ENHANCING QUALITY CONTROL IN THE TESTING OF MILITARY APPLICANTS.

M. A. Fischl and Robert M. Ross

PERSONNEL UTILIZATION TECHNICAL AREA

U. S. Army Research Institute for the Behavioral and Social Sciences

Approved for public release; distribution unlimited.
NOTICES

DISTRIBUTION Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to U. S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI TP, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION This report may be destroyed when it is no longer needed. Please do not return it to the U. S. Army Research Institute for the Behavioral and Social Sciences.

NOTE 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.
ENHANCING QUALITY CONTROL IN THE TESTING OF MILITARY APPLICANTS

M. A. Fischl and Robert M. Ross

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, VA 22333

Deputy Chief of Staff for Personnel
Washington, DC 20310

April 1980

Approved for public release; distribution unlimited.

Diverse sources of error must be controlled for aptitude tests to have substantial validity. This paper describes a highly cost effective procedure for immediate verification of the veridicality of operational test scores.

A continuous need exists to maintain the high quality of testing procedures and of operational test scores used in selecting and classifying.
enlisted personnel. In a large-scale testing program such as the one that uses the Armed Services Vocational Aptitude Battery (ASVAB), the risk of test compromise is always present. A cost effective procedure for detecting the incidence of spurious scores was developed, consisting of (a) comparison of scores on two ASVAB subtests to detect any large differences between them, (b) administration of a 10-minute retest to examinees showing the large difference, and (c) comparison of original and retest scores to verify the incidence of likely test compromise. Tryout of the procedure indicated that the 10-minute retest of fewer than 20% of all examinees could detect approximately 70% of all cases of test compromise.
ENHANCING QUALITY CONTROL IN THE TESTING OF MILITARY APPLICANTS

M. A. Fischl and Robert M. Ross

Submitted by:
Joyce L. Shields, Chief
PERSONNEL UTILIZATION TECHNICAL AREA

Approved by:
E. Ralph Dusek
PERSONNEL AND TRAINING RESEARCH LABORATORY

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333
Office, Deputy Chief of Staff for Personnel
Department of the Army

April 1980

Army Project Number
2Q163101A768

Enlisted Selection

Approved for public release; distribution unlimited.
ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.
This research is related to ongoing ARI efforts to make the most efficient use of the Armed Services Vocational Aptitude Battery (ASVAB) for selection and classification of enlisted recruits. The report describes a quality-control technique that can be used to detect cases of possible ASVAB compromise. The technique is based on psychometric properties of ASVAB sub-tests. The ASVAB research is responsive to requirements established by the Deputy Chief of Staff for Personnel, Department of the Army, and was conducted under Army Project 2Q163101A768.
ENHANCING QUALITY CONTROL IN THE TESTING OF MILITARY APPLICANTS

BRIEF

Requirement:

The Armed Services Vocational Aptitude Battery (ASVAB) is the principal test battery used to select and classify recruits. A test cannot provide accurate information if it has been compromised (that is, if an examinee learns the questions and answers in advance). This research provides an operational method of detecting a substantial proportion of individual cases of ASVAB compromise.

Strategy:

The basic strategy analyzed the statistical relationships between the ASVAB subtest most vulnerable to compromise and other subtests, using ASVAB scores from 1,000 enlistment applicants native to determine the range of normal and abnormal score patterns. The Word Knowledge (WK) subtest is the most likely to be compromised, in part because vocabulary words are fairly easy to remember, look up, and discuss. The Arithmetic Reasoning (AR) subtest, on the other hand, is not easily compromised. Most people score in the same range on both tests. Cases in which the WK score is more than 10 points higher than the AR score are suspect. Retests with the 10-minute WK subtest from the 1973 Army Classification Battery (ACB-73), which is no longer in use for the active Army and therefore unlikely to be compromised, show that a difference of 11 to 14 raw score points between the two WK subtests would confirm cases of test compromise.

Field Tryout:

Several months after the nationwide data collection, a sample of 111 enlistees whose ASVAB scores had been recorded in that collection were retested with the ACB-73 at the Fort Jackson, S.C., Reception Station. Comparing their recorded ASVAB WK and AR scores flagged 20 cases; comparing the ASVAB WK and ACB WK scores for these 20 cases identified 9 as highly suspect. Both sets of WK scores were then compared for the entire sample, and 13 highly suspect cases were identified in all. That is, retesting 18% of the sample (20 out of 111) identified about 70% of the compromise cases (9 out of 13).

