Predicting Criterion Performance

in

U.S. Navy Apprentice Technical Training

Stephen E. Watson, Ph.D.
Navy Selection and Classification (CNO N132)
Washington, DC.

&

Martin J. Ippel, Ph.D.
CogniMetrics Inc.,
San Antonio, TX

Paper presented at the 50th Annual Conference of the
International Military Testing Association
Amsterdam
September 29 – October 3, 2008

SUMMARY

Criterion development and the measurement of individual performance are generally considered critical for the evaluation of the utility and efficiency of a personnel selection and classification (PS&C) system. Training data often pose a fundamental problem for PS&C systems in that training performance tests seldom are designed to discriminate competence levels beyond the minimum requirements. This is, in particular, the case with computer-based training systems such as the Navy's Apprentice Technical Training (ATT) program. The ATT program is a modular training system. Each module consists of a series of lessons. Students work through the lessons at their own pace, take a test at the end of the lesson, and, if they pass the criterion score, move on to the next lesson. At the end of a module the student has to take two tests: a test of factual knowledge and a test of skills learned in the module (i.e., two post-test scores: a D-score and a P-score, respectively). When the student fails on either of these tests, he or she is expected to redo the lesson and take the test again. The result is a set of tests with score distributions that are highly compressed and negatively skewed.

An analysis of the criterion-referenced reliability of the ATT D-scores and P-scores demonstrated a complete failure to discriminate between the criterion score and individual performance (r_{tt}' s between 0.00 and 0.38). Using the data of a sample of 3195 Navy recruits we constructed a new set of efficiency-based measures from the existing training data which assess the probability of passing the criterion score on the post tests per module as an increasing function of the of the time taken to achieve mastery of individual lessons of the training modules. In this way, a set of two scores were constructed per module (i.e., new D-scores and P-scores). Reliability (internal consistency) analysis of the variables suggested a sufficient level of reliability for research purposes. Results of predictive utility analysis of ASVAB and ITAB presented show a significant and substantial incremental validity of the ITAB over the ASVAB selection composites in predicting ATT criterion scores.

Criterion development and the measurement of individual performance are generally considered critical for the evaluation of the utility and efficiency of a personnel selection and classification (PS&C) system (e.g., Campbell & Knapp, 2001). As a potential source for criterion development, training data often pose a fundamental problem for PS&C systems in that training performance tests seldom are designed to discriminate competence levels beyond the minimum requirements (Wolfe, 1997). This is, in particular, the case with computer-based training systems such as the Navy's Apprentice Technical Training (ATT) program at the Great Lakes Recruit Training Command in Chicago, IL.

The ATT program is a modular training system. Table 1 gives a listing of various training modules of the ATT program. Each module consists of a series of lessons. Students work through the lessons at their own pace, take a test at the end of the lesson, and, if they pass the criterion score, move on to the next lesson. At the end of a module the student has to take two tests: a test of factual, or declarative, knowledge and a test of skills, or procedural knowledge, learned in the module (i.e., two post-test scores: a D-score and a P-score, respectively). When the student fails on either of these tests, he or she is expected to redo the lessons and take the tests again. This procedure necessarily results in a set of test score distributions that are highly compressed and negatively skewed.

[Insert Table 1 about here]

Some of the training modules were designed for all ATT job ratings, others were designed for certain specialties.[1] The primary tool for selection and placement in the Armed Services is the Armed Services Vocational Aptitude Battery (ASVAB), test battery consisting of nine tests. Very early in the recruitment process, would-be recruits are screened with the Armed Forces Qualification Test (AFQT), a subset of ASVAB tests measuring verbal (tests: Word Knowledge (WK) and Paragraph Comprehension (PC)) and mathematics (tests: Arithmetic Reasoning (AR) and Math Knowledge (MK)) abilities. The AFQT was designed as a measure of trainability for jobs in the Armed Forces. The AFQT has high loadings on general intelligence and is to a large extent a measure of past learning (crystallized intelligence). Table 2 lists the ASVAB tests and their measurement claims.

[Insert Table 2 about here]

While the AFQT score, derived from the ASVAB, serves as a screening test for all Services, the Services combine the various ASVAB tests into "aptitude area" composites, which are used to

[1] The Navy and the Coast Guard refer to their enlisted jobs as ratings.

assign new recruits to military occupations (Sellman, 2004). Table 3 lists Navy ratings and corresponding ASVAB Selection Composites and their minimum scores for admission to technical training included in this study. These aptitude area composites have a classification function -- optimally matching jobs and available recruits.

