Validity and Predictive Utility of the

I.T. Aptitude Battery

For

U.S. Navy Selection and Classification

Contract number:	**USGS-296600**
Principal Investigator:	**Martin J. Ippel, Ph.D.**
Institute:	**CogniMetrics Inc.**
Telephone:	**(210) 481 – 0261**
Web site:	**www.cognimetrics.com**
Published:	**December 18, 2008**

CogniMetrics, Inc. is a privately owned company specializing in assessment and decision-support technologies involving human performance.

TABLE OF CONTENT

List of Tables

Page:

List of Figures

EXECUTIVE SUMMARY

The U.S. Navy Selection and Classification (CNO N132) has funded two relatively small-scale studies to investigate the usefulness of the I.T. Aptitude Battery (ITAB) as a candidate for inclusion in the Navy's S&C Rating Identification Engine (RIDE). The interest of Navy Selection and Classification (CNO N132) for the ITAB as an option to complement its selection and classification instruments concurred with an initiative in 2004 of the Defense Manpower Data Center (DMDC) to conduct a review of the Armed Services Vocational Aptitude Battery (ASVAB) to determine if changes in ASVAB content and methodology were warranted given the demands on the Armed Forces in the 21st century. The review panel recommended, among other things, *to develop and evaluate one or more non-verbal reasoning tests* (*NVR*) *for inclusion in the ASVAB*. One of the tests reviewed in the subsequent HUMRRO report is the ITAB.

This study involves a comprehensive evaluation of the validity and predictive utility of the ITAB for Navy selection and classification purposes. The ultimate question for this study was whether the ITAB would show incremental validity over the ASVAB in predicting success in Navy technical training.

The ASVAB is a very high standard to be held against. The value of the ASVAB is well proven. It was scientifically developed and validated on thousands of recruits to ensure a fair chance for every enlistee to successfully complete a military career. It has been the principal tool for selection and placement in the U.S. Armed Forces for several decades. The ASVAB is composed of nine different tests measuring verbal, mathematical and technical knowledge. Administration of the ASVAB tests takes several hours.

The ITAB is a new test, first published in 2004; consisting of two tests measuring fluent intelligence in a new manner. The tests are computer-based and completely interactive; it takes on the average 20 minutes in total to have both tests administered.

This report approaches the evaluative task so as to find answers to four questions. We will use these questions as a heading to shortly summarize the outcomes of the study.

1. Are the ITAB scores based on measures that are sufficiently reliable and do they measure individual attributes of sufficient stability?

The first chapter deals in detail with the issue of the structural models that organize the basic variables into summary scores. In the evaluation of the quality of the models the emphasis was

on their stability over 5 independently drawn random samples of Navy recruits. The standard errors in the estimation of the model parameters appeared very small. The reliabilities (Cronbach's alpha index) of the basic variables appeared satisfying. No indications were found that the learning affected seriously the quality of the measurements. The resulting scores, that is, one predictor score and two diagnostics scores, showed high levels of reliability.

2. Do the ITAB scores reflect individual differences in effectiveness of underlying processes of fluent intelligence?

The ITAB tests are radically different from conventional psychometric tests. The question of studying the underlying process that is supposedly being measured could never be directly answered in traditional tests. Chapter 2 explains how this is very well possible with the ITAB tests. However, thus far no large-scale efforts have been made to answer this question.

3. How do ITAB scores relate to the ASVAB?

The ASVAB is not a comprehensive test of cognitive abilities. It basically measures general knowledge and crystallized intelligence. Therefore, it cannot provide a direct basis for confirmation or refutation of measurement claims in the domain of general and fluid intelligence. Chapter 3 reports that the ITAB confirmed the expectation of exclusively loading on the general factor. The magnitude of the g-factor loading was in line with expectations for a test of fluid intelligence on a factor comprising the common variance of tests of common knowledge and crystallized intelligence.

4. How useful are the ITAB tests in predicting success in the Navy Apprentice Technical Training program? In particular, do the ITAB tests add to the predictive utility of certain ASVAB selection composites?

