Group 1 model of relative $\dot{V}O_2$ max.

Variation of two standardized regression coefficients with bias.

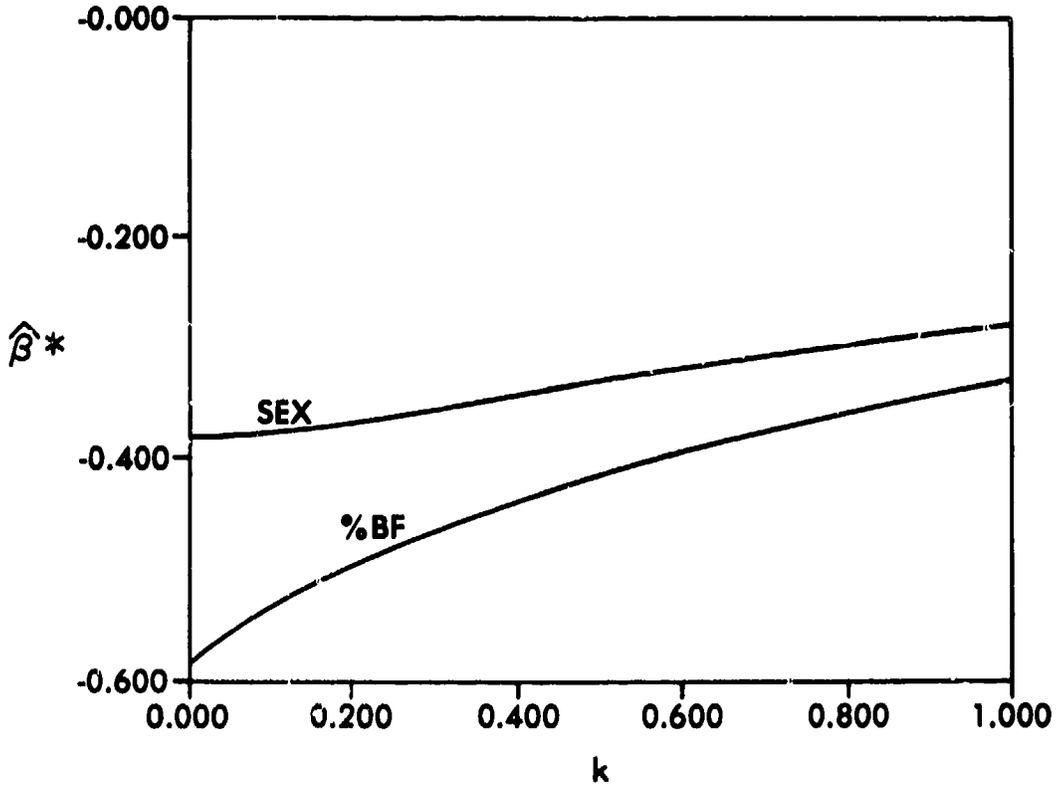


Figure 15 - Group 2 model of relative $\dot{V}O_2$ max.

Variation of two standardized regression coefficients with bias .

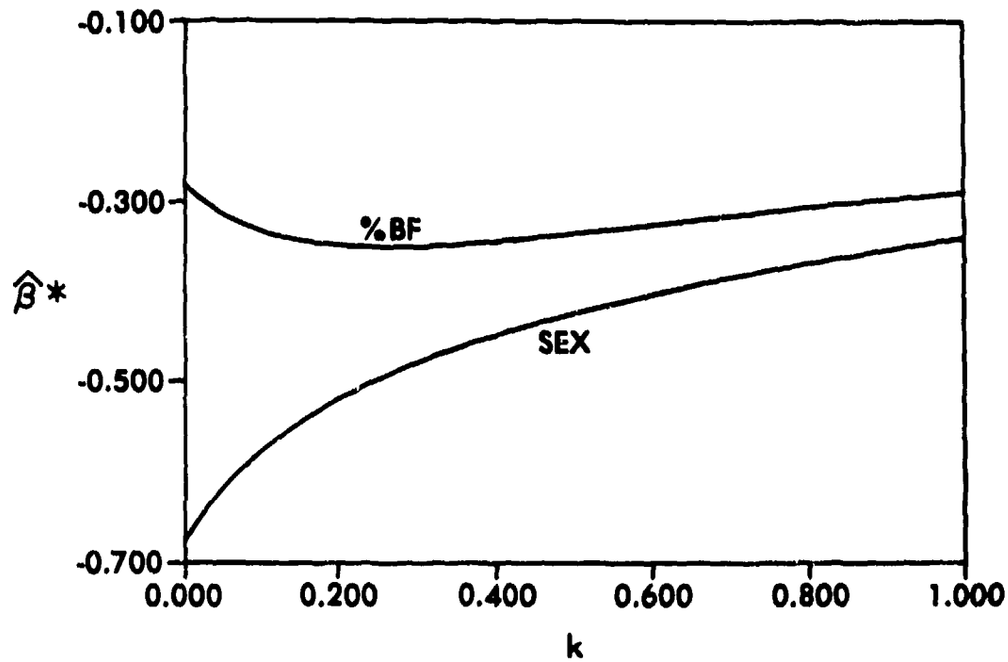


Figure 16 - Group 2 predictor data in two predictor Group 1 model for relative $\dot{V}O_2$ max. Variation of prediction standard deviation with bias.

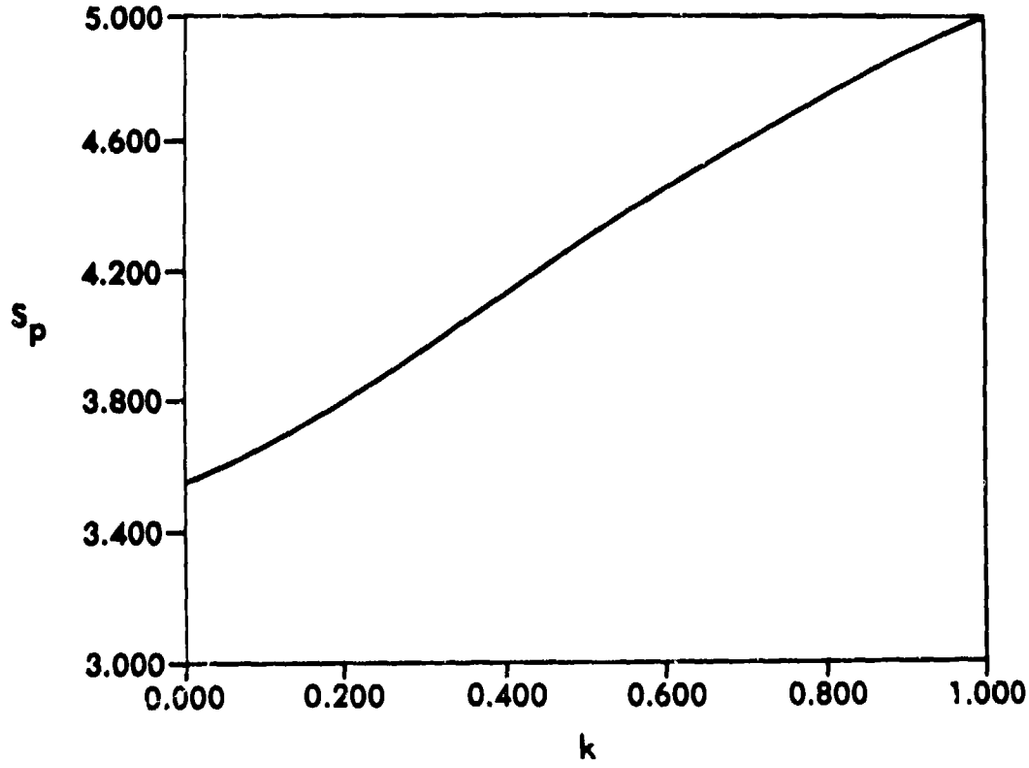


Figure 17 - Group 1 predictor data in two predictor Group 2 model for relative $\dot{V}O_2$ max. Variation of prediction standard deviation with bias.