[Insert Table 3 about here]

In summary, the ATT program is designed to successfully put through all students entering the system. The perspective of a PS&C system is different by nature. Its effectiveness is measured as the degree in which it ensures that instructional effort is allocated to students who can profit maximally in the shortest time possible. A precondition for such a system is the availability of achievement scores that reliably differentiate between individual accomplishments. The first goal of this study is to investigate whether data produced by the ATT program can be used to measure the effectiveness of the Navy's PS&C system. We will focus on the power of these measures to reliably differentiate between individual training results and how well these criterion measures can be predicted with various ASVAB selection composites. The second goal is to investigate whether the I.T. Aptitude Battery, a new test battery measuring procedural skill learning ability (fluid intelligence), would be a useful addition to the ASVAB selection composites for predicting outcomes of technical training.

Development of New Criterion Measures

Levels of minimum competence

Post-tests in the ATT system were designed to certify that the student has reached at least a minimum level of competence. A minimum competence level (MCL) is defined in relation to (1) the particular domain of knowledge and skills that is being trained, and (2) the requirements of the job(s) for which this competence is being required. These requirements are not defined by characteristics of the population distribution of Navy recruits (i.e., the passing score is not defined in reference to the mean of the population of a particular Navy rating), but follows from an absolute standard and, ideally, is determined by domain experts.

The analyses in this paper accept the ATT modules system at face value as a set of work definitions providing (a) a demarcation of the content of various domains of knowledge and skills for which Navy recruits are being trained, and (b) a benchmark for the levels of competence minimally required for different Navy ratings.

The standard of minimal competence for each of the tests following the lessons in the ATT modules and for the post-test scores (i.e., P-tests and D-tests) is determined at 70 on a scale ranging from 1 to 100. A score range of 1 to 100 suggests a fine grain assessment of students' competence; however a criterion-referenced reliability analysis demonstrated a complete failure to differentiate between the criterion of minimum competence and individual performance levels. Ippel and Seals (2008) report Cronbach's alpha coefficients between 0.00 and 0.38.[2]

New MCL measures

The purpose of this investigation is to design and test a new set of MCL measures that describe the probability of passing the criterion of minimum competence (i.e., the passing scores of the post-tests) as some function of the ease with which the student advances from one lesson to the next.

The notion of 'ease of advancing" has several possible interpretations. For example, it can be interpreted as passing at the first trial of each test, including the post-tests. Alternatively, it can mean the total number of trials needed to achieve a passing score on the tests that follow each lesson of an ATT module. Both interpretations require converting the scores of the ATT scoring system into a set of dichotomous variables (pass = 1, fail = 0).

The new score model builds on the distinction between the lesson test scores, which will be referred to as *observed* scores, and the post test scores. The model estimates the probability of passing the criterion scores as an increasing function of the observed scores.

Let $P_k(X=1|X_O)$ be the probability that students with observed score X_O pass a post-test of module k, where X designates a post-test, either a D-test (X_D) or a P-test (X_P). $P_k(X=1|X_O)$ provides a test characteristic function for the post-tests of each module, which specifies that as the observed score, X_O, increases, the probability of a passing score at the post-test (X_D or X_P) of module k increases. A distribution function that is often used in the analysis of dichotomous outcome variables is the logistic distribution function (Hosmer and Lemeshow, 1989). Let π(x) be a shorthand notation for $P_k(X=1|X_O)$, where X equals X_D or X_P. The logistic regression model has a linear form for this probability,

$$\text{Logit}\,[\pi(x)] = \log\big(\pi(x) / (1 - \pi(x))\big) = \alpha + \beta\, x \qquad [1]$$

[2] Ippel and Seals (2008) used a method suggested by Lovett (1977). While reliability is usually defined as the ratio of true variance and observed variance (Lord and Novick, 1968) and the true and observed scores usually are defined in relation to the population mean (i.e., norm-referenced reliability), Lovett (1977) suggests to define both variance components in relation to the passing score (i.e., criterion-referenced reliability).

Subsequently, the logit score obtained with Equation [1] can be transformed into the estimated probability that X = 1 at a fixed value x of X_O by

$$\text{Est. } \pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x)) \qquad [2]$$

For reasons explained elsewhere (see: Ippel and Seals, 2008), we used the first interpretation of "ease of advancing", that is, whether or not a student passed the tests following the lessons of a particular module at the first trial as a predictor for X_D (designated as X_{sumD}). The alternative interpretation, that is, the total number of trials needed to achieve a passing score on the tests that follow each lesson of an ATT module, was used to predict X_P (designated as X_{sumP}) .[3]

[Insert Table 4.a and Table 4.b about here]

Table 4.a. and Table 4.b. show the results of the fit of the linear model (Eq. 1) with $P(X_D=1)$ and $P(X_P=1)$ as dependent variables and X_{sumD} and X_{sumP} as regressors. The linear model fits of $P(X_D)$ with regressor X_{sumD} generally were very good and produced only incidental outliers. The fits of $P(X_P)$ with regressor X_{sumD} were comparable to the first set of measures with one exception: the modeling of the P-test for module 1 failed. The model fit was low and the resulting variable correlated negatively with the other variables. The columns four in each table shows the fits after the outliers were removed from the data.

