

LEVEL 4

①
mc

⑨

⑬ ARI-RM-

Research Memorandum 75-8

AD A 076786

⑥

**DEVELOPMENT OF PERFORMANCE TESTS AS
SUPPLEMENTARY ENLISTMENT SCREENING
MEASURES: AN INTERIM REPORT.**

⑩

Leonard C. Seeley and M. A. Fischl

DDC
NOV 15 1975
REGULATED
E

⑮ 29163101A733

MANPOWER DEVELOPMENT AND UTILIZATION TECHNICAL AREA

DDC FILE COPY

⑫ 14



This document has been approved for public release and sale; its distribution is unlimited.

U. S. Army

Research Institute for the Behavioral and Social Sciences

⑪
July 1975
79 11 13 383
408010

B

DISPOSITION FORM

For use of this form, see AR 340-15, the proponent agency is TAGCEN.

REFERENCE OR OFFICE SYMBOL

PERI-TP

SUBJECT

Clearance and Transmittal of Reports to DTIC

TO DDC-DAA-1
ATTN: Mr. Schrecengost

FROM ARI Rsch Pub Group
*for the Behavioral +
See 5c*

DATE 8 Nov 79
Ms Price/48913

CMT 1

1. The reports listed on Inclosure 1 are approved for public release with unlimited distribution (50 numbered ARI Research Memorandums, 74-1 thru 76-30).

2. These are among the previously unrecorded ARI reports which you identified to us 22 June 1979 as not in your retrieval system. The accompanying box contains at least one copy of each report for your retention and reproduction.

1 incl
List of reports, 1974-76

Helen S. Price

HELEN S. PRICE
Research Publications Group
Army Research Institute

x 4

Accession For	
ERIC	<input checked="" type="checkbox"/>
ERIC INS	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or special
A	

Army Project Number
2Q163101A733

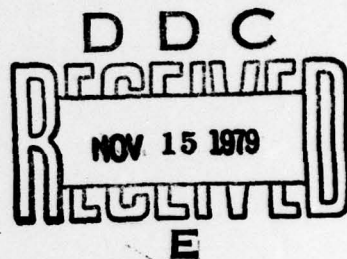
Enlisted Personnel
Recruitment and Classification

①

Research Memorandum 75-8

DEVELOPMENT OF PERFORMANCE TESTS AS
SUPPLEMENTARY ENLISTMENT SCREENING
MEASURES: AN INTERIM REPORT

Leonard C. Seeley and M. A. Fischl



Submitted by:
Ralph R. Canter, Chief
MANPOWER DEVELOPMENT AND UTILIZATION TECHNICAL AREA

July 1975

Approved by:

E. Ralph Dusek, Director
Individual Training and Performance
Research Laboratory

J. E. Uhlener, Technical Director
U. S. Army Research Institute for
the Behavioral and Social Sciences

Research Memorandums are informal reports on technical research problems. Limited distribution is made, primarily to personnel engaged in research for the Army Research Institute.

This document has been approved
for public release and sale; its
distribution is unlimited.

DEVELOPMENT OF PERFORMANCE TESTS AS SUPPLEMENTARY ENLISTMENT SCREENING MEASURES: AN INTERIM REPORT

BACKGROUND

Since the inception of personnel research programs in the Army and the other military services, research results have consistently revealed that marginal personnel, as defined by scores in the Category IV range on military aptitude measures, produce more than their share of disciplinary problems and unsatisfactory performance in training and on the job. Hence a persistent research problem has been to find ways to screen out as many poor risks as possible and to accept all potentially satisfactory soldiers..

Accordingly, the Army Research Institute (ARI) has had the task of developing and validating measures designed to supplement the current battery of operational screening instruments for use with Category IV or other score ranges as desired. Because the present operational measures are reasonably effective in covering the major intellectual areas, this task objective was to concentrate on the noncognitive, non-intellectual areas, involving motivational, attitudinal, and personality variables insofar as practicable.

Many marginal scorers on the aptitude test batteries are school dropouts without substantive academic achievements and possibly with little fondness for paper and pencil work. For this reason it was judged that tests and measures developed in this project should not resemble school-type measures. Instead, the development effort should be directed to performance tasks calling upon physical activity, where the intellectually limited might have more interest and feel more confident. Further, the effort should develop feasible situations which would generate overt behavior from which an observer could infer (and quantify) the noncognitive, attitudinal, and motivational characteristics of the examinee.