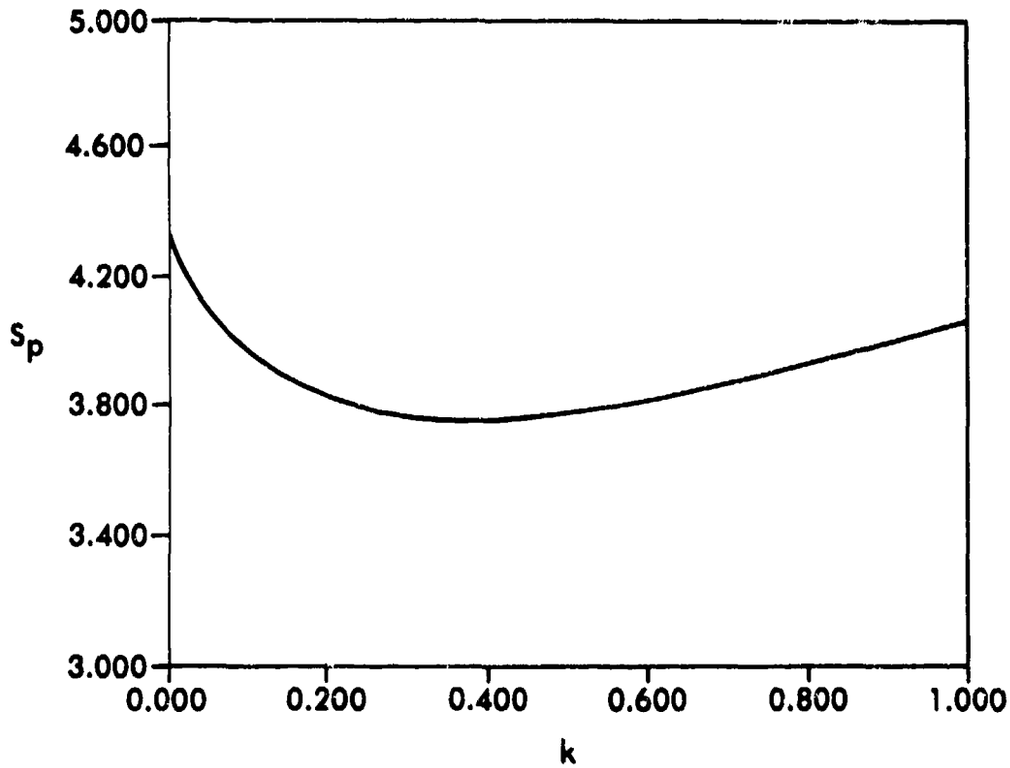


Figure 18 - Combined groups two predictor model for relative MOC.

Variation of standardized regression coefficients with bias.

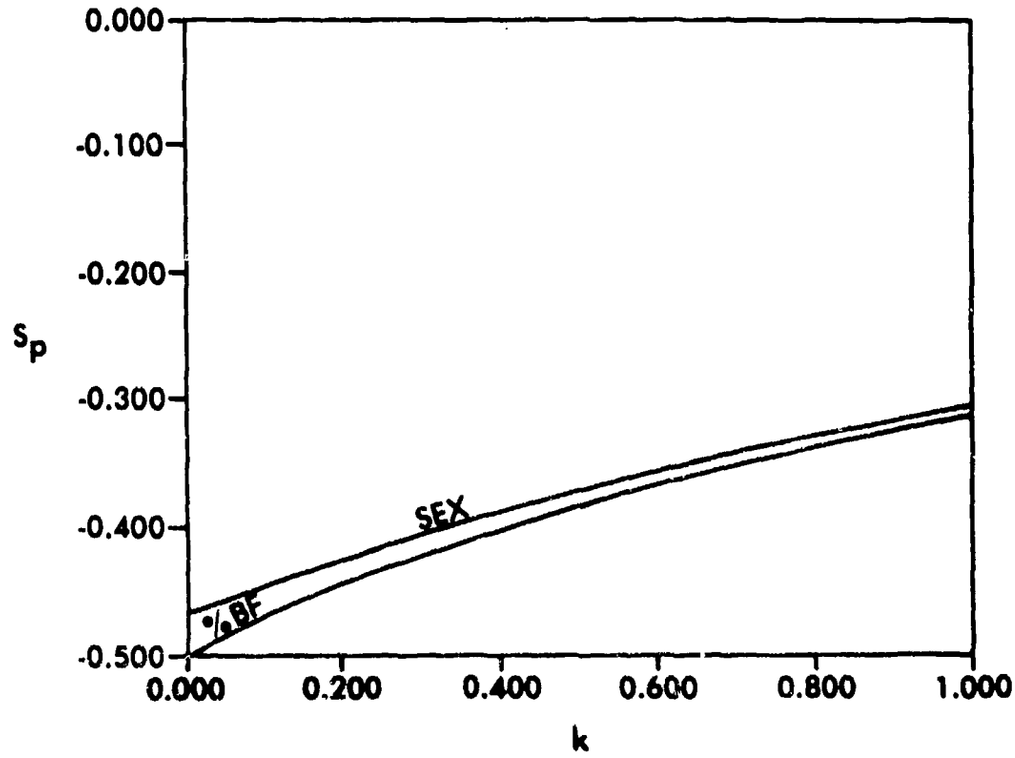


Table 17

Standardized regression coefficients and squared multiple correlation coefficients for two models and a combined groups model of $\dot{V}O_2$ max for two predictors.

model group:	1	2	combined		1	2	combined
k	0.0	0.3	0.0	estimator R^2	0.781	0.853	0.821
β weights:				new sample R^2	0.756	0.836	--
SEX	-0.384	-0.481	-0.467	predictor R^2	0.775	0.811	--
% BF	-0.586	-0.350	-0.502	n	48	47	95

Model of MLC

Analysis of Covariance

The data from the Fort Stewart Study are summarized in Table 7. The table depicts the sample characteristics of the two model groups for each sex. Table 8 summarizes the t values, and has been discussed above. The parallel and coincidental behavior of the Fort Stewart data for ML132 vs a number of predictor variables is summarized in Table 18. The t values for both tests of parallel behavior and coincidence are not significant at a level of 0.05 for intergroup comparisons within the same gender. Tests between sexes within the same group are presented in Table 19 and show consistent parallel behavior, but noncoincidence. These features support the utility of a single model for both genders with a gender designator as a constituent variable.

Table 18

Test of Fort Stewart data for parallel and coincidental behavior using t tests, and homogeneity of variance using F test. Comparisons are between groups for the same gender.

variable with ML132	Females					Males				
	n_1	n_2	t_p	t_c	F	n_1	n_2	t_p	t_c	F
LBM	21	22	0.15	1.07	1.15	92	90	0.74	0.15	1.29
LEG	21	22	1.07	0.06	1.07	92	91	0.98	0.16	1.26
TR	21	22	0.35	0.20	1.04	91	91	0.62	0.40	1.46
UT	21	22	0.72	0.07	1.43	92	91	1.29	0.23	1.12
HG	21	22	0.18	0.34	1.34	92	91	0.87	0.10	1.46
UP38	21	22	0.85	0.73	1.24	92	91	1.13	0.01	1.37
UP132	21	22	0.44	0.06	1.26	92	91	0.56	0.35	1.32

Table 19

Test of Fort Stewart data for parallel and coincidental behavior using t test, and homogeneity of variance using F test. Comparisons are between sexes in the same group.

variable with ML132	Group 1					Group 2				
	n_f	n_m	t_p	t_c	F	n_f	n_m	t_p	t_c	F
LBM	21	92	1.54	1.54	3.22**	22	90	1.60	3.97**	2.16*
LEG	21	92	0.53	7.90**	3.71**	22	91	0.77	8.88**	3.16**
TR	21	91	0.49	7.27**	4.42**	22	91	0.12	8.44**	2.91**
UT	21	92	1.36	3.04**	4.63**	22	91	0.15	4.10**	2.90**
HG	21	92	0.90	3.94**	4.85**	22	91	0.32	5.54**	2.49**
UP38	21	92	0.76	3.77*	4.26**	22	91	0.40	6.60**	3.86**
UP132	21	92	0.04	7.04**	4.69**	22	91	0.03	9.19**	2.80**

*significant at 0.05

**significant at 0.01

The summary of F tests for homogeneity of variance for the Fort Stewart data is presented in Tables 18 and 19. Comparisons between groups for the same sex support homogeneity of variance by consistently non-significant F values at the 0.05 level. Ten of the 14 F values are less than the F values at the 0.25 level lending strong support for the randomization procedure in sorting into groups. Comparison of the sexes within the same group reveal F values highly significant with 13 of the 14 F values significant at the 0.01 level. It is readily apparent that the variance of the residuals is significantly greater for the males in these two groups of data. This feature detracts from the use of a combined gender model using this set of data where confidence limits could be used in establishing predicted score cutoffs. Because of the low number of females in this sample, it is difficult to ascertain whether this heterogeneity in variance between sexes truly reflects the characteristics of the population as a whole.