Subsequently, using Eq [2] the probabilities that the students would pass the criterion of minimum competence at the first trial of a post test were estimated. These estimates were the data for the reliability analysis of the new criterion measures. The final four columns in Tables 4.a. and 4.b. display reliability estimates, which were obtained as follows. The first estimate is based on the domain-sampling model. It estimates the average correlation of the measure with all the measures in the domain. The square root of that estimate is the correlation between the measure and the true score in that domain (i.e., the reliability). The second estimate was based on the common factor model, i.e., it is the ratio between the common variance and the total variance of a measure. The third estimate is an item-total correlation, where the "total" refers to the set of post tests, either D-tests or P-tests. This is not strictly a reliability coefficient. The fourth column displays the average value over the estimates. The median value of the average reliability coefficient of the D-tests was 0.51 and for the P-tests the median value was 0.61.

3 Note, X_D as well as X_P were dichotomous variables measuring whether the student passed the particular post-test at the first trial.

In addition, we estimated the internal consistency (Cronbach's alpha) over all general modules D-tests and P-tests. The results were 0.751, and 0.824, respectively.

Prediction of Criterion Measures

Predictors (1): ASVAB Selection Composite scores

The U.S. Navy uses various selection composites to optimally match available jobs and available talent. In this study we investigate two ASVAB selection composites that the U.S. Navy uses. The first one (ASC01) consists of the four tests comprising the AFQT score, but in a different weighing[4] plus Mechanical Comprehension (MC). The passing score for ASC01 equals 209 over these five tests. This minimum score makes recruits in principle fitting for training for the following ratings: Electricians Mate (EM), Gas Turbine Systems Technician (GSE) and Interior Communication man (IC). The second selection composite (ASC02) consists of an equal weighed linear combination of Math Knowledge (MK), Arithmetic Reasoning (AR), General Science (GS) and Electronics Information (EI). The minimum passing score equals 222 / 223. This minimum score makes recruits in principle fitting for training as Aviation Electrician's Mate (AE), Aviation Electronics Technician (AT) Electronics Technician (ET), Fire Control man (FC) Sonar Technician (STG).

Predictors (2): I.T. Aptitude Battery

In this analysis we investigated the incremental predictive utility of the I.T. Aptitude Battery (ITAB) in predicting training outcomes in the Navy Apprentice Technical Training program. The ITAB is an assessment system that radically differs from existing tests. The purpose of the ITAB tests is to measure the aptitude to learn procedural skills. In order to do that the tests provide a task environment in which the examinee has to develop procedures (or algorithms) to achieve a goal. The test measures how examinees incorporate feedback from the system into their follow-up actions and how quickly this leads to the build up of a more or less efficient algorithm. The test scores reflect the efficiency of these procedures and estimate how much exposure (i.e., training) the examinee would need to be able to develop a maximally efficient procedure.

Two basic innovations are: (1) the test provides an interactive environment, and (2) actions of the examinee are not scored as singleton answers to distinct problems, but are analyzed as sequence patterns.

- Complete interactivity is achieved by creating an internal representation of the task-

[4] AFQT = 2VE + AR + MK, where VE = WK + PC.

environment. Artificial Intelligence technology is used to compute the "intelligence" of each step taken by the examinee.

- Unlike the present generation of computerized tests, the ITAB tests *do not* consist of items with a standard set of response alternatives. Within the task-environment created by the ITAB tests, the examinee is free to act. The examinee must produce sequences of actions to achieve a certain goal.

ITAB 01: Hidden Target Test

Task: The Interface of the Hidden Target test generates a two-dimensional search space consisting of a rectangular grid of equally spaced horizontal and vertical lines. The subject is required to determine where a target is located in as few steps as possible. The subject can move a cursor across the two-dimensional space from junction to junction using a set of arrow keys. To indicate his or her guess regarding the target location the subject clicks at the <TEST> key. The feedback of the task-environment indicates the distance between the tested location and the location of the hidden target in City Block metric. Based upon this feedback the examinee makes his or her next guess until the target has been located.