The ITAB met the challenge of incremental validity over the ASVAB with flying colors. The conclusions of Chapter 4 are:

- All tests of incremental validity of the ITAB over two ASVAB selection composites for the prediction of success in technical training at Navy schools were significant. The estimates were made under assumptions more conservative than in comparable studies.
- The estimates of the percentages of improvement in predictive validity were made with corrections for criterion unreliability. This is a recommended procedure (see footnote 9) and standard practice. The estimates were substantially higher than in comparable studies.
- The improvement effects of predictive utility were very general. It was shown to occur in all training modules under study. At the same time, the effect were the largest for modules that require a relatively high level abstract thinking (i.e., Digital Logic Functions).

- The ASVAB is a very high standard to be held against. Most experts would give any well-designed test low odds on improving the predictive utility of the ASVAB. We believe that ITAB could, because the ITAB is designed to measure the aptitude to learn procedural skills and makes optimal use of information technology to realize that goal. Conventional tests in the domain of cognition, such as the ASVAB, were not designed to measure procedural skill learning. Even when technical knowledge is the measurement objective, it is being treated as declarative knowledge (e.g., ASVAB subtests: Auto/Shop, Mechanical Comprehension, Electronics Information).

INTRODUCTION

Since 2006 U.S. Navy Selection and Classification (CNO N132) has funded two relatively small-scale studies to investigate the usefulness of the I.T. Aptitude Battery (ITAB) as a candidate for inclusion in the Navy's S&C Rating Identification Engine (RIDE).[1]

The interest of Navy Selection and Classification (CNO N132) for the ITAB as a possible test to addition to its selection and classification instruments concurred with an initiative in 2004 of the Defense Manpower Data Center (DMDC) to conduct a review of the Armed Services Vocational Aptitude Battery (ASVAB) to determine if changes in ASVAB content and methodology were warranted given the demands on the Armed Forces in the 21st century. The review panel recommended, among others, *to develop and evaluate one or more non-verbal reasoning tests* (*NVR*) *for inclusion in the ASVAB*. One of the tests reviewed in the subsequent HUMRRO report (Waters, Russell, and Sellman, 2007) is the ITAB.

The purpose of the first of the two studies initiated by Navy Selection and Classification (CNO N132) was to answer two questions. First, how does the ITAB, a set of two tests of fluid intelligence, fit in with the ASVAB factor structure? Second, do the ASVAB and ITAB have invariant meaning across the entire range of scores? The latter question is of great practical importance. It addresses the issue of whether the population of interest in Department of Defense (DoD) selection and classification models should necessarily be the general population of applicants for the Air Force, Army Corps, or Navy, with a mean AFQT score of 50 and a standard deviation of 10. Intuitively it is hard to imagine that a candidate for the highly complex training course for Navy Electronics Technician-Submarine (ET SS_N) can be considered sampled from the same population of interest as a candidate for the Navy Engine Man (EN) course. Recent studies with the ASVAB and other general cognitive ability tests suggest a lower involvement of the g factor in test performance at higher levels of cognitive ability.

The purpose of the second of the two studies initiated by Navy Selection and Classification (CNO N132) was to investigate whether inclusion of the ITAB tests would improve the predictive

---

[1] Contract number: N0014005C1566-05 title: Measurement Invariance of ASVAB and ITAB in three clusters of Navy Ratings.
Contract number: N00189-07-D-Z015; title: Pilot Study Criterion Performance Measures U.S. Navy Advanced Technical Training.

utility of the RIDE for various Navy technical training programs. Navy technical schools provide training in procedural skills of all sorts and levels of complexity. Traditional tests of cognitive abilities such as the ASVAB were not designed to measure aptitude for procedural skill learning. Even when technical knowledge is the measurement objective, it is being treated as declarative knowledge (e.g., ASVAB tests: Auto/Shop, Mechanical Comprehension, and Electronics Information). The ITAB tests measure the basic mechanisms of procedural skill learning directly without assuming domain-specific knowledge such as programming knowledge or mechanical knowledge (Ippel & Zaal, 2004).

The present study involves a comprehensive evaluation of the validity and predictive utility of the ITAB for Navy selection and classification purposes. The validity of a test score is the evidence in support of its measurement claims in the populations for which the test is used. The ITAB tests claim to measure fluid intelligence. In this report we will present and discuss the evidence in support of the ITAB as a measure of fluid intelligence. In particular, we will substantiate this claim in four different chapters demonstrating that the ITAB scores:

I. are based on variables with sufficient homogeneity to be considered indicators of a single attribute (i.e., evidence of reliability of measurements);
II. reflect individual differences in the effectiveness of the search control cycle underlying test performance (i.e., evidence of the process);
III. relate to other variables consistent with theoretical expectations (i.e., external relations as evidence for measurement claims);

A separate issue is the predictive utility (McDonald, 1999) of the ITAB scores, that is, the usefulness of the tests scores in predicting criterion variables of practical relevance. For example, school achievements, outcomes of technical training, job performance. Chapter IV of this report involves the evaluation of the contribution of the ITAB in predicting success in the Navy Apprentice Technical Training program (ATT) at the Great Lakes Recruit Training Command in Chicago, IL.

