

NO-A101 377

MARINE CORPS PROJECT TO VALIDATE THE ASVAB (ARMED
SERVICES VOCATIONAL APT. (U) CENTER FOR NAVAL ANALYSES
ALEXANDRIA VA M MAIER MAY 87 CNA-PP-454

1/1

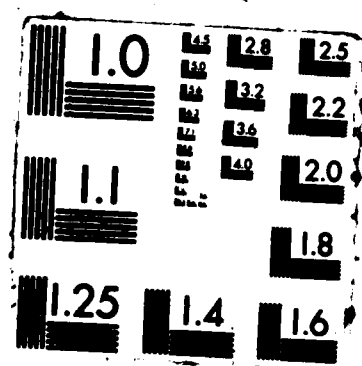
UNCLASSIFIED

NO0014-87-C-0001

F/G 5/9

NL





AD-A181 377

Marine Corps Project To Validate The ASVAB Against Job Performance

by

Milton Maier

DTIC
ELECTE
JUN 03 1987

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited

D

N00014-87-C-0001

A Division of

CNA

Hudson Institute

CENTER FOR NAVAL ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

87 6 2 014

Approved for Public Release. Distribution Unlimited.

**The ideas expressed in this paper are those of the author.
The paper does not necessarily represent the views of the
Center for Naval Analyses.**

MARINE CORPS PROJECT TO VALIDATE THE ASVAB AGAINST JOB PERFORMANCE

→ The Marine Corps conducted an initial validation study in 1981 to evaluate the predictive validity of the Armed Services Vocational Aptitude Battery (ASVAB) against hands-on and written job performance tests. The results were presented to the Joint Service Working Group and the National Academy of Sciences Advisory Committee in briefings and in a series of CNA reports. Now it is time to revisit our initial study in light of the services' experience since 1981.

DESCRIPTION OF THE MARINE CORPS INITIAL STUDY

* Three occupational specialties were included in the initial Marine Corps study. These are Infantry Rifleman, with relatively low technical demands; Automotive Mechanic, with moderate technical demands, and Radio Repairer, with high technical demands. Development of the tests for all three specialties was completed in about 8 months, and the administration of the tests in about 4 months. With such a tight schedule, there was little opportunity to go through the rigorous process that the Working Group subsequently endorsed.

▷ An issue about selecting test content was whether the performance tests should be primarily descriptions of the current proficiency of examinees on a representative set of job requirements, or primarily predictive in indicating how well examinees would be expected to perform on the full set of job requirements. The performance test for the riflemen was descriptive, in that the current proficiency of examinees

COPY
INSPECTED
■

Codes

Dit

over and/or
Special

A-1

fig 1
was evaluated in a broad range of job requirements. The performance test for the mechanics was primarily descriptive, but the test content was limited to the ~~quarter-ton jeep~~; proficiency on maintaining trucks was not tested. The performance test for the radio repairer, was primarily predictive; the entire hands-on test was devoted to troubleshooting a circuit board for a new piece of equipment that none of the examinees had seen before. Since then the Working Group has endorsed the position that the performance tests should be descriptive of the current proficiency of examinees.

Perhaps the greatest problem with the initial study was that the test administrators received little training on how to administer and score the hands-on tests, and they were not monitored during the study to ensure that they maintained the same scoring standards. Since then the quality of the test administrations has been a primary concern of the Working Group.

Predictive Validity of the ASVAB

The predictive validity of the ASVAB was evaluated against hands-on and written performance tests and grades in the training courses for the occupational specialties. The validity coefficients are shown in table 1. These coefficients are comparable to those traditionally found by the services for predicting grades in occupational specialty training courses.

TABLE 1

VALIDITY OF THE ASVAB FOR PREDICTING PERFORMANCE
IN THREE MARINE CORPS OCCUPATIONAL SPECIALTIES

<u>Specialty</u>	<u>Hands-on test</u>	<u>Written test</u>	<u>Training grades</u>
Infantry Rifleman	.58	.69	.29
Automotive Mechanic	.56	.65	.83
Radio Repair	.59	.73	.75

The validity coefficients for predicting the hands-on tests are very respectable, .56 to .59. Values of this magnitude justify using the ASVAB in making selection and classification decisions. The coefficients of course need to be verified, and the accruing results by the other services are providing comparable values.

The values reported in table 1, however, did not spring full blown from the first pass through the computer. Three steps were required to improve the quality of the data and make the results more generalizable: cleaning up the data; adjusting for scale differences among the test administrators; and correcting for range restriction. A brief description about each follows:

Cleaning Up the Samples

In contrast to laboratory experiments, studies conducted in a field environment inherently produce dirty data. The data need to be cleaned up.