ITAB 02: Battery Test

Task: The test requires the individual to test whether the presented batteries work. Each new trial starts with a set of nine batteries to be tested. A battery tester is designed such that batteries have to be tested in pairs. The goal of the task is to identify the defective batteries in as few steps as possible. Each set of nine batteries contains two defective batteries.

Analysis

Incremental validity analysis

Increments in validity of ITAB over the ASVAB selection composites were computed as the difference between two validity coefficients (R^2s, the percentages of explained variance by regression models with and without the additional predictor (ITAB)). For each ATT criterion score, the probability associated with this difference was tested using the F distribution with degrees of freedom equal to 1 and N – (1 + 1) – 1, where N equals the number of observations.

$$\Delta R^2 = R^2_{ASVAB\ +ITAB} - R^2_{ASVAB}$$

$$F_{1, N-3} = (N - 1)\left(\frac{\Delta R^2}{1 - R^2_{ASVAB + ITAB}}\right)$$

To reduce the likelihood of Type I errors that results from multiple significance tests, the significance was tested of the incremental validity of the ITAB over the ASVAB Selection Composites in predicting the combined criterion tests for declarative and for procedural knowledge. Significance of the incremental validity for the combined criterion tests, either D-test or P-test, was condition for detailed analysis of the incremental validity of the ITAB in relation to criterion tests of each module separately.

Correction for restriction of range

In different degrees the Navy recruits participating in the ATT program are a selected sample. The first step in the selection process is based on a minimum AFQT score, which determines whether a would-be recruit is sufficiently trainable to join the U.S. Armed Forces. The minimum AFQT score for the Navy is 35 and excludes 33.6 percent of the national youth population from serving in the Armed Forces. The 1997 National Youth Population corrected for the by U.S. Congress defined AFQT lower bound of 35 for the U.S. Navy served as the reference population to determine the magnitude of the effects of restriction of range in the various samples in our study.

The additional selection effect of ASC01 was minimal, less than 0.05 percent. ASC02 excludes an extra 42.5 percent on top of the AFQT selection threshold; only 26.9 percent of the national youth population has a score equal to or higher than 222 on this selection composite.

To account for these selection effects, the sample correlations were corrected using a multivariate procedure based on Lawley (1943) and implemented by Johnson and Ree (1994). Strictly speaking, using a single-score selection composition of ASVAB tests as first predictor exerting an incidental selection effect on the second predictor (i.e., the ITAB) does not represent a multivariate configuration, but an instance of the (univariate) case 3 of Thorndike (1940).

Further corrections of estimates

All reliabilities were corrected for the reduction in variance in the various samples using the following formula:

$$R_{xx} = 1 - (s_x^2 \,/\, S_x^2)\,(1 - r_{xx}),$$

Where r_{xx} is the uncorrected reliability, s_x^2 is the uncorrected population variance, and S_x^2 is the corresponding corrected population value (Gulliksen, 1987).

We followed the convention of upward correcting for (negative) sampling bias using the Wherry (1937) formula to estimate the shrunken coefficients from a single sample:

$$\rho = 1 - [(N - 1) / (N - p - 1)] (1 - r^2),$$

Where ρ is the corrected correlation, N is the sample size, p is the number of predictors and r^2 is the squared multiple correlation.

Finally, the predictor-criterion correlations and multiple Rs were corrected for unreliability in the criterion variables by dividing the correlations by the squared root of the estimated reliability of the criterion.[5]

RESULTS

Reliability Estimates

Tables 4.a and 4.b show the original reliabilities as estimated in the initial sample (N = 2773). Table 5 shows the corrected reliabilities for the same sample and the corrected values for the ratings cluster ASC02 with the cut-off score of 211.

[Insert Table 5 about here]

The values in Table 5 were produced by Johnson and Ree's (1994) program for a multivariate correction for restriction of range. The reliabilities in the column under "Total Sample" were established in the sample of 2773 recruits. The sample was representative with respect to the National Youth Population with AFQT scores equal to or larger than 35. As might have been expected no correction was necessary with respect to the values displayed in Tables 4-1.a and 4-1.b. However, the cut-off score of 221 on the ASVAB Selection Composite ASC02 produced an extremely large downward in estimated variance in the ASC02 sample. This resulted in very low reliability

---

5 The procedure as described in this section has been recommended by a National Sciences committee (Dunbar & Linn, 1991).

estimates for the specific module tests. For this reason we decided not to pursue a test of the sequential hurdle model of selection.