Utilization of Findings:

This quality-control procedure for aptitude testing is highly cost effective because of its simplicity, short testing time, and screening effectiveness.
ENHANCING QUALITY CONTROL IN THE TESTING OF MILITARY APPLICANTS

CONTENTS

INTRODUCTION .......................................................... 1
APPROACH .............................................................. 2
IMPLEMENTATION ....................................................... 2
EMPIRICAL TEST ......................................................... 4
SUMMARY AND CONCLUSIONS ........................................ 5
DISTRIBUTION ........................................................... 7

LIST OF TABLES

Table 1. Statistical description of AFQT subtests of ASVAB-6 ........ 3
2. Statistical description of Word Knowledge subtest
in ACB-73 (WK 2) ..................................................... 4
3. Results of empirical test ........................................... 5
ENHANCING QUALITY CONTROL IN THE TESTING OF MILITARY APPLICANTS

INTRODUCTION

All testing is subject to influences, lasting and temporary, general and specific, that cause the aptitude test score an individual attains to vary from the theoretical true score. For purposes of prediction in selection and classification through the use of testing, all reasons that would increase this variance over a group may be considered error.

Such semipermanent influences as the ability to deal with instructions on tests, or general examinee strategies for answering test questions, vary widely with individuals. The services have used several methods in attempts to reduce error attributable to this "test wiseness." Instructions are easy to understand and are targeted to low levels of reading ability, and sample test items and sample instructions are provided in an information pamphlet intended to familiarize everyone concerned with the nature of the test.

Temporary influences on test scores may also affect measurement. A person's physical and emotional condition and the physical testing environment may cause variation from true scores. To reduce these temporary effects that add to measurement error, care is taken to excuse from the testing session persons who are clearly ill or excessively fatigued, or persons who are disturbing others; regulations prohibit testing for long periods without breaks, or testing in places without proper lighting and temperature conditions.

Scoring and recording errors occur either as transitory human errors or, at times, as semipermanent conditions when, for example, an undetected malfunction develops in equipment used to score tests. Generally, the variety of scoring aids now used in Armed Forces Examining and Entrance Stations (AFEES), including optical scanning equipment, not only reduces errors but saves time as well.

Another source of measurement error is test compromise. These measurement errors, rather than being randomly distributed, usually operate in one direction—to yield overestimates of qualifications. Although compromise probably would not affect the measurement of very large numbers of enlistees as could other measurement errors, its nonrandom character makes test security of great importance.

In the past, the most common means of coping with test compromise has been by use of alternate test forms. There are two types of alternate forms, and they differ in cost of production and in the kind of protection they provide. One type uses the same items, but arranged in different sequences in different test booklets. This type remedies situations in which the compromise has taken the form of examinees being provided with a key to the correct answers, but not the content of those answers (for example: 1a, 2c, 3d, etc.). This type of compromise is believed to be relatively uncommon. The other type of alternate test form is very much more costly to produce but also very much more comprehensive in its protection. It consists of two tests with similar (but not identical) content, matched in difficulty and other statistical properties. The protection afforded is not just for cases that include
applicants having the key, but for applicants having the full answers to one of the forms. Both of the two types of alternate test form are now used in the test quality control programs of the services.

The parallel forms approach provides reasonable protection, but at an extremely high cost of production. That approach also does not, in and of itself, identify cases of suspect scores.

This paper describes an alternative approach to test quality control that involves minimum test development costs as well as minimum examining time onsite.

APPROACH

The objective of this development was to provide an operational tool to detect a substantial percentage of enlistment qualification test compromise cases. The general strategy was to capitalize on what is known or can be deduced logically concerning the differential compromise vulnerability of the various parts of the battery (ASVAB), and to combine that information with known statistical relationships among the subtests so as to "flag" highly unusual score patterns for subsequent followup.

Operational experience has shown that the main target for compromise has been the AFQT portion of the test battery. AFQT has been in joint services use the longest; for some of the services, AFQT is the principal selection standard. The nature of its contents--vocabulary, arithmetic problems, and geometric figures--are generally the best known of all military tests.