The riflemen sample of examinees required only nominal editing. Some people were tested with forms 8, 9, and 10 of the ASVAB, and some

with forms 5, 6, and 7. If we would have mixed these test forms, the results would have been disastrous. At the end of the paper, we mention some of the problems that have arisen in the past 8 years with the ASVAB score scale; failure to take these various score scales into account during the analyses could have serious consequences on the computed validity coefficients.

For automotive mechanics, some people had missing ASVAB scores or missing training course grades. When the latter group was deleted from the sample, the ASVAB validity coefficient increased from .24 to .29 (table 2). These examinees tended to have much job experience in the Marine Corps, and the other factors besides aptitude could have an effect on their proficiency, which would tend to pull down validity coefficients.

For the radio repairers, cleaning up the sample was essential. Examinees from three occupational specialties ended up being tested with the same performance tests; they differed in training and job experience, and consequently in their performance test scores. The cleaned-up sample is restricted to those people with the same training and job experience (as ground radio repairers). The validity coefficient for the unedited sample was only .01; by cleaning up the sample, it was increased to .13 (table 2).

TABLE 2

EFFECTS OF IMPROVING THE QUALITY OF THE DATA

<u>Specialty</u>	<u>Dirty data</u>	<u>Clean samples</u>	<u>Standardize scoring</u>	<u>Range restriction</u>
Infantry Rifleman	.39	-	-	.58
Mechanic	.24	.29	.49	.56
Radio Repairer	.01	.13	.21	.59

Standardizing Test Administrators

In hands-on testing, test administrators are the key to the accuracy of the scores. They simultaneously administer and score the tests. Deviations from standard practices can only introduce errors into the scores. The most common error is that administrators are too lenient, and they are not consistent in applying the same standards. These differences in scoring standards reduce the predictive validity of the ASVAB, and other variables. In addition, they render competency, or absolute, interpretations of the scores meaningless. Hands-on test scores are direct measures of job proficiency only when the test administrators are known to employ accurate and consistent scoring standards.

In our initial study, the test administrators did not produce competency scores. For the riflemen sample, we could not identify the test administrators, and no evaluation of their scoring standards was possible.

For the other two specialties the test administrators were identified, and their scoring accuracy could be evaluated. Both sets were inconsistent among themselves, which increases the variance of the

hands-on test scores, but of course lowers reliability. The administrators used different scoring standards as evidenced by the different means and standard deviations of the scores they assigned. These differences were not related to characteristics of the examinees they tested, such as aptitude, job experience, or written test scores.

We adjusted the scoring standards for each administrator by standardizing his set of scores to have a common mean and standard deviation. This statistical adjustment resulted in a big jump in the validity of the ASVAB for the automotive mechanic sample, from .29 to .49 (table 2); the increase for the radio repairer sample was more modest but still noticeable, from .13 to .21. The validity coefficients after standardizing the hands-on test scores are assuming respectable values.

A word about this statistical adjustment is in order. The standardized scores do put all the test administrators on the same scale; differences in leniency, or difficulty, are statistically removed, and in that sense the administrators are comparable. In another significant sense, however, differences among the test administrators remain. The administrators are analogous to different forms in paper-and pencil testing. The intent is that they be parallel forms, which means not only equal means and variances, but also equal covariances. That is, the intercorrelation among the tasks or steps in the hands-on tests should be equal for all test administrators. Standardizing the scores to give them the same mean and variance does not affect the covariance. Differences among the test administrators in

what they look at when they make their judgments remain, and they serve to lower the reliability of the tests. The only way we know of ensuring that the test administrators are scoring the same thing is through careful training and monitoring of their performance. That is, the controls must be experimental and not statistical. In our initial study, the test administrators were not functioning as parallel forms, and the validity coefficients are reduced because of their unreliability.

Correction for Range Restriction

The final adjustment we made to enhance the generalizability of the results was to put all the validity and intercorrelation coefficients on a common metric by correcting for range restriction. The samples of examinees were subject to different degrees of selection when they were assigned to the occupational specialties and during their training courses. The radio repairers are highly selected; only people whose aptitude scores are above average are eligible for electronics repair training, and many students fail for academic reasons. The automotive mechanics go through a less restrictive selection process, and the sample was fairly representative of the population. For riflemen, there is relatively little selection of the low end of the ability continuum, but both the extreme top and bottom of the population are underrepresented.