An additional possibility is that the heterogeneity of variance represents a range effect. That is, "weaker" subjects show less variation than "stronger" subjects. This phenomenon is commonly seen in performance measurements possessing a closed bound on the low end of the scale and is unbounded on the high end. The observation that less variance is associated with the smaller number of women lends support to this interpretation. An opposite association would be expected if the heterogeneity effect were due simply to a disproportionate number of women. The issue could be addressed by testing additional females.

In spite of this defect in the sample data it was decided to pursue a combined gender model with a gender designator as a constituent variable. If in reality there is either a true difference in the variance characteristics between sexes or a range effect, then this model will have certain inherent defects. If it

is decided that a conservative approach is to be used in setting the predicted MLC standard then the subject will be expected to perform on the test battery with a higher score than the set cluster standard. The deficiency of the model would manifest itself by slightly increasing the number of false positives and slightly decreasing the number of false negatives for strong subjects (i.e., males). The defect would affect weaker subjects (i.e., females) by increasing the false negatives (i.e., a sizable number of women would be denied qualification for a cluster when they truly qualified), and decreasing the false positives.

If a "liberal" approach is used by setting the predictive MLC score below the true cluster standard, then the model defect would manifest itself differently. For stronger subjects the effect would be to slightly increase the number of false negatives and slightly decrease the number of false positives. For weaker subjects the effect would be to more markedly increase the number of false positives and decrease the number of false negatives.

In pursuing the "conservative" use of the model it could be construed that one is willing to live with a high degree of false negatives in order to minimize the false positive. The opposite effect is the case in the "liberal" approach to the use of the model. If the heterogeneity of variance is real, then the model developed for this sample data and used in the conservative mode could be accused of discriminating against weak subjects. In the liberal mode, however, the model would discriminate against strong subjects and give a selective advantage to weak subjects in meeting the true MLC cluster standard.

Determination of Predictive Model

Table 29 summarizes the intercorrelation matrix for predictor variables and the criterion measure for each group. All the correlations are significant at

the 0.01 level. The gender designator accounts for the significant amount of the variance with the criterion measure, ML132. Again, the model to be developed for ML132 will have its ratio of range to resolution determined to a large extent by the gender designator.

Table 20

Intercorrelation matrix for criterion and predictor variables for each group in the Fort Stewart data.

Group 1

$n = 112,\ n_f = 21,\ n_m = 91$

	LBM	UT	LEG	TR	HG	UP132	UP38	SEX	ML132
LBM	1.000								
UT	0.804	1.000							
LEG	0.490	0.434	1.000						
TR	0.573	0.664	0.330	1.000					
HG	0.821	0.798	0.427	0.578	1.000				
UP132	0.647	0.676	0.496	0.582	0.629	1.000			
UP38	0.808	0.798	0.599	0.701	0.768	0.750	1.000		
SEX	-0.745	-0.747	-0.484	-0.583	-0.705	-0.534	-0.729	1.000	
ML132	0.875	0.780	0.484	0.522	0.756	0.594	0.741	-0.695	1.000

Group 2

$n = 112,\ n_f = 22,\ n_m = 90$

	LBM	UT	LEG	TR	HG	UP132	UP38	SEX	ML132
LBM	1.000								
UT	0.779	1.000							
LEG	0.569	0.609	1.000						
TR	0.597	0.753	0.437	1.000					
HG	0.806	0.750	0.503	0.584	1.000				
UP132	0.569	0.603	0.479	0.557	0.593	1.000			
UP38	0.731	0.814	0.551	0.681	0.753	0.690	1.000		
SEX	-0.728	-0.790	-0.582	-0.549	-0.682	-0.484	-0.636	1.000	
ML132	0.878	0.771	0.519	0.602	0.799	0.566	0.758	-0.754	1.000

The results of the initial ridge regression analysis for the two groups are presented in Table 21 and Figures 19 and 20. Contrast of the first and last eigenvalues for groups 1 and 2 reveals approximately 50 fold differences for each. This suggests multicollinearity to be a significant problem in both groups. Inspection of Figures 19 and 20 show that three of the β weights are driven relatively more rapidly to zero than the others. Those are LEG, TR and UP132 for both groups. In keeping with the constraints mentioned previously, these three predictor variables were eliminated from the ridge repression problem, and the regression repeated with the reduced set.

Figure 19 - Group 1 model of MLC.

Variation of eight standardized regression coefficients with bias

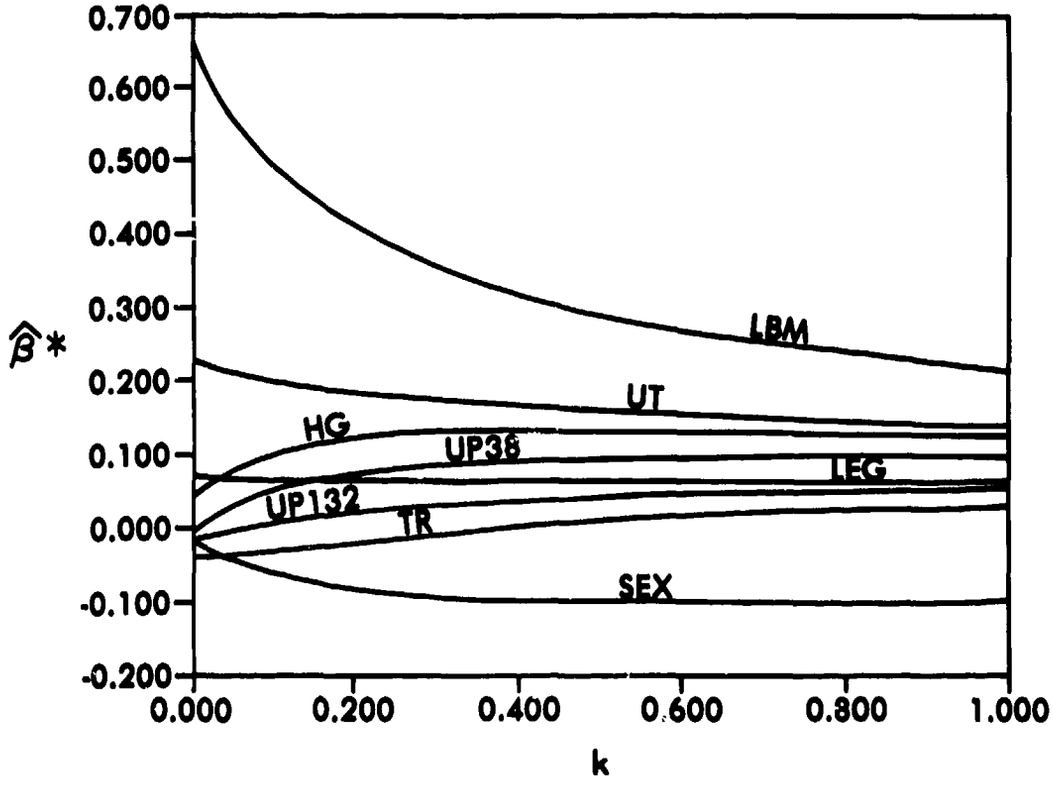


Figure 20 - Group 2 model of MLC.