DATA COLLECTION

The research described in this report was conducted at the Reception Station at Fort Dix, New Jersey. While it was recognized at the outset that any such screening measures developed for operational use would most probably be administered at the Armed Forces Examining and Entrance Stations (AFEES) with other currently employed selection and classification devices, several considerations indicated the desirability of beginning with a reception station: adequate numbers of examinees would be available for experimental testing and could be more readily followed through training, physical space and materiel would be more easily obtained, and proximity to ARI would result in time and cost savings. Further refinement and standardization would be undertaken later at a suitable AFEES.

Data collection for the measures described in this report took place between the first part of September 1973, just after TRADOC 635-1 was implemented to facilitate early release of unsatisfactory recruits, and the end of February 1974. Before this period several months had been devoted to preliminary development and modification of a variety of items, which led to those reported here.

The tests were administered to new recruits during their reception station processing, in space provided by the reception station. This included office space plus four separate private testing rooms. Test administration and follow-up criterion data collection were handled by a team of enlisted men (E-3 through E-7) selected from Fort Dix personnel, who were trained by ARI scientists and were supervised throughout this period by a contractor's representative with periodic visits by ARI research personnel.

THE SAMPLE TESTED

A total of 225 Category IV men and 88 low Category III men--those scoring below 50 on the Armed Forces Qualification Test (AFQT)--were tested and included in the analysis. The Category III men were used to complete test groups when not enough Category IV men were available for a group test. Men tested but subsequently discharged for medical reasons before completing Basic Combat Training (BCT) were excluded from the sample.

PERFORMANCE VARIABLES

Development of the performance items described below was preceded by a certain amount of trial and error effort. For each item reported on, one or more additional items were invented at Fort Dix, tried out briefly, found wanting in some respect, and dropped from further consideration. During this early research it also became apparent that some motivational and personality characteristics showed more promise than others of being tapped by suitable performance items. Thus three general non-intellectual areas emerged within which experimental items were produced. These could be called dependability, cooperativeness, and willingness to push oneself physically.

In the items developed, which are described in detail in the next section, three aspects of dependability were tapped.

1. Willingness to remain on the job, and to work conscientiously, in the absence of direct supervision. This was measured by an Ammunition Sorting item and a Bottle Cap Sorting item.

2. Punctuality. The ability to keep an appointment on time, despite minor distractions, was measured by the Punctuality item.

3. Observational thoroughness. The examinee's ability to study a military scene briefly and render a complete and accurate account of his observations was measured by the Sand Table item.

In the case of the second non-intellectual behavioral characteristic for which test items were constructed, cooperativeness, two aspects were measured: individual helpfulness, and participation in a group task.

1. Individual helpfulness was assessed by the Chairs item, in which an individual is given an opportunity to help a soldier, whom he does not know, do light physical work for a few minutes.

2. Group participation was assessed by the Layout item, which provided a group task in which individuals could help other members of the group to a greater or lesser degree or not at all, and by the Forgotten test, which offered the examinee an opportunity to avoid further testing by leaving the group.

The third behavioral area for which a performance test was developed, willingness to push one's self physically, was assessed by the Exercise item, in which the examinee was told to do a particular exercise as many times as he could, and when he stopped was asked to do still more. This item was scored both on total number done and on willingness to do more.

DESCRIPTION OF THE TEST ITEMS

In the following individual descriptions, the sequence in which the items are presented to an examinee is not necessarily fixed, except in the case of the Forgotten Test.

1. Ammunition Sorting

Examinee is seated in front of eight empty ammunition cans and one large can containing 320 expended shells, 40 of each of eight sizes. He is told by the examiner to sort the shells and that the examiner must leave but will return in less than half an hour. The examiner leaves and closes the door, leaving the examinee completely alone; he returns in 10 minutes. Score is number of shells properly sorted.

2. Bottle Cap Sorting

Examinee is seated in front of a large carton of bottle caps with seven smaller empty boxes. He is told by the examiner to sort the caps and that the examiner will return in less than half an hour. Examiner leaves and closes door, leaving examinee alone; he returns in 10 minutes. Score is quantity of caps properly sorted, measured by volume.

3. Punctuality

Examinee is sent to snack bar in nearby building, where various distractions exist, and is told to report back at a specified time, usually 15 to 30 minutes later. Examinee is scored dichotomously on whether he returns on time or late.