Within the AFQT portion of the battery, experience has indicated that, if compromise takes place, the compromise usually involves the vocabulary items. This is not surprising because vocabulary words are easy to remember and to look up after the examination. The other two subtests do not lend themselves to this kind of compromise: the arithmetic problems are relatively long prose paragraphs, and there is no readily available source of the right answers; and the totally pictorial test of spatial relations is nearly impossible to compromise through memory.

Given (a) that Word Knowledge is probably the key ASVAB subtest compromised, that (b) the other components are relatively hard to compromise, and that (c) the psychometric relationships among these subtests are stable and known: Likely compromise can be detected by comparing discrepancies in score between the Word Knowledge subtest and one or both of the other AFQT components (Arithmetic Reasoning, Space Perception).

IMPLEMENTATION

The numeric values needed to begin to implement the logic of this approach were derived from a national sample of 1,000 AFEES applicants drawn in January 1976. These 1,000 cases were stratified on AFQT to conform to the standard mobilization reference population, and the statistics shown in Table 1 were obtained. As may be seen, the correlation of Word Knowledge (WK) with Space Perception (SP) is 0.43. This means that fairly sizable
score discrepancies between WK and SP can be expected just by chance. On the other hand, the correlation of WK with Arithmetic Reasoning (AR) is high enough to be usable, 0.68. As a result, development focused on use of the WK/AR discrepancy.

Table 1

Statistical Description of AFQT Subtests of ASVAB-6

<table>
<thead>
<tr>
<th>Subtests</th>
<th>Mean</th>
<th>SD</th>
<th>Correlation with WK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Knowledge (WK), 30 items</td>
<td>17.5</td>
<td>7.5</td>
<td>-</td>
</tr>
<tr>
<td>Arithmetic Reasoning (AR), 20 items</td>
<td>11.7</td>
<td>4.8</td>
<td>0.68</td>
</tr>
<tr>
<td>Space Perception (SP), 20 items</td>
<td>10.3</td>
<td>4.1</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Standard error of estimate of WK on AR = 5.5
Regression line: WK = 5.07 + 1.06 (AR)

Note. N = 1,000 AFEES applicants tested in January 1976.

The intention was to develop an initial screening procedure that would "flag," as suspicious, cases in which the WK raw score exceeded the AR raw score by some amount greater than chance expectation. The regression line of WK on AR can predict the expected WK score from any AR score (Table 1). The prediction has confidence limits defined by the standard error of estimate of WK on AR and the confidence level selected. A somewhat low one-tailed confidence level of $p < 0.80$ was chosen in consideration of maximizing detectability for subsequent followup. Using the regression formula and standard error of estimate it was found that a difference of 10 raw score points between WK and AR is unlikely to occur by chance, i.e., outside the confidence interval. The 10-point difference is appropriate through the score range AR < 15 since the regression coefficient was so close to 1.0. Therefore, 15% to 20% of cases that have a difference equal to or greater than 10 points and with AR less than 15 are flagged as unusual cases. These cases will be the only ones used in further screening for possible compromise detection.

A group exhibiting the unusual score pattern consists of two types of individuals: (a) those for whom the abilities measured by the WK subtest are truly well in excess of their abilities in the domains measured by AR, and (b) those whose WK scores are artificially inflated through some breach of test security. The next step, then, is to separate these types.

The simplest way to sort the compromise cases from the genuine, though unusual, ones is to administer a 10-minute retest consisting of WK items known to be secure, and to compare performance on the WK retest with performance on the original WK. For some cases, the original WK score (WK 1) will replicate, plus or minus a calculable chance error effect; for others,
the second WK score (WK 2) will be so much lower as to be virtually unex-
plainable through normal chance variation.

Just as the initial screen utilized the values shown in Table 1 to de-
fine the critical WK/AR difference, values in the same table plus those in
Table 2 were used to set the limits for the WK 1/WK 2 difference. For this
step a confidence level of $p < 0.95$ was set to minimize the risk of false ac-
cusation and to identify cases virtually unexplainable by the hypothesis of
chance variation.