Incremental validity of ITAB over ASVAB Selection Composites

The incremental validity analysis followed the procedure as outlined in the Analysis section. The sample in this analysis was the ASVAB – ITAB – ATT sample, that is, a sample representing the population of reference, only restricted by whether or not the ITAB had been administered. This can be considered a random selection effect, which was not supposed to effect the sample variance. Therefore, no restriction of range correction was applied in the analysis summarized in Tables 6.a. and 6.b. Table 6.a. shows the results of significance testing of the incremental validity of the ITAB over the ASVAB Selection Composites (i.e., ASC01 and ASC02). The ITAB caused an increment of explained variance in all four models tested.

[Insert Table 6.a about here]

The correlation between the ITAB and ASC01 was 0.409 and the correlation with ASC02 was 0.391. Table 6.b. displays the percentages of improvement in predicting criterion variance after correction for attenuation due to unreliability of the criterion.

[Insert Table 6.b about here]

The incremental validity appeared fairly substantial. Notice that this study adopted an approach to incremental validity testing that was more conservative than previously published incremental validity studies with Navy recruits. First, the reference population in this study was *not* the National Youth Sample as in some other studies (e.g., Carey, 1994), but the National Youth Population corrected for the minimum AFQT score required to serve in the U.S. Navy. We considered this to be a more realistic (but also more restrictive) population of reference. Second, we also did not only correct the multiple Rs for unreliability of the criterion (e.g., Wolfe, 1997), but also corrected the single predictor-criterion correlations for these effects. This negatively effected ΔR^2. Finally, the predictions were based on (uncorrected) fallible predictor variables. Correction for attenuation of the predictor scores would not have made much of a difference since both the ASVAB subtests and the ITAB tests are highly reliable tests (Ippel, 2008).

[Insert Table 7.a about here]

Tables 7.a. and 7.b. present the incremental validities, expressed as percentages of improvement of explained variance in criterion scores for the specific ATT modules. Bear in mind that these improvements assume perfectly reliable criterion scores. Notice the extreme improvement of the percentages of explained variance in the Digital Logic Function criterion scores (24-D) in Table 7.b. These were the only predictor-criterion configurations in which the ITAB had a larger contribution in the prediction than the ASVAB Selection Composites. A similar configuration occurred in the prediction of procedural skill criterion scores of the Digital Logic Function (DLF) Module (24-P), albeit less extreme.

[Insert Table 7.b. about here]

Discussion

Educational technology (ET) staff and personnel selection and classification (PS&C) staff often work with different perspectives in mind. The ET perspective on a successful ET design is one that maximizes the number of students that complete a training successfully. For PS&C staff success is assigning educational means to students that can profit maximally with the least effort.

In this pilot study we tested a set of criterion measures that did not impose any alterations upon the extent ET system – the ATT program. However, the measures could with an acceptable level of dependability differentiate between individual accomplishments. The median value of the average reliability coefficient of the D-tests was 0.51 and for the P-tests the median value was 0.61. This is considered and acceptable level of reliability for the purpose of the measures, that is, system evaluation, not individual diagnosis (Nunnally and Bernstein, 1994). Notice that this was a pilot study in which only score distributions were manipulated. Research into the content and method of what is measured could further improve the dependability of these criterion measures.

As far as the prediction of the criterion measures is concerned, the largest contribution in the prediction of Apprentice Technical Training performance can be attributed to the ASVAB Selection Composites (ASC01 and ASC02) that were object of investigation in the study. The notable exception was the module ‘Digital Logical Functions’, the most abstract module of the training program. The ITAB contribution was larger in the prediction of the knowledge test results and consequently the estimate increment in explained variance in the knowledge test variance was very high. Somewhat weaker, but still very large, was the effect on the prediction of the DLF skill test.

The incremental utility effect of the ITAB appeared to be a general effect, that is, it showed in every module of the Apprentice Technical Training program. The effects were fairly substantial. Certainly if compared with similar studies with the Enhanced Computer-Administered Test (ECAT). If anything the studies shows the importance of criterion development and improvement studies.

Acknowledgements

The opinions expressed in this paper are those of the authors, are not official, and should mot be construed as reflecting the views of the U.S. Department of Defense.

Research for this paper was performed under contract number N00189-07-D-Z015.

References

Campbell, J. P., & Knapp, D. J. (2001). *Exploring the Limits in Personnel Selection and Classification,* New Jersey: Lawrence Erlbaum Associates.

Carey, N.B., (1994). Computer predictors of mechanical job performance: Marine corps findings. *Military Psychology,* 6, 1-30.