The final column of validity coefficients in table 2 shows the effects of correcting for range restriction. The increase over the

previous validities in table 2 illustrates the degree to which the samples were subject to selection on the basis of their ASVAB scores. The NAS Advisory Committee and the Joint Service Working Group endorse correcting for range restriction. The base population for correcting the correlation coefficients is the 1980 Youth Population. The primary purpose of the correction is to put all coefficients on a common metric, which makes them comparable across specialties and services. Incidentally, they also increase the magnitude of the coefficients.

IMPACT OF THE INITIAL STUDY ON THE DESIGN OF THE FOLLOW-ON MARINE CORPS PROJECT

The lessons learned from the initial study have had a profound effect on the design of the follow-on Marine Corps efforts. Three facets of the research design have been improved.

- o The test content covers the full range of job requirements.
- o Test administrators are being carefully trained and monitored.
- o The sample size has been expanded to include second-term Marines and to permit analysis of subgroups.

The first follow-on effort is for the Infantry Occupational Field. We expect that the results will go a long way to address concerns and criticisms of using the ASVAB for making selection decisions for infantrymen.

Coverage of Test Content

The job performance tests for the infantrymen are descriptive of the current level of proficiency. They will cover the full range of job requirements, from pay grades E1, private, through E5, sergeant, for the sample of first-term Marines, and pay grades E1 through E6, staff sergeant, for the sample of second-term Marines. The domain of job requirements is defined by the Marine Corps Individual Training Standards, which are comparable to the Army Soldier's Manuals. The test content has been randomly selected from the Individual Training Standards. The random selection is generally at the subtask level. The judgments of the job experts are used in refining the randomly selected subtasks, or training steps as they are called by the Marine Corps, to help ensure the job relevance of the tests. Because testing time is severely limited, it needs to be filled with activities that provide information relevant to making inferences about job proficiency.

Training and Monitoring of Test Administrators

Because of the crucial role that test administrators fulfill in hands-on testing, their performance cannot be left to chance. The full-scale data collection for the infantrymen is scheduled to begin

August 1987. Before administering and scoring their first test, each test administrator will undergo rigorous training for up to 2 weeks. During this time, they will learn to administer the entire hands-on test, and they will demonstrate that they can do so objectively and reliably. Teams of three or four administrators will work independently in scoring individuals during the training period. The training will continue until they agree with each other. The test administrators will be recently retired Marines who served as infantry unit leaders.

During the full-scale administration, the test administrators will be monitored daily to make sure that the proper scoring standards are being maintained. We expect to rotate administrators across testing stations, which will randomize idiosyncratic administrator's effects and hopefully minimize boredom and fatigue.

Sample Size

The initial study produced the tantalizing finding that the ASVAB predictive validity drops dramatically for Marines with 2 or more years of time in service. If this result is verified in the follow-on study, we will need to rethink the tradeoff between aptitude and training and how we justify enlistment standards.

The sample will consist of 1,200 infantry riflemen in pay grades E1 through E4, private through corporal, and 600 riflemen in pay grade E5, sergeant. In addition, 300 people will be tested in each of the other infantry specialties - Machine Gunner, Mortar Man, and Assaultman. With samples of this size, we can evaluate the validity of

the ASVAB for people with different levels of job experience and pay grade. Also, by testing sergeants we can evaluate how aptitude is related to proficiency on the more demanding tasks of unit leaders. At least for the vital Infantry Occupational Field we will know how valid the ASVAB is as a predictor of job performance.

DISCUSSION

The greatest lesson learned from the initial study is so obvious that it should not require mentioning - we need to exercise greater quality control over the entire process. The concept is easy, but the execution is difficult. We now address some specific points that require attention, and for some of these points there is no consensus on the right procedure.

Test Development

Hands-on testing in the past 2 decades has grown up in the training community, with relatively little scrutiny or input from the measurement point of view. From a training point of view, the job task is the natural unit, and people are trained and tested on job tasks. If the person can perform the task, training stops; if not, then training continues. Classification errors are generally of no great consequence because of the redundancy in the training process. In the testing environment, however, time is limited, and classification errors can have serious consequences.

From a testing point of view, the task is not necessarily the natural unit. The hands-on test for mechanics in our initial study, for example, had a strange mixture of testing time for a task and the number of scoring units that the task contributed to the total score. The wheel-and-brake task required over 1 hour of testing time, but produced only 13 points of discrimination. The coil task, by contrast, required only 30 minutes, but produced 18 points of discrimination. Surely the test developers intended that the wheel-and-brake task should have a greater impact on total score, yet by simply adding up the scores, the coil test would contribute more to the total score variance.