Chapter 1
EVIDENCE OF DEPENDABILITY OF ITAB BASIC VARIABLES

Innovative features

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-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.

Thus, the ITAB tests are drastically different from existing psychometric test technology. Each ITAB test provides an interactive task environment. An examinee action (user action) changes this task environment (system action) and the examinee receives feedback information about the new state of the task environment (system output). The cognitive diagnosis component of the tests measures these changes per user action and analyzes the sequential aspects of it (Markov model) and the informational content of each new system state.

An interactive test generates a multitude of information. To structure the many measurements generated by the tests, the diagnostic components organizes them into three scores the meaning of which is intuitively accessible, viz., INTELLIGENCE (efficiency in extracting information) and WORK STYLE (measuring whether an individual follows a steady approach and is careful in his/her inferences) and ITAB_B (a multifaceted performance score for fluid intelligence). The scientific basis for the structuring of test information in this way is an empirically tested structural equations model. Figure 1 displays the LSE-model for diagnostic structure of the Battery Test. A similar model was developed and tested for the Hidden Target Test.

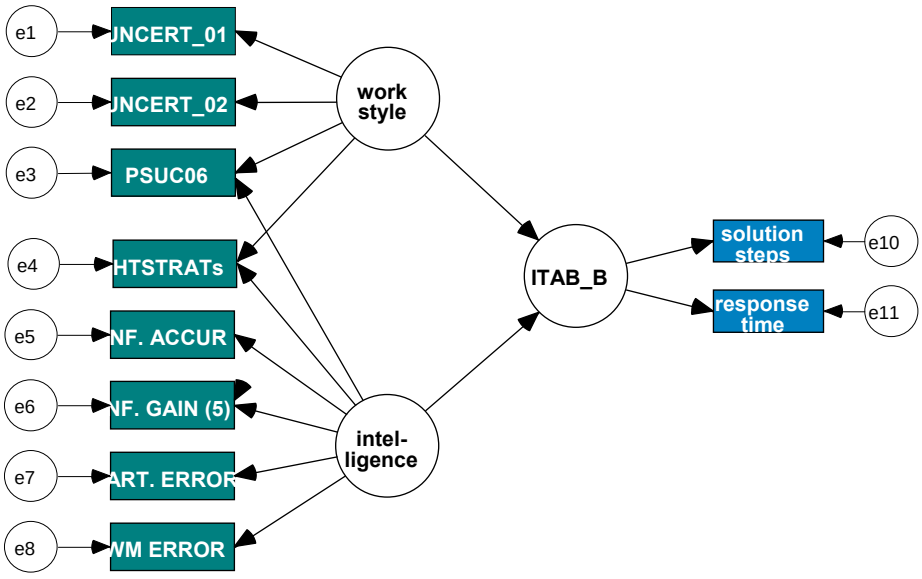

Figure 1-1: Diagnostic structure of the Battery Test

The status of the three scores is different from a practitioner's point of view. The ITAB score is a predictive score, whereas INTELLIGENCE and WORKSTYLE have diagnostic purposes.

Three types of variables

Figure 1 distinguishes three categories of variables by different colors (white, blue and green). The white colored variables are the latent variables ITAB_B (or ITAB_H in case of the HTT), IN-

TELLIGENCE and WORK STYLE. The blue color indicates the multiple indicators for the latent variable ITAB (i.e., the predictor score). The green color denotes the multiple indicators for the latent variables INTELLIGENCE (INT) and WORK STYLE (WS) (i.e., the diagnostic scores). The character of the green colored indicators differs from the blue colored indicators. The blue colored indicators are directly observable, viz., number of steps to solution (STS) and a response time measure, the average time per step in the solution process.