Variation of eight standardized regression coefficients with bias

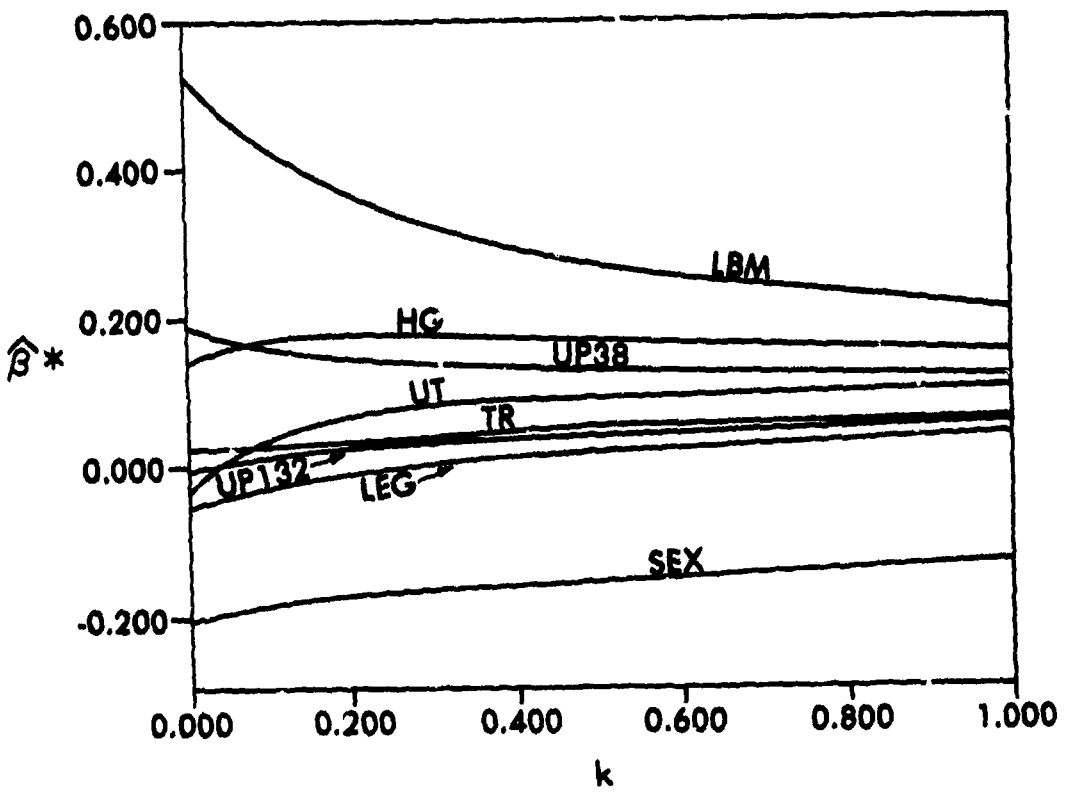


Table 21

Eigenvalues and unbiased standardized regression coefficients for the prediction of ML132 from LBM, UT, LEG, TR, HG, UP132, UP38, and SEX.

Group 1 model

variable	β weight	eigenvalue	degree
LBM	0.669	5.541	1
UT	0.224	0.729	2
LEG	0.068	0.540	3
TR	-0.040	0.427	4
HG	0.033	0.262	5
UP132	-0.019	0.185	6
UP38	-0.015	0.177	7
SEX	-0.018	0.140	8

Group 2 model

variable	β weight	eigenvalue	degree
LBM	0.538	5.503	1
UT	-0.049	0.635	2
LEG	-0.060	0.552	3
TR	0.023	0.467	4
HG	0.134	0.306	5
UP132	-0.010	0.234	6
UP38	0.192	0.183	7
SEX	-0.216	0.119	8

Figure 21 - Group 1 model of MLC.
Variation of five standardized regression coefficients with bias

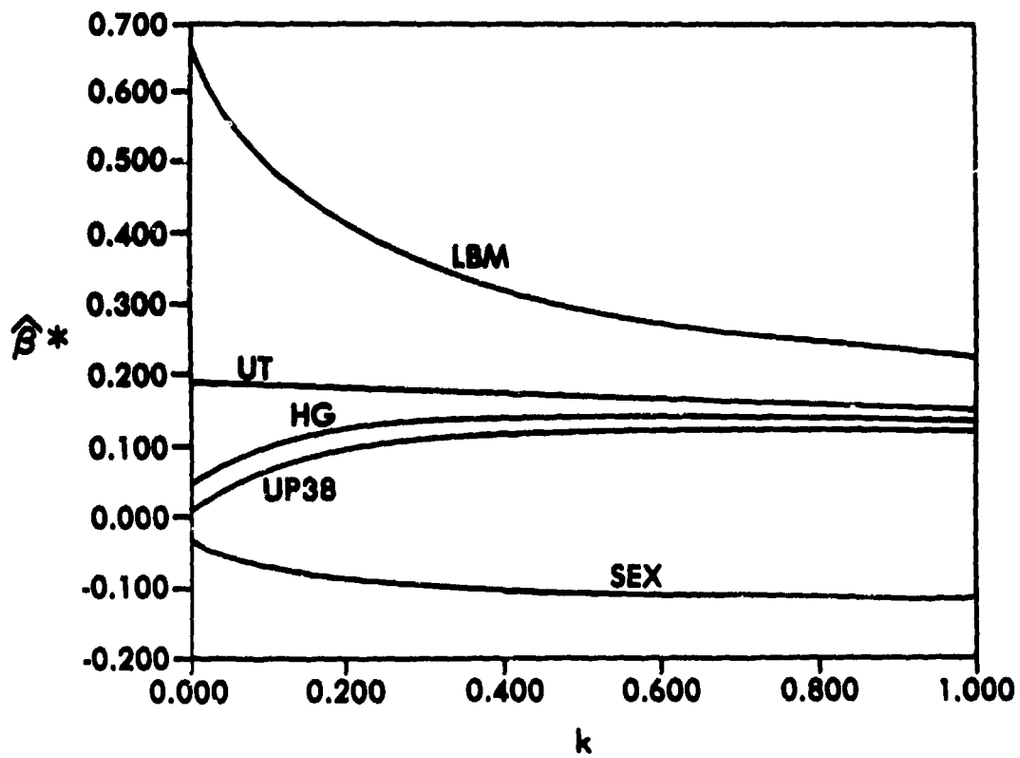


Figure 22 - Group 2 model of MLC.
Variation of three standardized regression coefficients with bias

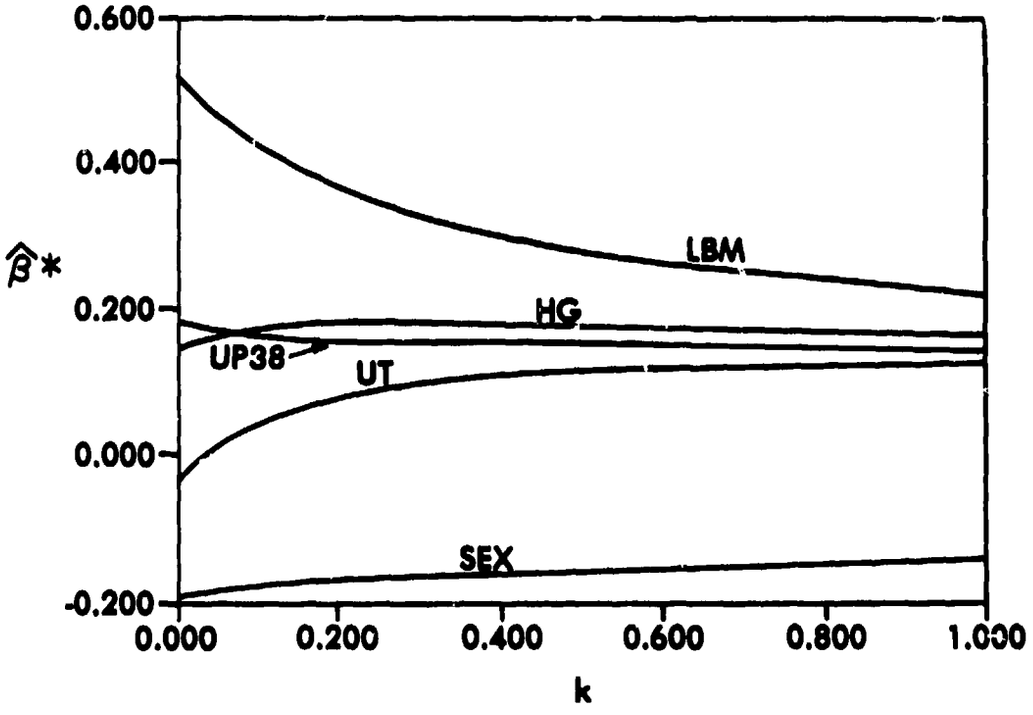


Table 22 and Figures 21 and 22 illustrate the results of the ridge regression for this reduced set of variables. Again, inspection of the first and last eigenvalues for each group suggests a significant multicollinearity problem. Inspection of Figure 21 suggests that the β weight for LBM is excessively high, and that the weights for HG and UP38 are underestimated. In fact the β weight for UP38 is driven from a slightly negative value to a more significant and realistic positive value. Inspection of Figure 22 for group 2 again suggests the β weight for LBM to be overestimated. Also, the weight for UT is driven from a negative value to a physically meaningful positive value.