4. Sand Table

Concealed from examinee's view is a 4' x 4' table set up to display a military field situation (tanks, trucks, artillery, etc., in broken terrain). Examinee is told to act as a scout, to observe the situation for a specified brief time (45 seconds), and report back all the information he can. He does this; then examiner makes specific changes in the situation and examinee returns to observe for 30 seconds and report on the changes. Score is total of items reported on both observations.

5. Chairs

Examinee, presenting himself for test, is told to wait in adjacent room until called. Staff member, dressed as recruit, is busy in the room moving 30 or more chairs from stack in the corner into classroom arrangement. Examinee is awarded 0-3 points depending on whether, and how soon, he offers to help.

6. Layout

This test is given to groups of four. Each man is provided with a bag containing the equipment of a full field layout. Contents of another bag are already laid out. Examinees are shown the display briefly and told they should make four displays, all alike but not necessarily like the demonstration display. They are encouraged to work together. Each man is scored for the number of times he helps with another man's display either physically or by giving oral advice.

7. The Forgotten Test

This test is also given to groups of four after all other items are completed. As examiner is preparing to dismiss group, another staff member interrupts, announcing one more test. Offers to excuse those who want to drop out. Each man is scored on whether or not he is willing to stay for another test.

8. Exercise Test

Examiner demonstrates leg-up exercise (lie on back on mat, knees stiff; raise legs to vertical position, forming 90-degree angle with torso, then return legs to mat). Examinee is told to do as many leg-ups

as he can, and report when he can do no more. When he so reports, examiner says "Try to do 5 more." If he does 5 more, examiner says "Try one more." Two scores are obtained: (1) Exercise Total, the number of leg-ups done, which is an indication of physical condition; and (2) Exercise Effort, the willingness to try or not try to do "one more" (the attempt does not need to be completed in order to be credited).

TIME REQUIREMENTS FOR THE TESTS

Time required for administration of this battery of performance test items ranges from one-and-one-half to two hours. During the data collection at Fort Dix the usual procedure was to test eight men during a two hour period, with all but the Layout and Forgotten Test items being done individually. Approximate times for each measure are as follows:

Chairs	5 minutes
Forgotten Test	5 minutes
Sand Table	6 minutes
Exercise	6 minutes
Ammunition Sorting	12 minutes
Bottle Cap Sorting	12 minutes
Punctuality	20 minutes
Layout	30 minutes

THE PERFORMANCE CRITERION

Since the Army is concerned with enlisting men who will become good soldiers, the ideal criterion against which to judge the validity of the performance items would be an objective measure of the man's performance in BCT, AIT, and on all his subsequent job assignments. Practical considerations of time and costs, however, required limiting the criterion in the present study to the most immediately available, a measure of BCT performance.

The criterion used was a combination of the existing Enlisted Efficiency Report (EER), any action taken to separate the trainee under TRADOC 635-1, and a simple performance rating form devised as a supplement. The EER is a three-way letter grade evaluation routinely recorded for all Basic Trainees: A for outstanding, B for average, C for poor. The rating form consisted of an overall appraisal on a 5-point scale; it was completed by the drill sergeant best acquainted with the trainee and filled out under the supervision of a member of the ARI staff, usually the day before BCT graduation.

From these evaluations three criterion groups were formed: (1) A High Performance group consisted of those men rated A on the EER or "good"

or "excellent" by the drill sergeant; (2) a Low Performance group consisted of those men given an early release under TRADOC 635-1, or rated C on the EER or "fair" or "poor" by the drill sergeant; and (3) the Average Performance group were those who were neither in the High nor the Low group and who were rated "average" by the drill sergeant.

ANALYSES

One way to examine the data produced in this study is to compare the mean score attained on each predictor item by the High Performance group with the mean score attained on that item by the Low Performance group. This was done, and the size and the statistical significance of the difference between the respective means were noted. In this comparison the Average Performance cases were omitted. Correlation coefficients were also computed between each test item and the criterion, and for this computation all cases were included. Table 1 shows the results of both of these procedures.

Another approach to evaluating the data is to select a possible qualifying score on each item, apply it to both High and Low criterion groups, and compare the results in terms of the percentage of each group qualifying. In each instance this was done with a view to maximizing the differences between High and Low groups. Table 2 presents the results of this comparison. As may be seen, the item producing the greatest difference between those who did well in BCT and those who did poorly was the Bottle Cap Sorting measure. When a qualifying score for this item was set at 5.5, 81% of the High Performance group passed, while 50% of the Low Performance group did. This difference is statistically significant and large enough to be useful.