Gulliksen, H., (1987) Theory of Mental Tests. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hosmer, D.W., & Lemeshow, S. (1989). *Applied Logistic Regression.*New York: John Wiley & Sons.

Ippel, M.J., & Seals, J.S. (2008). Pilot Study Criterion Performance Measures U.S. Navy Advanced Technical Training *CogniMetrics Research Report, CCR. 01.2008.* (49 pp).

Ippel, M.J. (2008). Validity and Predictive Utility of the I.T. Aptitude Battery for U.S. Navy Selection and Classification *CogniMetrics Research Report,* (48 pp).

Johnson, J. T., & Ree, M.J. (1994). Rangej: A pascal program to compute the multivariate correction for range restriction. *Educational and Psychological Measurement*, 54: 693-694.

Lawley, D.N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society Edinburgh, Section A., 62*, 28-30.

Linn, R.L., Baker, E.L., & Dunbar, S.B., Complex, performance-based assessment: expectations and validation criteria, *Educational Researcher,* 20, 8, 15-21.

Lord, F.M. & Novick, M.R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley Publishing Company.

Lovett, H.T. Criterion-Referenced Reliability Estimated by ANOVA, *Educational and Psychological Measurement, 37*, 1, 21-29.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory (3rd Ed.).* New York: McGraw-Hill.

Sellman, W.S. (2004). Predicting Readiness for Military Service. How Enlistment Standards Are Established. National Governing Board.

Thorndike, R.L. (1940). Constancy of the IQ. *Psychological Bulletin*, *37*, 167-186.

Wolfe, J.H., (1997). Incremental validity of ECAT battery factors, *Military Psychology,* 9, 49-76.

Wherry, R.J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlations. *Annals of Mathematical Statistics*, 2, 446-457.

TABLES

Table 1: List of general A.T.T modules, number of lesson per module and type of post test available

	Module	# Lessons	Post Test D	Post Test P
1	Intro to Electricity	7	1	1
2	Multi-meter Measurements	6	1	1
3	Basic DC Circuits	10	1	1
6	Intro to AC	4	1	0
7	AC Test Equipment	4	1	1
12	Transformers	3	1	1
23	Intro to Digital Circuits	6	1	1
24	Digital Logic Functions	8	1	1

Table 2: List of ASVAB tests and their measurement claims

ASVAB tests	Measurement claims
• **General Science (GS)**:	a 25 item knowledge test of physical and biological sciences.
• **Arithmetic Reasoning (AR)**:	a 30 item arithmetic word problem test.
• **Word Knowledge (WK)**:	35 items testing knowledge of words and synonyms.
• **Paragraph Comprehension (PC)**:	15 items testing the ability to extract meaning from short paragraphs.
• **Auto and Shop Information (AS):**	a 25 item knowledge test of automobiles, shop practices, tools and tool use.
• **Mathematical Knowledge (MK):**	a 25 item test of algebra, geometry, fractions, decimals, and exponents.
• **Mechanical Comprehension (MC):**	a 25 item test of mechanical and physical principles and ability to visualize how illustrated objects work.
• **Electronics Information (EI)**:	a 20 item test measuring knowledge about electronics, radio, and electrical principles.
• **Assembling Objects (AO)**:	a 16 item spatial visualization test.

Table 3: Listing of Navy Job Ratings

Seaman (SEA): None Established
Boatswain's Mate (BM): None Established
Quartermaster (QM): VE+AR=96
Damage Controlman (DC): VE+AR+MK+AS=200 or MK+AS+AO=150
Hull Technician (HT): VE+AR+MK+AS=200 or MK+AS+AO=150
Machinary Repairman (MR): VE+AR+MK+AS=200 or MK+AS+AO=150
Mineman (MN): VE+MC+AS=161
Engineman (EN): VE+AR+MK+AS=195 or VE+AR+MK+AO=200
Gas Turbine Systems Technician (GSM): VE+AR+MK+AS=195 or VE+AR+MK+AO=200
Machinist Mate (MM): VE+AR+MK+AS=195 or VE+AR+MK+AO=200
Gunner's Mate (GM): AR+MK+EI+GS=205
Electricians Mate (EM): VE+AR+MK+MC=209
Gas Turbine Systems Technician (GSE): VE+AR+MK+MC=209
Interior Communication man (IC): VE+AR+MK+MC=209
Aviation Electrician's Mate (AE): AR+MK+EI+GS=222 or VE+AR+MK+MC=222
Aviation Electronics Technician (AT): AR+MK+EI+GS=222 or VE+AR+MK+MC=222
Electronics Technician (ET): AR+MK+EI+GS=223
Fire Controlman (FC): AR+MK+EI+GS=223
Sonar Technician (STG): AR+MK+EI+GS=223