Table 22

Eignvalues and unbiased standardized regression coefficients for the prediction of ML132 from LBM, UT, HG, UP38, and SEX.

Group 1 model

variable	β weight	eigenvalue	degree
LBM	0.674	4.092	1
UT	0.186	0.315	2
HG	0.033	0.228	3
UP38	-0.002	0.196	4
SEX	-0.034	0.168	5

Group 2 model

variable	β weight	eigenvalue	degree
LBM	0.526	3.990	1
UT	-0.047	0.379	2
HG	0.138	0.305	3
UP38	0.182	0.189	4
SEX	-0.199	0.136	5

The results of the ridge analyses of this reduced set of predictor variables suggest LBM to be the most significant predictor, gender to play a significant role, and the three isometric measures to be similar in importance to a predictive model. Because of the operational constraints of the AFEES it was decided to eliminate HG and UT as predictor variables and retain UP38. The basis for keeping UP38 rested mainly on its face validity and the simplicity of the measure. Little set-up is required of the subject and/or the device as compared to the other two variables. Retention of some measure of strength performance was deemed teleologically important enough in the prediction of strength capacity to justify its inclusion. The predictive model to be developed rests then on three variables - lean body mass, gender, and the 38 cm isometric upright pull.

Figure 23 - Group 1 model of MLC.

Variation of three standardized regression coefficients with bias.

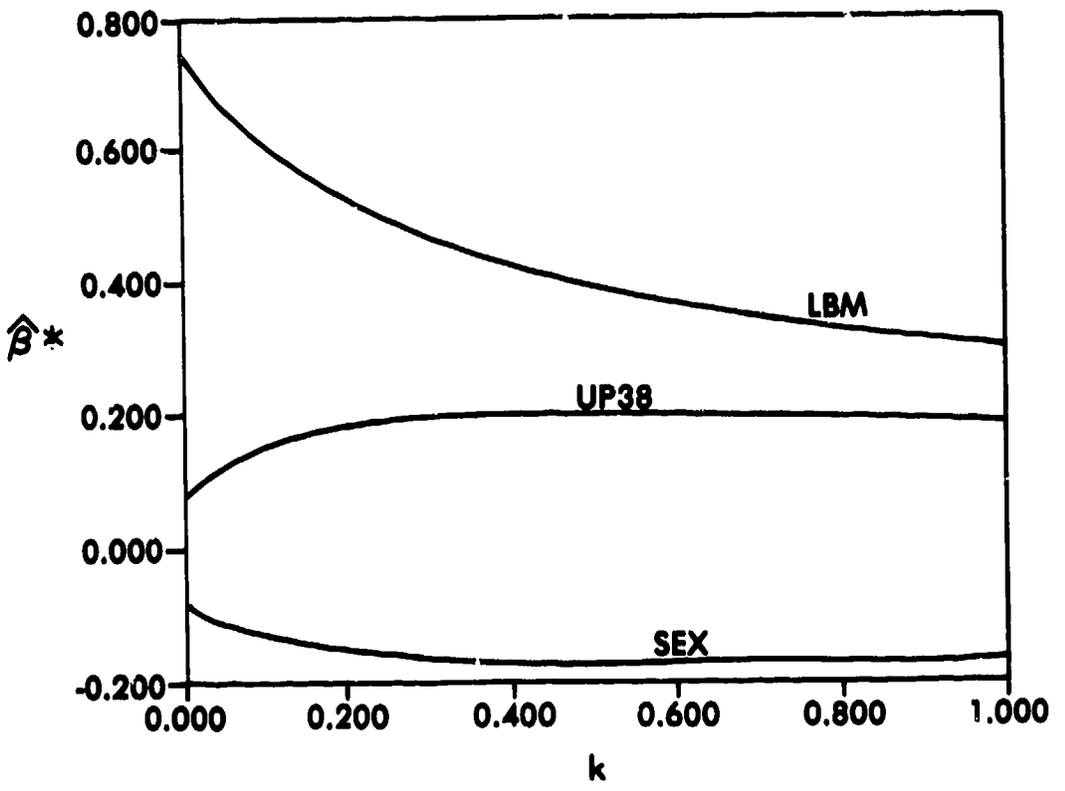


Figure 24 - Group 2 model of MLC.

Variation of three standardized regression coefficients with bias.

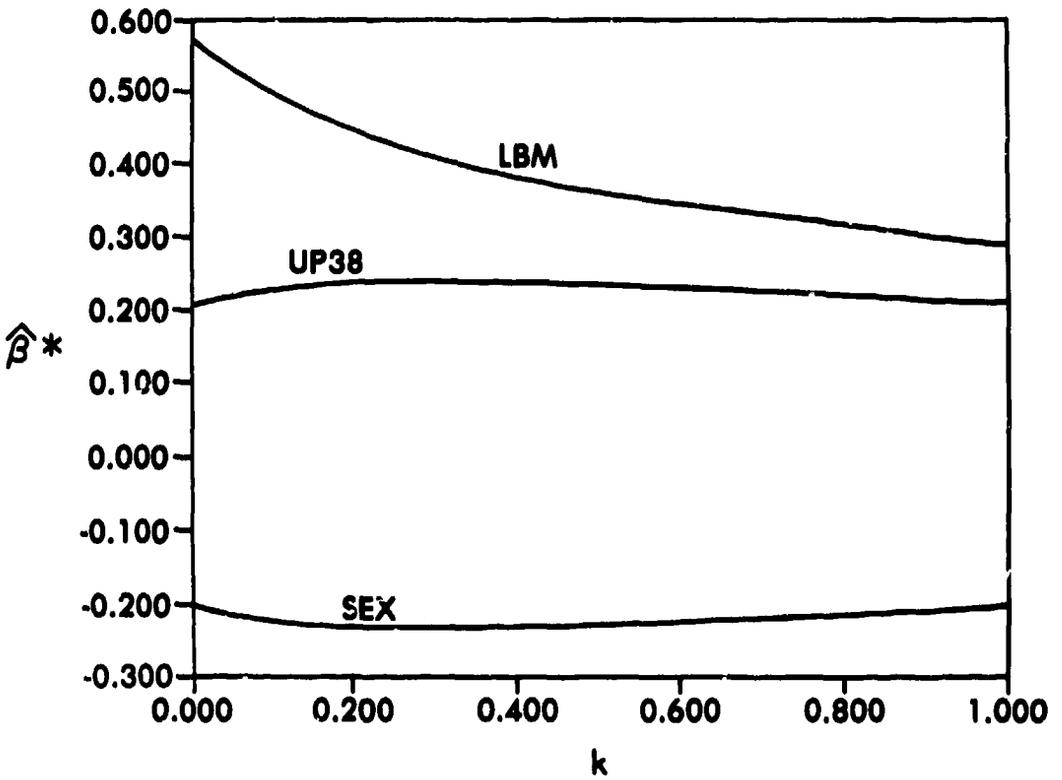


Table 23 and Figures 23 and 24 illustrate the results of the ridge regression analysis for this three predictor model. The first and third eigenvalues differ by approximately factors of 13 and 10 for groups 1 and 2 respectively. Figure 23 suggests a range of 0.2 to 0.4 for the bias coefficient in the group 1 data. A range of 0.0 to 0.2 is suggested by inspection of Figure 24 for group 2. Figures 25 and 26 depict the cross validation procedure. For the group 1 model using group 2 data a range 0.05 to 0.2 is suggested for the bias coefficient. The S_p vs k plot for the group 2 model depicted by Figure 26 suggests a value of k = 0.0.

Table 23

Eigenvalues and unbiased standardized regression coefficients for the predication of ML132 from LBM, UP38, and SEX

Group 1 model

variable	β weight	eigenvalue	degree
LBM	0.759	2.522	1
UP38	0.069	0.289	2
SEX	-0.081	0.190	3

Group 2 model

variable	β weight	eigenvalue	degree
LBM	0.583	2.397	1
UP38	0.205	0.364	2
SEX	-0.199	0.239	3

Figure 25 - Group 2 predictor data in three predictor Group 1 model of MLC. Variation of prediction standard deviation with bias.