The values of the green colored indicators are the outcome of more complex theory-based calculations. For example, UNCERT_01 and UNCERT_02 are two parallel tests measuring the residual uncertainty at the moment the individual decides that a pair of batteries is defective. This value is the outcome of a theoretical analysis of the steps in the solution process and is expressed in the metrics of mathematical information theory (binary units, or bits, of information) (Ash, 1965; Edwards, 1964, Garner, 1962; Krippendorff, 1982). Another example, PSUC6 indicates the probability of a successful solution in 6 steps or less for the individual. This value is based on a Markov analysis of the total sequence of solution steps during test taking.

The previous two examples are from the diagnostic structure of the Battery Test. The next example is from the diagnostic structure of the Hidden Target Test. This structure includes the variables URAT_3 and URAT_4. These are two parallel tests measuring the mean information loss per solution step if that step was less than maximally "intelligent". This information loss is also expressed in bits of information (per step).

In summary, the diagnostic structures of the tests are identical at the level of the latent variables and the outcome measures. This structural isomorphism is the basis for combining corresponding latent variable scores into a single score. For example, the ITAB_B score and the ITAB_H score are combined into the ITAB score. The same holds for the INT_B score and INT_H score and the WS_B score and WS_H score.

METHOD

Goal of the study

The ITAB scores are latent trait scores which are derived from a total of ten more basic variables per test. Each of those variables aggregates information collected during the interaction of the examinee with the test environment. The following analyses investigated in a bottom up fashion (1) whether observations subsumed under the same basic measure are sufficiently homogeneous to be considered indicators of a single attribute; (2) whether these attributes change as a result of interaction with the test environment; (3) the stability of the structural relations between the variables over different samples.

INSTRUMENTS

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.

Table 1-1: Basic Variables of the Hidden Target Test

Variables	Description
STS_H	expected value number of steps to solution. (Markov Analysis).
TEMP_H	mean time per solution step. Measured per trial (20 trials).
mcons_3	mean consistency score (odd trials).
mcons_4	mean consistency score (even trials).
gtime	expected value holding time guessing state (Markov Analysis).
pstime	expected value holding time problem solving state (Markov Analysis).
h_ipcap3	information processing resources employed; expressed as percentage of maximum information processing possible (odd trials).
h_ipcap4	idem (even trials).
urat_03	utilization of heuristic information (expressed as bits of information per solution step) (odd trials).
urat_04	idem (even trials).

Diagnostic structure: Table 1-1 presents the basic variables and a short description. Together these variables are part of a linear structural equations (LSE) model that turns the basic measurements into three test scores. Figure 1-2 presents a graphic representation of the LSE model of the Hidden Target Test. Figure 1-2 incorporates all constraints on model parameters either graphically (an arrow implies a relationship between two variables, the direction of the arrow implies a causal direction and bidirectional arrow indicate a mere correlation; no arrow implies a zero correlation) or symbolically (e.g., when two variables are accompanied by a symbol v with identical numerical suffix, say, v1, this mean that their variances were constrained to be

equal; the symbol a with identical numerical suffix, say, a1 implies identical intercepts; if two arrows are accompanied by symbol b with identical numerical suffix, say, b1, this implies that two regression coefficients were constrained to be equal. Each latent variable comes with two parameter values placed right above it, the first one indicating a value for the mean and the second one indicates the variance. In the LSE models of both ITAB tests the latent variables INTELLIGENCE and WORKSTYLE were standardized, that is, the mean was set equal to zero and the variance was set equal to one.