Table 4.a: Model Fits and Reliabilities of the new MCL measures of declarative knowledge (N = 2773)

module	model fit			reliability estimates			
	R^2	outliers	R^2 adjusted	domain	ω	r_{it} value	average
Intro to Electricity (1D)	0.91	1	0.96	0.49	0.62	0.41	0.51
Multi-meter Measurements (2D)	0.97	0	0.97	0.52	0.60	0.48	0.53
Basic DC Circuits (3D)	0.92	0	0.92	0.59	0.56	0.58	0.58
Intro to AC (6D)	0.93	1	0.97	0.56	0.45	0.51	0.51
AC Test Equipment (7D)	0.80	0	0.80	0.47	0.23	0.34	0.35
Transformers (12D)	0.99	0	0.99	0.54	0.43	0.44	0.47
Intro to Digital Circuits (23D)	0.88	0	0.88	0.56	0.60	0.48	0.55
Digital Logic Functions (24D)	0.84	1	0.83	0.49	0.54	0.35	0.46

Table 4.b: Model fits and reliabilities for the new MCL measures of procedural knowledge (N = 2773)

module	model fit			reliability estimates			
	R^2	outliers	R^2 adjusted	domain	ω	r_{it} value	average
Intro to Electricity (1P)	0.12	4	0.58				
Multi-meter Measurements (2P)	0.76	1	0.86	0.63	0.45	0.52	0.53
Basic DC Circuits (3P)	0.47	6	0.72	0.68	0.58	0.64	0.63
Intro to AC (6P) *)	No test available						
AC Test Equipment (7P)	0.38	1	0.94	0.67	0.57	0.62	0.62
Transformers (12P)	0.87	1	0.95	0.68	0.58	0.63	0.63
Intro to Digital Circuits (23P)	0.52	4	0.94	0.67	0.56	0.59	0.61
Digital Logic Functions (24P)	0.67	4	0.83	0.64	0.46	0.53	0.54

*) no Post Test available

Table 5: Reliability estimates of criterion scores in the reference population based on a sample of N = 2773 and corrected reliability estimates in a sample with restricted variance due to selection on ASC02 with cut off score at 221 (N = 189)

Module	Variable	Selection	
		Total Sample (N=2773)	ASC02 (cut-off score =221)
D-Test (combined)		0.75	0.83
P-Test (combined)		0.82	0.83
Intro to Electricity	1D	0.51	0.16
Multi-meter Measurements	2D	0.53	0.19
Multi-meter Measurements	2P	0.53	0.24
Basic DC Circuits	3D	0.58	0.40
Basic DC Circuits	3P	0.63	0.41
Intro to AC	6D	0.51	0.53
AC Test Equipment	7D	0.35	0.26
AC Test Equipment	7P	0.62	0.29
Transformers	12D	0.47	0.23
Transformers	12P	0.63	0.42
Intro to Digital Circuits	23D	0.55	0.27
Intro to Digital Circuits	23P	0.61	0.27
Digital Logic Functions	24D	0.46	neg. estim.
Digital Logic Functions	24P	0.54	neg. estim.

Table 6.a: Significance tests of the incremental validity of the ITAB over two ASVAB Selection Composite scores (ASC01 and ASC02) in prediction of combined Apprentice Technical Training (ATT) criterion scores [6]

Selection Composite	modules (C)	N	R predictors		Multiple R			Incremental Validity		
			ASC..	ITAB	R^2	F	sign.	ΔR^2	F	sign.
ASC01										
	D-Tests	391	0.301	0.203	0.104	22.60	p<.001	0.014	6.05	p<.05
	P-Tests	384	0.315	0.210	0.114	24.54	p<.001	0.015	6.43	p<.05
ASC02										
	D-Tests	399	0.297	0.214	0.107	23.66	p<.001	0.019	8.40	p<.01
	P-Tests	392	0.324	0.222	0.124	27.03	p<.001	0.019	8.42	p<.01

[6] D-test and T-test refer to the combined (declarative) knowledge tests and (procedural) skill tests of the ATT modules, respectively.