Table 2

Statistical Description of Word Knowledge Subtest in ACB-73 (WK 2)

<table>
<thead>
<tr>
<th>Number of items</th>
<th>Mean</th>
<th>SD</th>
<th>Correlation with ASVAB-6 WK 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>11.8</td>
<td>4.6</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Standard error of estimate of ASVAB-6 WK on ACB-73 WK = 4.87
Regression line: WK 1 = 2.88 + 1.24 (WK 2)

A difference of 11 to 14 raw score points (depending on the level of the
WK 2 score) between WK 1 and WK 2 is the critical difference, i.e., beyond
that point score differences are probably not due to chance. Individuals ex-
hibiting a larger difference are identified as most likely having received
improper pretest assistance.

EMPIRICAL TEST

In the spring of 1976 a sample of 111 enlistees who had been tested with
ASVAB at AFEES was retested at the Fort Jackson, S.C., Reception Station with
ACB-73. ACB-73 contains WK and AR subtests and was the Army's basis for com-
puting AFQT scores until it was replaced by ASVAB-6 and -7 in January 1976.
At the time the test sample was drawn, ACB-73 was no longer operational, and
hence its WK subtest could be considered as completely secure.

The first step in the test was to calculate the one-sided difference of
ASVAB-6 WK minus AR and to refer it to the specified critical difference of
10 points. This step identified 20 cases.

The second step was to calculate the one-sided difference of ASVAB-6 WK
minus ACB-73 WK and refer that difference to the specified critical differ-
ence of 11 to 14 raw score points. This procedure identified 9 of the 20
flagged cases as highly suspect compromise cases. These and other important
relationships are summarized in Table 3. As may be seen, when the retest
scores of the entire sample were examined, 13 cases were identified as highly
suspect. Under operational conditions, only 18%—the 20 flagged cases—would have been retested, and only 9 of the 13 highly suspect cases caught; that is, retest of less than 20% of the sample caught about 70% of the compromise cases.

Table 3
Results of Empirical Test

<table>
<thead>
<tr>
<th>Flagged by WK-AR</th>
<th>Passed by WK-AR</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Clean&quot;</td>
<td>11</td>
<td>87</td>
</tr>
<tr>
<td>Highly suspect</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>91</td>
</tr>
</tbody>
</table>

A final empirical test was performed to assure maximum certainty of the percentage of the input which would have to be retested under the rule of WK-AR ≥ 10 points. It may be recalled that 10 points implements a confidence level of 0.80—i.e., about 15% to 20% of the population flagged for retesting—and one sample, at Fort Jackson, yielded 18% so flagged. In mid-1976, another sample of AFEES data was drawn, of size 500, and the WK minus AR criterion was again applied. Results in this sample flagged 17% of the cases.

SUMMARY AND CONCLUSIONS

In recognition of the fact that the Word Knowledge subtest is the most vulnerable to compromise of all the tests in the selection and classification battery, a simplified procedure was developed to detect WK compromise. The procedure has two steps:

1. At the time of scoring the AFQT portion of the battery, separate those papers in which the AR raw score is less than 15, and the WK raw score is 10 or more points greater than the AR score. This step will flag, as potentially suspect, some 15 to 20% of the cases.

2. To only those flagged by step one, administer a 10-minute retest consisting of a completely secure WK and separate those papers in which the WK retest score is at least 11 to 14 raw score points lower than the original WK score (checklist tables can easily be prepared to accomplish all conversions and all comparisons with critical differences). This combination of steps will identify, as highly suspect, approximately 70% of all cases of likely test compromise.
An alternative to the two-step procedure is to administer the WK retest to everyone and apply the rule of an 11 to 14 raw score point drop. This will detect more compromise cases, but at five to seven times the cost (that is, retesting 100% of AFEES applicants instead of between 15% and 20%).

Another alternative is to enlarge the requisite WK/AR difference so as to retest 10% of the input. In the Fort Jackson sample, this detected about 40% of the likely compromise cases.

For any of these alternatives, the conclusion may be drawn that a simple and cost effective procedure for enhancing quality control in the testing of military applicants has been developed.
LIBRARY OF CONGRESS UNIT DOCUMENTS EXPEDITION PROJECT
4 EDITION R AND O MAGAZINE ATI NR URLUE-LMN
1 US GOVERNMENT PRINTING OFP LIBRARY, PUBLIC DOCUMENTS DEPARTMENT
1 US GOVERNMENT PRINTING OFP LIBRARY AND STATUTORY, L19 DIV (SLL)
1 IMP ARMY LIBRARY