The other sorting item, Ammunition Sorting, ranked next to Bottle Cap Sorting in yielding a difference between good and poor BCT performers, based on the data in Table 2. Forty-three percent of the High Performance group sorted more than 300 shells, while 20% of the Low Performance group did.

Both of these measures were hypothesized to be essentially the same, and in the sample tested each individual was given one of these measures but not both. The fact that both showed considerable prediction of BCT performance suggests their possible usefulness in alternate forms of an eventual test. In terms of the psychological areas that ARI is attempting to measure, dependability as reflected by willingness to work conscientiously without direct supervision would appear to be both measurable and positively related to performance in BCT.

Of those items administered to the entire sample, the Exercise Total score showed the most predictive value. When a qualifying score was placed at 19, 46% of the High Performance group passed, while 24% of

Table 1

VALIDITY STATISTICS OF PERFORMANCE TEST ITEMS

Test Items	Means of High and Low Criterion Groups		p	Correlation of All Groups with BCT Criterion	
	N	High		Low	r
BOTTLE CAP SORTING	IV ^a	7.0	5.9	<.01	.35
	III & IV ^b	6.9	6.0	<.01	.27
AMMUNITION SORTING	IV	259	239	ns	.11
	III & IV	257	240	ns	.09
EXERCISE TOTAL	IV	18.57	15.82	<.05	.20
	III & IV	19.07	15.33	<.01	.23
SAND TABLE	IV	13.47	12.28	<.05	.16
	III & IV	13.87	12.58	<.05	.16
LAYOUT	IV	2.09	1.72	ns	.11
	III & IV	2.04	1.58	ns	.11
CHAIRS	IV	1.79	1.70	ns	.06
	III & IV	1.90	1.77	ns	.07
PUNCTUALITY ^c	IV	.61	.68	ns	-.06
	III & IV	.63	.70	ns	-.06
EXERCISE EFFORT ^c	IV	.59	.53	ns	.05
	III & IV	.58	.49	ns	.08
FORGOTTEN TEST ^c	IV	.85	.81	ns	.01
	III & IV	.84	.79	ns	.01

^a All examinees in Mental Category IV.

^b Includes additional cases in lower part of Mental Category III.

^c Means for Punctuality, Exercise Effort, and Forgotten Test based on dichotomous scoring: 0 or 1.

Table 2

COMPARISON OF TEST ITEM SCORES OF CATEGORY IV MEN FROM THE HIGH PERFORMANCE GROUP WITH SCORES OF MEN FROM THE LOW GROUP

TEST	POSSIBLE QUALIFYING SCORE	PERFORMANCE GROUP	N	% QUALIFYING	PERCENTAGE POINTS DIFFERENCE	
					10	30
BOTTLE CAP SORTING	5.5+	HIGH	64	81		31**
		LOW	36	50		
AMMUNITION SORTING	300+	HIGH	56	43		23*
		LOW	20	20		
EXERCISE TOTAL	19+	HIGH	125	46		22**
		LOW	58	24		
SAND TABLE	11+	HIGH	125	74		18*
		LOW	57	56		
LAYOUT	2+	HIGH	125	36		13 ns
		LOW	57	23		
PUNCTUALITY	1+	HIGH	124	60		10 ns
		LOW	56	70		
CHAIRS	1+	HIGH	125	55		8 ns
		LOW	57	47		
EXERCISE EFFORT	1+	HIGH	125	59		6 ns
		LOW	57	53		
FORGOTTEN TEST	2	HIGH	108	85		4 ns
		LOW	52	81		

* $p < .05$

** $p < .01$

the Low Performance group did. This outcome was not fully anticipated. The main intent of the Exercise item was to tap willingness to make an extra effort to push oneself physically. However, the finding may indicate that actual physical condition at time of enlistment, as revealed by such a specific exercise as leg-ups, has a bearing on success in BCT.

The Sand Table score was nearly as discriminating as Exercise Total score. When a qualifying score of 11 was established, 74% of the High Performance group passed while 56% of the Low group did. This measure of observational thoroughness seems to have relevance to BCT performance. The contribution of short-term memory to success on this item may deserve eventual examination.

The Layout item was also useful, but to a slightly lesser degree. Use of a qualifying score of 2 resulted in 36% of the High Performance group passing and 23% of the Low group. Participation in a group task--cooperativeness as measured by the Layout item--may be related to BCT success to a degree approaching statistical significance.