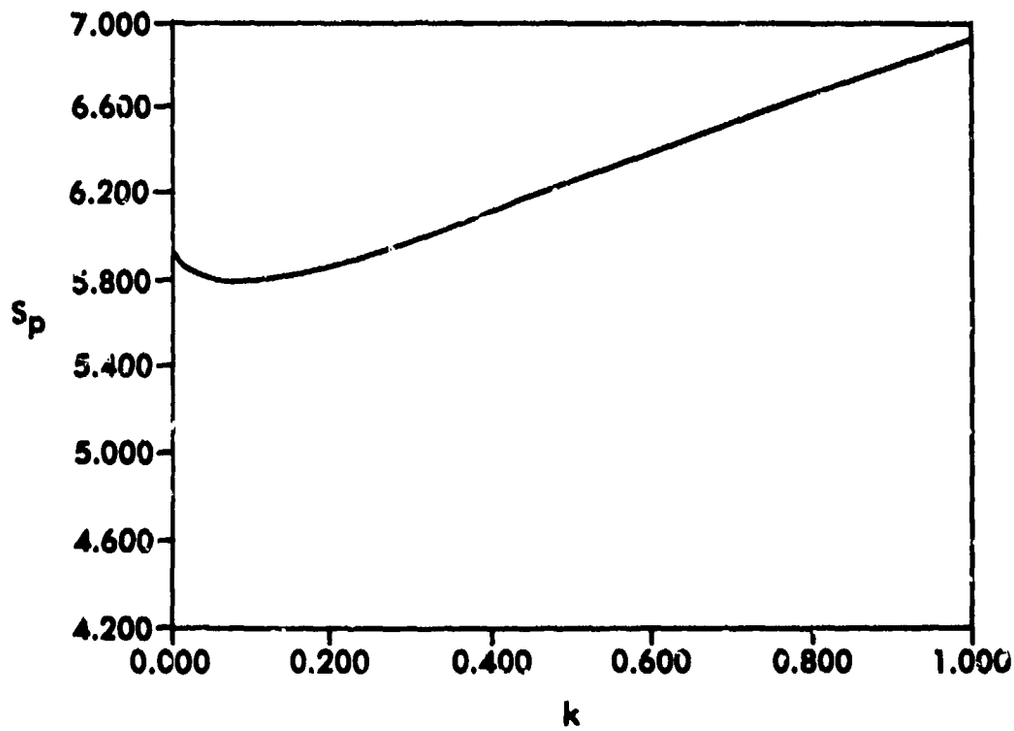
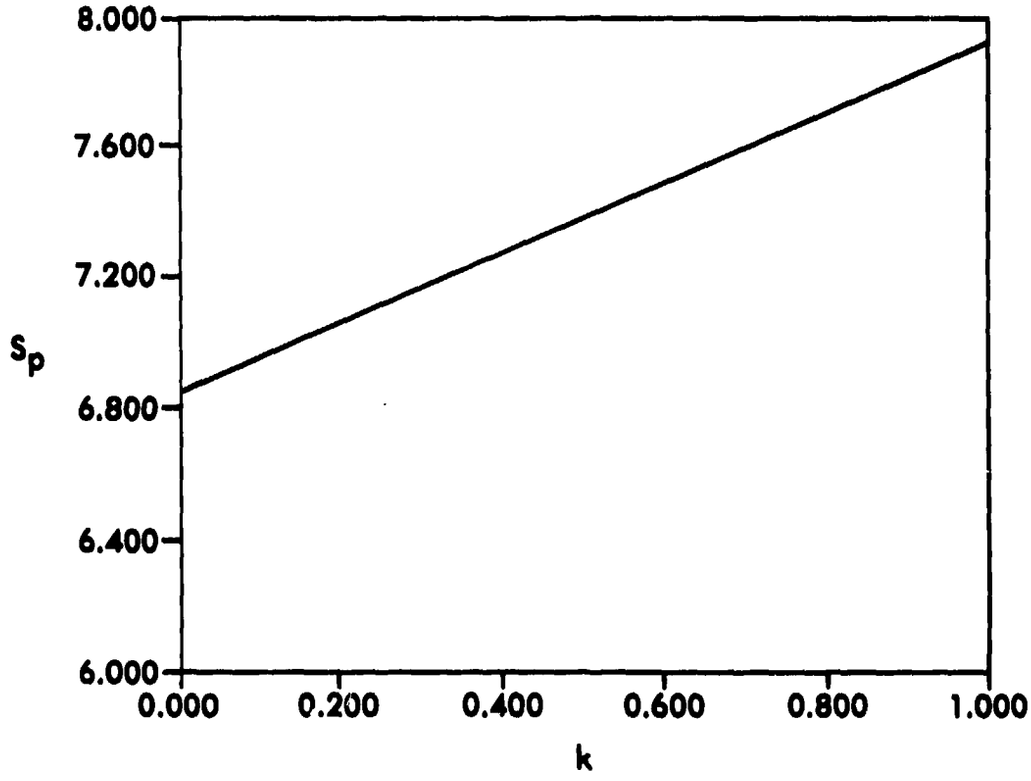


Figure 26 - Group 1 predictor data in three predictor Group 2 model of MLC. Variation of prediction standard deviation with bias.



As a result of these observations values of k= 0.2 and k = 0.0 were chosen for group 1 and 2 respectively. Table 24 depicts the standardized regression coefficients for the two groups for the chosen values of k. It is readily apparent that the β weights of group 2 are consistently greater in magnitude than those of group 1. However, the percentage of relative importance as calculated by the ratio of the square of the β weight to the sum of squares of the weights are quite comparable.

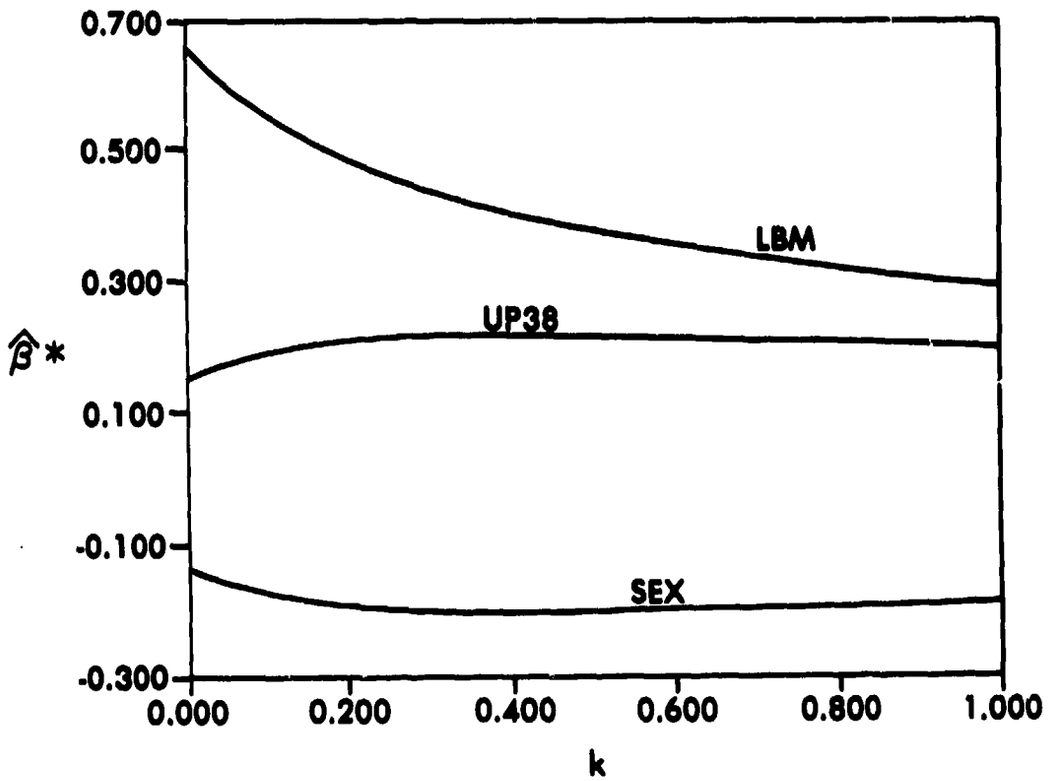
Also presented in Table 24 are squared correlations reflecting the estimator model R^2, the new sample R^2, and the cross validation R^2 for both groups. Although the cross validation R^2 for the group 2 model is less than the expected new sample R^2 the difference is not significant enough to detract from the model.