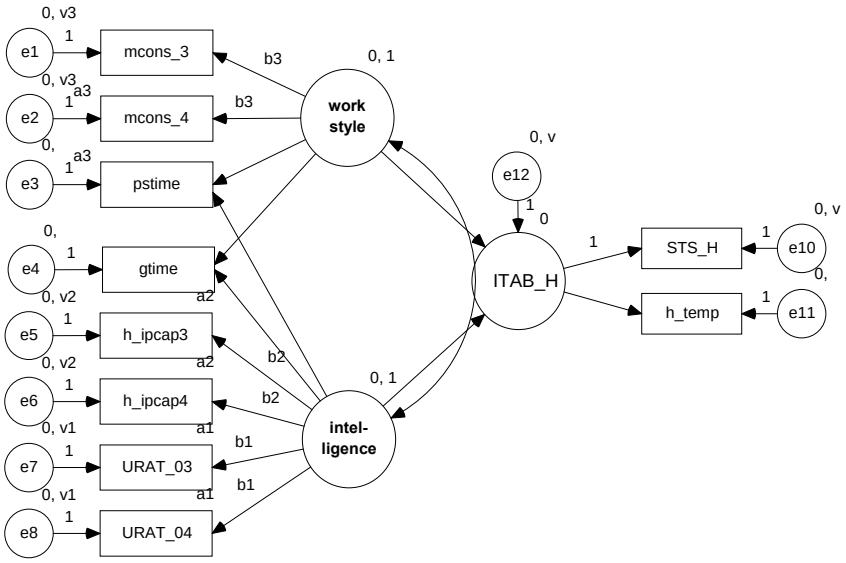

Figure 1-2: Graphic representation of the Hidden Target Test LSE-model

Notice that three pairs of indicators can be considered strict parallel tests with equal means and equal variances, viz., MCONS_3, MCONS_4, H_IPCAP3, H_IPCAP4, and URAT_03, URAT_04. In addition, their respective regression coefficients to the latent variables were constrained to be equal. An initial version of the model had the latent variables INTELLIGENCE and WORKSTYLE uncorrelated, but this constrained had to be relaxed in confrontation with empirical data.

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.

Table 1-2: Basic variables of the Battery Test

Variables	Description
STS_B	expected value number of steps to solution (Markov Analysis).
TEMP_B	mean time per solution step (20 trials).
MN_H_03	mean uncertainty (H) at first rejection decision (expressed in bits of information) (odd trials).
MN_H_04	idem (even trials).
PSUC06	Probability of solving problem in 6 steps or less (Markov Analysis).
REPTAPPL	sum total of expected holding times for partitioning and testing approaches expressed as percentages of STS_B.
MN_INFEX	mean number of problem solving steps after H = 0 in fact has been reached.
H_RED05	mean reduction of uncertainty (H) at step 5.
MN_NE1	mean number of partitioning errors (Type 1 error: no uncertainty reduction).
MN_NEM	mean number of tests performed on the same pair of batteries.

Diagnostic Structure:

The LSE model of Figure 1-3 represents the diagnostic structure of the Battery Tests and has a few special features that should be noticed. First, the two (diagnostic) latent variables, that is, INTELLIGENCE and WORKSTYLE are independent. This is a valuable feature for practical use of the test battery. Second, notice the two variables top left in Figure 1-2, namely, MN_HR03 and MN_HR04. These variables measure the residual uncertainty (H) in the problem solving situation at the moment the examinee takes the decision to reject the batteries that as being defective. Ideally, the residual uncertainty equals zero. To ensure that this important characteristic would strongly determine the WORKSTYLE latent variable, it was represented by two strict parallel tests, viz., MN_HR03 and MN_HR04. Figure 1-2 shows the means and variances constrained to be equal. In addition, the regressions to the latent variable WORKSTYLE were constrained to be equal.

SAMPLING

Participants were 2609 U.S. Navy recruits from schools with a focus on technical ratings in the Engineering and Combat Systems areas of the Great Lakes Navy training facility near Chicago, IL. The recruits were administered two tests of the I.T. Aptitude Battery (ITAB), viz., the Hidden Target Test (HTT) and the Battery Test (BT). Due to initial technical problems related to the firewall at the Chicago's Great Lakes Recruit Training Command not all recruits completed both tests. We were able to collect complete test protocols of both tests of a total of 1420 recruits. In

this analysis 5 independent and random samples of 300 recruits each were drawn from this grand sample.