Contribution of task scores to the total score can be handled after the fact by statistically weighting the tasks and scores. Our preferred solution is to control the weighting experimentally. Our rule is that equal units of time should generally produce equal units of scores. The more important tasks would receive more time and hence make a larger contribution to total score. We conform to traditional measurement practice.

Editing the Data

We have already dwelled at length on the need to clean up the data from performance tests. The sources of dirty data are legion, and we have only mentioned the most obvious ones. What may be less well recognized is that the ASVAB scores themselves are dirty. A short description of problems with them may be illuminating.

The natural tendency of researchers may be to retrieve the ASVAB scores of record and use them in the regression analysis. After all, these scores have official status, and they are used in making personnel decisions. Such an easy solution could have a serious impact on the computed validity coefficients.

In our sample of second-term Marines, many of them will have been tested with forms 5, 6, and 7 of the ASVAB administered before October 1980. These scores of record are seriously inflated because of problems with the conversion tables. In this case, the subtest raw scores need to be retrieved and placed on the correct score scale.

Most examinees in our study, as for the other services, will have been tested after 1 October 1980, which avoids the gross inflation of test scores, but still there are problems. On 1 October 1984, the ASVAB score scale was changed when the World War II Mobilization Population was dropped as the reference for the ASVAB score scale and replaced by the 1980 Youth Population. Mixing these two score scales would introduce unwanted variance into the ASVAB scores; the AFQT scores, for example, could shift up to 4 percentile score points. In 1986, an adjustment was made to the conversion tables for forms 11, 12, and 13, which changed the AFQT scores by about 2 percentile score points. Again, mixing the scores of record for score before and after July 1986 would introduce unwanted variance.

Another concern is that the services permit retesting on the ASVAB by inservice personnel. They can improve their ASVAB scores obtained at time of enlistment by retesting when they have gained more experience.

For purposes of evaluating enlistment standards, the ASVAB scores obtained at time of enlistment should be used, and not the retest scores.

The solution to the ASVAB score problem is to retrieve the subtest raw scores used at time of enlistment and convert them to a common metric. The raw scores cannot be used as they are because the examinees took different forms that have different raw score means and standard deviations. Conversion tables are available to transform scores on the World War II scale to the 1980 scale, and this needs to be accomplished before the ASVAB scores from pre-October 1984 was combined with those from past October 1984.

Our conclusion about quality control is that data collected in the natural or field environment cannot be taken at face value. Even though our scientific ethos may tell us not to mess with the data, the consequences of using them as they are given to us usually are worse than employing judicious care in cleaning them up.

A final observation about placing our initial study in the emerging context of the Joint Service Job Performance Measurement Project. Since 1981 we have analyzed and developed our data, usually from a critical point of view because we wanted to learn how to improve our research design. Now we think we have accomplished that.

CNA PROFESSIONAL PAPER INDEX¹

PP 407²

Laird, Robbin F. *The French Strategic Dilemma*, 22 pp., Nov 1984

PP 415

Mizrahi, Maurice M. *Can Authoritative Studies Be Trusted?* 2 pp., Jun 1984

PP 416

Jondrow, James M., and Levy, Robert A. *The Displacement of Local Spending for Pollution Control by Federal Construction Grants*, 6 pp., Jun 1984 (Reprinted from *American Economic Review*, May 1984)

PP 418

Reslock, Patricia A. *The Care and Feeding of Magnetic Tapes*, 7 pp., Jul 1984

PP 420

Weiss, Kenneth G. *The War for the Falklands: A Chronology*, 32 pp., Aug 1982

PP 422

Quester, Aline, and Marcus, Alan. *An Evaluation of The Effectiveness of Classroom and On the Job Training*, 35 pp., Dec 1984. (Presented at the Symposium on Training Effectiveness, NATO Defense Research Group, Brussels, 7-9 January 1985)

PP 423

Dismukes, N. Bradford, and Weiss, Kenneth G. *MARE MOSSO: The Mediterranean Theater*, 26 pp., Nov 1984. (Presented at the Seapower Conference, Washington, D.C., 26-27 November 1984)

PP 424

Berg, Dr. Robert M., *The CNA Ordnance Programming Model and Methodology*, 27 pp., Oct 1984. (Presented at the ORSA-MAS/MDRS Symposium, Washington, Aug 1984)

PP 425

Horowitz, Stanely A., and Angier, Bruce N. *Costs and Benefits of Training and Experience*, 18 pp., Jan 1985. (Presented at the Symposium on Training Effectiveness, NATO Defense Research Group, Brussels, 7-9 January 1985)

PP 427

Cavalluzzo, Linda C. *OpTempo and Training Effectiveness*, 19 pp., Dec 1984. (Presented at the Symposium on Training Effectiveness, NATO Defense Research Group, Brussels, 7-9 January 1985)

PP 428

Matthes, Greg, Cdr., USN and Evanovich, Peter *Force Levels, Readiness, and Capability*, 24 pp., Nov 1984. (Presented at the ORSA-TIMS 26-28 November Meeting, Washington, D.C.)