Table 24

Standardized regression coefficients and squared multiple correlation coefficients for two models of ML132.

model group:	1	2		1	2
k	0.2	0.0	estimator R^2	0.754	0.817
β weights:			new sample R^2	0.738	0.805
LBM	0.514	0.583	predictor R^2	0.804	0.760
UP38	0.180	0.205			
SEX	-0.152	-0.199			

Figure 27 - Combined groups three predictor model for MLC.
Variation of standardized regression coefficients with bias.



With these results the groups were combined to generate the final model. Table 25 and Figure 27 present the results of the ridge regression analysis. The first eigenvalue is about 10 times greater than the last, suggesting a possible multicollinearity problem. Inspection of the ridge plot of Figure 27 suggest fairly stable coefficients, however. Without any bias the β weights do not fall into the range suggested by the data in Table 24. A bias of k = 0.1 drives all the β weights within the range suggested by the separate groups. This bias was chosen in order to generate the final MLC model.

Table 25

Eigenvalues and standardized regression coefficients for a single combined groups model of ML132.

variable	$R^2 = 0.790$ β weight @ k = 0.0	$R^2 = 0.785$ β weight @ k = 0.1	eigenvalue	degree
LBM	0.664	0.546	2.456	1
UP38	0.145	0.191	0.324	2
SEX	-0.138	-0.175	0.220	3

Final Models for $\dot{V}O_2$ max and MLC

Table 26 presents the final model coefficients for raw score scaled data for both the prediction of relative $\dot{V}O_2$ max and the prediction of safe MLC to 132 cm. The standard error of the estimate is also presented.

Table 26

Raw score scaled coefficients, standard error of the estimate (SEE), and sample size for combined groups data for the prediction of ML132 in kg and relative $\dot{V}O_2$ max in ml/kg/min. (males = 1, females = 2 for SEX)

ML132 (kg)

SEE = 6.61 kg, $n = 225$, $n_f = 43$, $n_m = 182$

$$ML132 = -8.466 + 0.9933\ (LBM) + 0.006349\ (UP38) - 4.777\ (SEX)$$

$\dot{V}O_2$ max (ml/kg/min)

SEE = 3.49 ml/kg/min, $n = 95$, $n_f = 48$, $n_m = 47$

$$\dot{V}O_2\ max = 68.04 - 0.5725\ (\%\ BF) - 7.598\ (SEX)$$

Repetitive Lift and Carry Performance

The remaining issue to be addressed is the characterization of the lift and carry performance in terms of strength capacity and endurance capacity. Table 27 presents the results of a multiple regression analysis where the criterion measure is the number of repetitions over the ten minute period and constituent variables are ML132 and relative $\dot{V}O_2AR$. The highest correlations with the lift and carry performance at both weights are with ML132. All multiple R^2's are significant at the 0.01 level although moderately weak with the exception of the female 43 kg lift and carry performance with $R^2 = 0.640$. The addition of $\dot{V}O_2AR$ significantly increases the amount of variance accounted for by the regression model, although the increase is not large.

Table 27

Regression analysis for the prediction of lift and carry performance at two loads for each gender separately from ML132 and $\dot{V}O_2AR$ predictors.

43 kg over 10' for males (n = 182) and females (n = 42)

step	variable	males simple r	males multiple R	females simple r	females multiple R
1	ML132	0.335	0.335	0.602	0.602
2	$\dot{V}O_2AR$	0.129	0.357	0.173	0.640

25 kg over 10' for males (n = 182) and females (n = 42)

step	variable	males simple r	males multiple R	females simple r	females multiple R
1	ML132	0.322	0.322	0.306	0.306
2	$\dot{V}O_2AR$	0.153	0.353	0.036	0.312

This analysis confirms the importance of both a strength component and an endurance component in repetitive lift and carry peformance. Large or strong correlations cannot really be expected in this sample data for two reasons. The first is due to the sizable effect of motivation in the performance of the task. No reward system was utilized to enhance motivation. Less important is the use of an indirect and relatively imprecise measure of aerobic capacity as reflected in the step test. The strong correlation between lift and carry performance and ML132 for females at the 43 kg weight would suggest that strength capacity alone plays a much more significant role in women (or more objectively, "weak" subjects) than men for repetitive lifting of a relatively heavy external mass.

CONCLUSIONS AND SUMMARY

Two models have been developed to predict criterion measures reflecting aerobic and strength capacities. These models have been based on the

relationships between criterion measures having high face validity with real world Army physical performance requirments and simple measures of anthropometry and isometric strength performance. A statistical methodology has been used in developing these models with both teleological arguments and practical constraints playing roles in the choice of predictor variables.

The choice of relative maximal oxygen consumption ($\dot{V}O_2$ max) as the criterion variable reflecting aerobic capacity is based on well understood physiological principles. Using ridge regression techniques and a two group cross validation procedure a model for relative $\dot{V}O_2$ max was developed using a gender designator and percent body fat calculated by the sum of four skinfolds as predictors. This model was developed on a sample of 47 male and 48 female recruits from the Fort Jackson Basic Training Center. This sample and its distribution characteristics can be considered to reflect the population characteristics of recruits although no overt randomization procedure was pursued. The model would be strengthened both in terms of its use probabilistically and its distribution characterisitics by an increase in the sample size - probably in the range of 300 to 400 subjects. If the model in its present form were used over a period of four years, over one million U.S. Army inductees would be screened. The use of the model and its distribution characteristics to initially describe physical performance characteristics of the recruit population would be strengthened by an increase in sample size.

The effect of an eight-week basic training program was demonstrated to be significant in increasing the sample's $\dot{V}O_2$ max on an absolute basis (i.e., liters O_2/minute). However, although statistically significant, the improvement was small enough to be impractical in incorporating this training effect into a model used for individual screening.

The criterion measure of strength capacity was chosen to be the safe maximum lifting capacity (MLC) to a 132 cm platform representing the bed of a cargo truck. An administrative survey of job tasks by experienced personnel representing the diverse military occupational specialities of the Army revealed that in excess of 90% of job tasks having sizable strength requirements had lifting and/or repetitive lift and carrying solely as the strength demanding task. This observation greatly simplified the development of a conclusive criterion of strength capacity applicable to the military occupational environment.

Using the same statistical methodology as for the aerobic capacity model, a model of safe MLC was developed using a gender designator, an estimation of lean body mass, and performance on an isometric strength measure of upright pull at 38 cm. This model was developed from a sample of 182 males and 43 females at Fort Stewart, GA. These subjects were not enlistees, but were experienced military personnel. The subjects cannot be considered representative of enlistees in terms of their distributional characteristics. Similarly the small number of females in the sample is a weakness. In spite of the demonstration of consistent and significant differences in the residual variances between males and females of this sample data for regressions of MLC vs single predictor variables, a combined gender model was developed. The limitations in using this model as a screening device in a probabilistic manner were discussed. Use of the model in this manner could be misleading and may give selective advantages to either sex depending on its mode of use. The functional characteristics of the model can be applied to the enlistee population even though the model was not developed from that population. Less certain is the use of the model in a probabilistic manner in determining the predictive score cutoff for a cluster standard. Finally, the use of the sample in describing

the inductee population characteristics in terms of the criterion measures for purposes of manpower description and allocation is unjustified.

The methodology for both setting cluster standards and sorting MOS's into a cluster is discussed. For aerobic demanding tasks both the rate of energy expenditure and the duration of the task are factors in determining the aerobic demands of the task. Both of these factors can be accounted for in setting an aerobic cluster standard in terms of relative $\dot{V}O_2$ max. For strength demanding tasks both the absolute load lifted and repetition are factors in setting the strength cluster standard. It was demonstrated by Poulsen[23] that nothing is gained by having subjects repetitively lifting loads greater than 50% of their MLC in terms of work output. This information along with an accounting of injury risk and establishing "acceptable" rates of injury could be used in both setting the strength cluster standard and sorting job tasks into clusters.