The Punctuality and Chairs items were marginally predictive. They may prove to be more helpful in a different environment; at least their use should be investigated further. The Forgotten Test and Exercise Effort scores did not appear to be promising.

DISCUSSION

All of the relationships pointed out in the preceding section are based on Mental Category IV examinees (N = 225). A smaller group of low Category III examinees (N = 88) was tested with the same items, with practically the same results (see Table 1). Correlation coefficients for each item against the criterion of BCT performance, another index of the relative validity of the items, showed essentially the same outcome, ranging from zero (no validity) to .35, which is not only statistically significant but probably high enough to be operationally practical. When results obtained with Category III cases were combined with the Category IV findings the picture was also very similar (Table 1). At least four items--Sorting (both Bottle Caps and Ammunition), Exercise Total, Sand Table, and Layout--indicated prediction of the BCT criterion of sufficient magnitude to offer a likelihood of eventual usefulness as screening measures.

Despite reasonable precautions, differences occurred in test administration among the several administrators during the data collection for this research, timing of the items was not always precise, and, especially on the Layout item, scoring became somewhat subjective on occasion. These deficiencies should be overcome or at least reduced before further research is undertaken.

These performance test items need to be administered, cross-validated, and standardized at an AFEES, since if they are to be helpful as screening devices the AFEES is the most likely place for their administration. One of the most important questions to be answered at this point is whether or not applicants for enlistment, at an AFEES, will react in the same way to these items as did newly enlisted recruits in a Reception Station. In the Reception Station they were captive subjects, already committed for two or more years; at an AFEES they would be free to change their minds and walk away if they chose. Would any appreciable differences in motivation, willingness, or interest be reflected in test performance? Should any modifications be made in the test items as a result? Such questions must be answered.

In considering the validities of the individual performance tasks, it is relevant to observe that because most of the tasks are single actions that sample a narrow aspect of behavior, they have very low reliabilities, which in turn limits their validities. The tasks are directly comparable in this regard to the individual items of a conventional test such as the AFQT, which are recognized as possessing only modest validity. However, when a large number of the more valid conventional items are combined the resulting test yields a higher validity than any of its individual constituent items. Similarly, if several of the more valid performance items developed here were combined, the resulting performance test should be more valid than any of the individual items. Such a combination will be investigated in the subsequent cross-validation.

Apart from the technical considerations involved in introducing performance tests in AFEES, there are also administrative concerns. One is the physical space requirement; small but separate rooms are needed for administration of the items, most of which must be given individually--a requirement which could also produce scheduling problems of greater or lesser magnitude. A second concern, perhaps more difficult to handle, is test security. Most of the items used are novel and easily remembered by the examinee; they measure attitudes which can be faked, not abilities such as typing skill which cannot. Leakage, intentional or unintentional, could be expected, as tested applicants talk over their testing experiences with friends, potential applicants, and recruiters. To minimize this will require adequate numbers of additional items to alternate with the ones discussed here. Development of more items that may ultimately be suitable for such a purpose is underway.

In summary it may be reasonable to say that thus far the research on performance tests for use in screening marginal personnel has pointed to some problems, none of which appears insurmountable. More significantly, it has also resulted in the development of a number of items which in an initial tryout at Fort Dix Reception Station showed encouraging relationships with performance in BCT. The next step should be administration of these items at an AFEES for cross-validation and administrative feasibility. Successful results of such a tryout would indicate operational suitability as a supplementary screening test for marginal personnel.

FURTHER WORK

This progress report has been limited to the work performed in the initial phase at Fort Dix Reception Station. Active trials have been started at the Fort Jackson AFEEES to begin answering the questions identified in this report.

However, several other research analyses and new empirical research tasks are warranted. The intercorrelations among these performance test items and with Army Classification Battery scores need to be determined. Also, a number of new performance test items have been identified and need to be tried out. Other research findings need to be taken into account, such as the results of the University of Michigan Human Performance Center research program on performance measures,¹ which may contain substantial noncognitive, motivational variables that can be tapped and measured rather simply and quickly.

This area of research also should be extended and broadened in light of emerging Federal Government guidelines on selection and employment techniques that comply with equal employment opportunity requirements. In this area, it is necessary to insure that as many relevant human capacities as possible are taken into account--particularly the attitudinal, noncognitive ones.

¹Rose, A. M. Human Information Processing: An Assessment and Research Battery (U). Tech. Report No. 46. Human Performance Center, University of Michigan. January 1974.