Table 6.b: Incremental validity of the ITAB over two ASVAB Selection Composites (ASC01 and ASC02) expressed as increases in percentages of explained variance in ATT combined criterion scores [7]

Selection Composite	N	criterion variables name	criterion variables Reliability	R predictors ASC..	R predictors ITAB	(CMR) R^2	Incremental Validity ΔR^2	Incremental Validity % Improvement
ASC01								
	391	**D-Tests**	0.751	0.347	0.234	0.415	0.295	2.44%
	384	**P-Tests**	0.824	0.347	0.231	0.395	0.275	2.28%
ASC02								
	399	**D-Tests**	0.751	0.343	0.247	0.417	0.300	2.55%
	392	**P-Tests**	0.824	0.357	0.245	0.410	0.283	2.22%

[7] D-test and T-test refer to the combined (declarative) knowledge tests and (procedural) skill tests of the ATT modules, respectively; predictor–criterion correlations were corrected for criterion unreliability. The R^2 value was corrected for sampling shrinkage and for unreliability in the criterion.

Table 7.a: Incremental validities of ITAB over ASVAB Selection Composites (ASC) in a sample with unrestricted variance. Incremental validity is expressed as increase in percentages of explained criterion variance. Criterion variables are the declarative scores of ATT modules

Module	N	criterion variables		R predictors		(CMR)	Incremental Validity	
		name	Reliability	ASC	ITAB	R²	ΔR²	% Improvement
ASC01								
Intro to Electricity	539	**1-D**	0.51	0.238	0.158	0.340	0.284	5.0%
Multi-meter Measurements	539	**2-D**	0.53	0.310	0.154	0.423	0.327	3.4%
Basic DC Circuits	530	**3-D**	0.58	0.456	0.285	0.617	0.410	2.0%
Intro to AC	539	**6-D**	0.51	0.326	0.158	0.419	0.312	2.9%
AC Test Equipment	539	**7-D**	0.35	0.196	0.137	0.337	0.299	7.8%
Transformers	539	**12-D**	0.47	0.299	0.169	0.438	0.349	3.9%
Intro to Digital Circuits	539	**23-D**	0.55	0.135	0.030	0.165	0.147	8.1%
Digital Logic Functions	539	**24-D**	0.46	0.003	0.088	0.105	0.105	12124.5%
ASC02								
Intro to Electricity	548	**1-D**	0.51	0.269	0.172	0.385	0.312	4.3%
Multi-meter Measurements	548	**2-D**	0.53	0.334	0.162	0.456	0.345	3.1%
Basic DC Circuits	539	**3-D**	0.58	0.465	0.293	0.632	0.416	1.9%
Intro to AC	548	**6-D**	0.51	0.416	0.181	0.579	0.406	2.3%
AC Test Equipment	548	**7-D**	0.35	0.227	0.147	0.388	0.337	6.6%
Transformers	548	**12-D**	0.47	0.270	0.190	0.411	0.338	4.6%
Intro to Digital Circuits	548	**23-D**	0.55	0.125	0.038	0.149	0.133	8.5%
Digital Logic Functions	548	**24-D**	0.46	-0.037	0.074	0.118	0.117	85.9%

Table 7.b: Incremental validities of ITAB over ASVAB Selection Composites (ASC01 and ASC02) in a sample with unrestricted variance. Incremental validity is expressed as increase in percentages of explained criterion variance. Criterion variables are the procedural skill scores of ATT modules

Module	N	criterion variables		R predictors		(CMR)	Incremental Validity	
		name	Reliability	ASC..	ITAB	R^2	ΔR^2	% Improvement
ASC01								
Intro to Electricity								
Multi-meter Measurements	539	**2-P**	0.53	0.272	0.158	0.375	0.301	4.1%
Basic DC Circuits	530	**3-P**	0.63	0.420	0.275	0.549	0.373	2.1%
Intro to AC								
AC Test Equipment	516	**7-P**	0.62	0.259	0.135	0.325	0.258	3.8%
Transformers	539	**12-P**	0.63	0.272	0.241	0.386	0.312	4.2%
Intro to Digital Circuits	524	**23-P**	0.61	0.270	0.146	0.344	0.271	3.7%
Digital Logic Functions	503	**24-P**	0.54	0.144	0.128	0.208	0.188	9.0%
ASC02								
Intro to Electricity								
Multi-meter Measurements	548	**2-P**	0.53	0.293	0.166	0.404	0.319	3.7%
Basic DC Circuits	539	**3-P**	0.63	0.446	0.283	0.582	0.383	1.9%
Intro to AC								
AC Test Equipment	523	**7-P**	0.62	0.260	0.141	0.328	0.260	3.8%
Transformers	548	**12-P**	0.63	0.282	0.252	0.403	0.324	4.1%
Intro to Digital Circuits	533	**23-P**	0.61	0.282	0.160	0.362	0.283	3.6%
Digital Logic Functions	510	**24-P**	0.54	0.108	0.132	0.181	0.170	14.7%