It has been the purpose of this report to show the processes and methods chosen to develop a practical system to screen U.S. military enlistees for physically demanding MOS's. It should be readily apparent that the factors considered important for effective physical performance in the U.S. Army may not apply to civilian industry, or even other military services. In developing this system it has been necessary to focus on a number of critical issues involving work performance that are difficult to identify let alone quantify. The issue of what actually constitutes effective performance must be addressed. This task alone can be fraught with discord. Developing objective measures of performance and capacity, being able to test for these measures either directly or indirectly, and describing manpower distribution characteritics in terms of these measures is another awesome undertaking. The development of cost/risk/benefit standards and the effec. these will have in the efficient

operation of the enterprise are issues that can be particularly problematic. Physical capacity addresses only one aspect of effective job performance. It would be unwarranted to think that addressing this single aspect would resolve the larger issue of adequate job performance in the Army. The methods and factors discussed in this presentation offer the mechanisms by which some of these issues can at least be initially addressed.

Weaknesses in the sample data from which these models of physical capacity are developed limit the utility of the models for the purpose of describing the enlistee distributon characteristics in terms of the criterion measures. Use of the models probabilistically is weakened by the relatively low number of subjects, disproportionate number of females, and/or inappropriate sample population. A strong use of the models would be the description of physical capacity characteristics of the enlistee population, as defined by the criterion measures, and the use of this information to vary cluster standards. It would be inappropriate to utilize the models developed from these samples for this specific purpose.

The aforementioned limitations and weaknesses, however, may be relatively unimportant from the view of practicality. These limitations refer only to the use of the criterion measures as the mediators of effective physical occupational performance. It should be recalled that these criterion measures are in reality only simulators of the true physical performance requirements. Since they have been accepted as such, and it has been demonstrated that the predictor measures of anthropometry and isometric strength performance relate strongly to these criteria, it would be sufficient to deal solely with the predictor variables using manpower needs, injury rates, etc. to dynamically set cluster standards periodically. It would be exceedingly important to develop a mechanism by

which to monitor manpower distribution, injury rates, and any other variable deemed operationally relevant in affecting physical performance, and thereby provide the feedback necessary to vary cluster standards. Such flexibility would insure that the screening process would be responsive to changing needs and effects.

References

1. Linroth, K. (1976) Int. Rev. of the Ar., Navy., Air For. Med. Ser., 49., 405-416.

2. Tornvall, G. (1963) Acta Physiol. Scand., 58: Suppl. 201, 1-102.

3. Nordesjo, L.-O. & Schele, R. (1974) Forsvarsmedicin, 10, 1-14.

4. Gorbunov, V.V. (1978) Entsiklopediya Znachkisto GTO, Fizikul'tura i Sport, Moscow.

5. Riordan, J. (1977) Sport in Soviet Society, Cambridge University Press, New York, pp. 207-287.

6. Durnin, J.V.G.A. & Passmore, R. (1967) Energy, Work, and Leisure, Heinemann Educational Books Ltd., London.

7. Larson, L.A. (Ed.) (1974), Fitness, Health, and Work Capacity: International Standards for Assessment, MacMillan, New York.

8. U.S. Air Force Manual 39-1. Airman Classification Manual (1969).

9. Hermansen, L. & Saltin, B. (1969) J. Appl. Physiol., 26, 31-37.

10. Astrand, P.-O. & Saltin, B. (1961) J. Appl. Physiol., 16, 977-981.

11. Varnpliktsverket InsE (1977), Enlistment of Conscripts, Enrollment Board of the Armed Forces, Sweden.

12. Ahlborg, B. (1980) personal communication.

13. Nordesjo, L.-O. (1974) Upsala J. Med. Sci., 79, 45-50.

14. Astrand, P.-O. & Ryhming, I. (1954) J. Appl. Physiol., 7, 218-221.

15. Shephard, R.J. (1970) J. Sports Med.. 10, 206-210.

16. Margaria, R., Aghemo, P., & Rovelli, E. (1965) J. Appl. Physiol., 20, 1070-1073.

17. Astrand, I. (1960) Acta Physiol. Scand., 49: Suppl. 169, 1-92.

18. Buskirk, E. & Taylor, H.L. (1957) J. Appl. Apysiol., 11, 72-78.

19. Haisman, M.F. (1970) Human Biol., 42, 679-688.

20. Durnin, J.V.G.A. & Wormersley, J. (1974) Br. J. Nutr., 32, 77-97.

21. Durnin, J.V.G.A. & Rahaman, M.M. (1967) Br. J. Nutr., 21, 681-689.

22. Wormersley, J. & Durnin, J.V.G.A. (1977) Br. J. Nutr., 38, 271-284.

23. Poulsen, E. (1970) Comm. Dan. Nat. Ass. for Infant. Paral., 31, 3-15.

24. Jorgensen, K. & Poulsen, E. (1974) Ergonomics, 17, 31-39.

25. Poulsen, E. & Jorgensen, K. (1971) Appl. Ergonomics., 2, 133-137.

26. Rasch, P.J. & Pierson, W.R. (1963) Ergonomics, 6, 211-215.

27. Chaffin, D.B. & Park, K.S. (1973) Amer. Ind. Hyg. Assoc. J., 31, 511-525.

28. Chaffin, D.B., Herrin, G.D., & Keyserling, W.M. (1978) J. Occup. Med., 20, 403-408.

29. Drinkwater. B.L. (1973) Exercise and Sports Sciences Reviews, Wilmore, J.H., Ed., Academic Press, New York, pp. 126-153.

30. Forbes, G.B. & Amirhakimi, G.H. (1970) Human Biol., 42, 401.

31. Bink, B. (1962) Ergonomics, 5, 25-28.

32. Gleser, M.A. & Vogel, J.A. (1973) J. Appl. Physiol., 34, 438-442.

33. Grosse-Lordemann, H. & Muller, E.A. (1936) Arbeitsphysiologie, 9, 454-475.

34. Muller, E.A. (1937) Arbeitsphysiologie, 10, 67-73.

35. Taylor, H.L., Buskirk, E. & Henschel, A. (1955) J. Appl. Physiol., 8, 73-80.

36. Mitchell, J.H., Sproule, B.J. & Chapman, C.B. (1957) J. Clin. Invest., 37, 538-547.

37. Knapik, J., Kowal, D., Riley, P., Wright, J. & Sacco, M. (1979) U.S. Army Research Institute of Environmental Medicine, Natick, MA, Report No. T 2/79.

38. Armitage, P. (1971) Statistical Methods in Medical Research, Wiley, New York.

39. Snedecor, G.W. & Cochran, W.G. (1967) Statistical Methods, 6th ed., Iowa State University Press, Ames, Iowa.

40. Ehrenberg, A.S.C. (1968) J. Roy. Statist. Soc., A, 131, 280-302.

41. Hoerl, A.E. & Kennard, R.W. (1970) Technometrics, 12, 55-67.

42. Hoerl, A.E. & Kennard, R.W. (1970) Technometrics, 12, 69-82.

43. Freunde, R.J. & Minton, P.D. (1979) Regression Methods, Marcel Dekker, New York.

44. Marquardt, D.W. & Snee, R.D. (1975) The Amer. Statist., 29., 3-20.

45. Tatsuoka, M.M. (1976) Selected Topics in Advanced Statistics: An Elementary Approach, No. 5, Institute for Personality and Ability Testing, Champaign, Illinois.

46. Arkin, H. & Colton, R.R. (1963) Tables for Statisticians, Barnes and Noble, New York.

DISTRIBUTION LIST

5 copies to:

US Army Medical Research and Development Command
Fort Detrick
Frederick, MD 21701

12 copies to:

Defense Technical Information Center
ATTN: DTIC-DDA
Alexandria, VA 22314

1 copy to:

Superintendent
Academy of Health Sciences, US Army
ATTN: AHS-COM
Fort Sam Houston, TX 78234

1 copy to:

Dir of Biol & Med Sciences Div
Office of Naval Research
800 N. Quincy Street
Arlington, VA 22217

1 copy to:

CO, Naval Medical R&D Command
National Naval Medical Center
Bethesda, MD 20014

1 copy to:

HQ AFMSC/SGPA
Brooks AFB, TX 78235

1 copy to:

Director of Defense Research and Engineering
ATTN: Assistant Director (Environmental and Life Sciences)
Washington, DC 20301