PP 429

Perla, Peter P. and Barrett, Raymond T. LCdr., USN, *Wargaming and Its Uses*, 13 pp., Nov 1984. (Published in the *Naval War College Review*, XXXVIII, No. 5 / Sequence 311, September-October 1985)

PP 430

Goldberg, Matthew S. *The Relationship Between Material Failures And Flight Hours: Statistical Considerations*, 18 pp., Jan 1985

PP 431

McConnell, James M. *A Possible Change in Soviet Views on the Prospects for Anti-Submarine Warfare*, 19 pp., Jan 1985

PP 432

Marcus, Alan J. and Curran, Lawrence E., Cdr., USN. *The Use of Flight Simulators in Measuring and Improving Training Effectiveness*, 29 pp., Jan 1985 (Presented at the Symposium on Training Effectiveness, NATO Defense Research Group, Brussels, 7-9 January 1985)

PP 433

Quester, Aline O. and Lockman, Robert F. *The All Volunteer Force: Outlook for the Eighties and Nineties*, 20 pp., Mar 1984. (To be published in *Armed Forces and Society*, 1985)

PP 435

Levine, Daniel B. and Jondrow, James M. *Readiness or Resources: Which Comes First?* 12 pp., Mar 1985

PP 436

Goldberg, Matthew S. *Logit Specification Tests Using Grouped Data*, 26 pp., Jan 1985

1. CNA Professional Papers with an AD number may be obtained from the National Technical Information Service, U.S. Department of Commerce, Springfield, Virginia 22151. Other papers are available from the Management Information Office, Center for Naval Analyses, 4401 Ford Avenue, Alexandria, Virginia 22302-0268. An index of selected publications is also available on request. The index includes a listing of professional papers, with abstracts, issued from 1969 to December 1983).

2. Listings for Professional Papers issued prior to PP 407 can be found in *Index of Selected Publications (through December 1983)*, March 1984.

CNA PROFESSIONAL PAPER INDEX (Continued)

PP 438

Fletcher, Jean W. *Supply Problems in the Naval Reserve*, 14 pp., Feb 1986. (Presented at the Third Annual Mobilization Conference, Industrial College of the Armed Forces, National Defense University)

PP 440

Bell, Jr., Thomas D. *The Center for Naval Analyses Past, Present, and Future*, 12 pp., Aug 1985

PP 441

Schneiter, George R. *Implications of the Strategic Defense Initiative for the ABM Treaty*, 13 pp., Feb 1986. (Published in *Survival*, September/October 1985)

PP 442

Berg, Robert, Dennis, Richard, and Jondrow, James. *Price Analysis and the Effects of Competition*, 23 pp., Sep 1985. (Presented at the Association for Public Policy Analysis and Management - The Annual Research Conference, Shoreham Hotel, Washington, D.C., 25 October 1985)

PP 443

FitzGerald, Mary C., *Marshal Ogarkov on Modern War: 1977-1985*, 65 pp., Mar 1986

PP 445

Kober, Stanley, *Strategic Defense, Deterrence, and Arms Control*, 23 pp., Aug 1986. (Published in *The Washington Quarterly*, Winter 1986)

PP 446

Mayberry, Paul W. and Maier, Milton H., *Towards Justifying Enlistment Standards: Linking Input Characteristics to Job Performance*, 11 pp., Oct 1986. (Paper to be presented at 1986 American Psychological Association symposium entitled "Setting Standards in Performance Measurement".)

PP 448

Cymrot, Donald J., *Military Retirement and Social Security: A Comparative Analysis*, 28 pp., Oct 1986

PP 449

Richardson, Henry R., *Search Theory*, 13 pp., Apr 1986

PP 451

FitzGerald, Mary C., *The Soviet Leadership on Nuclear War*, 40 pp., Apr 1987

PP 452

Mayberry, Paul W., *Issues in the Development of a Competency Scale: Implications for Linking Job Performance and Aptitude*, 22 pp., Apr 1987

PP 453

Dismukes, Bradford, *Strategic ASW And The Conventional Defense Of Europe*, 26 pp., Apr 1987

PP 454

Maier, Milton, *Marine Corps Project To Validate The ASVAB Against Job Performance*, 14 pp., May 1987

END

7-87

DTIC