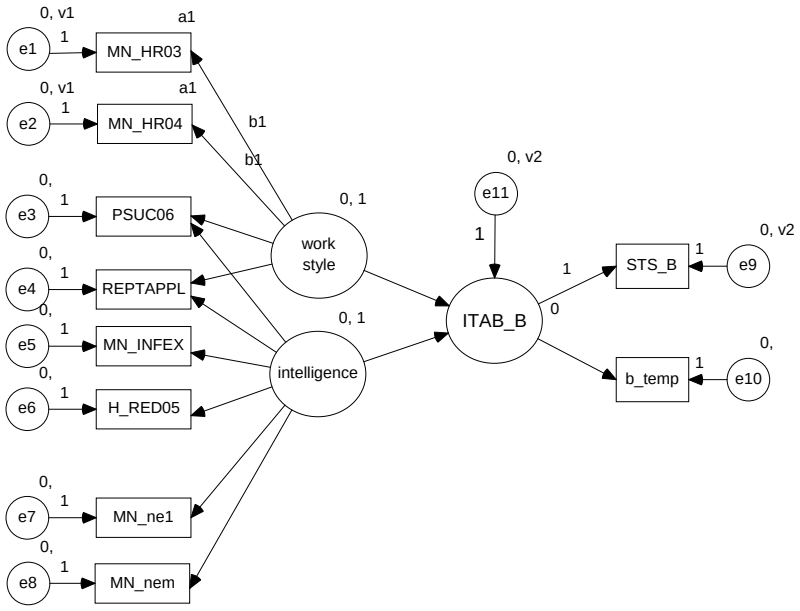

Figure 1-3: Graphic representation of the Battery Test LSE-model

ANALYSES

The first set of analyses involved the estimation of factor loadings and factor score coefficients. Although a confirmatory factor analysis method was used, the purpose of the analyses was not to test the linear structural equations (LSE) models, but to use a previously tested model to derive precise estimates of the loadings of the latent variables on the empirical indicators and the corresponding factor score coefficients. This was accomplished by fitting the LSE model representation of the diagnostic structure of each ITAB test to the data of five different random samples of 300 recruits and analyzing the means and variances of the resulting sampling distributions (for samples with N = 300) of these estimates. The sample means were used to calculate pooled estimates of the factor score weights for the latent variables.

The second set of analyses concerned the dependability of the empirical indicators of the LSE models. Under certain assumptions the reliability of the indicators can be directly derived from the LSE models (McDonald, 1999). That is why these analyses are presented first. In the second set of analyses we compared homogeneity estimates (Cronbach's alpha), which come conceptually closest to the omega estimates based on the LSE models, with split-half reliabilities where

the sets of empirical indicators was split into a first half (trials 1 – 10) and second half (trials 11 – 20). These split-half coefficients were referred to as S-H (F/S) and into sets of odd and even trials (referred to as S-H (O/E). Subsequently, the reliabilities for the various latent variable scores (or factor scores) were estimated.

Finally, the factors of the HTT (i.e., INT_H, WS_H, ITAB_H), BT (i.e., INT_B, WS_B, ITAB_B) and the ITAB (i.e., INT, WS, ITAB) were correlated to arrive at pooled estimates of the structural relations.

RESULTS AND DISCUSSION

Hidden Target Test

The LSE model of the Hidden Target Test (see Figure 1-2) was fitted to the data of five random samples (N = 300) of Navy recruits who had completed both ITAB tests. The model fits for the HTT LSE model were moderate (see Appendix). The Modification Indices did not provide any meaningful suggestion to further improve model fits. A salient feature of the best fitting model was that the covariance between INTELLIGENCE and WORKSTYLE could not be constrained to be zero. The mean correlation between these latent variables over the five samples was 0.44 (range: 0.38 – 0.49).

Table 1-3 presents the means and standard errors of the sampling samples of each of the model parameters distributions for N = 300 samples. Note that the factor loadings in Table 1-3 are expressed in standardized form.

Despite the moderate model fits the standard errors of estimation were very small and most indicators were well explained by the LSE model, that is, with two exceptions, all indicators had a H^2 larger than 0.60. The first exception was PSTIME; a measure resulting from a Markov analysis of the sequences of solution steps of an examinee. The HTT Markov model distinguishes three mental states, viz., guessing (G), problem solving (PS), and problem solved (S). The latter state is the terminal state. Once the examinee has reached state S the process terminates. PSTIME is the expected value of the holding time of the PS state, that is, the number of times in succession that the PS state is occupied after it is first entered. In general, the PS state is left, if the examinee either has reached the S state, or fell back into guessing (G state). H_TEMP was the second empirical indicator with a H^2 value lower than 0.60. H_TEMP measures the estimated time per solution step during the solution process.

Further, notice that INTELLIGENCE (INT_H), which indicates the efficiency of information processing, was roughly twice as important for the ITAB_H score as WORKSTYLE (WS_H).

Table 1-3: Means and standard errors (SE) of estimated factor loadings and communality (H^2) estimates of the Hidden Target Test LSE model over five samples of Navy recruits (each sample with N = 300)

Indicators	Factors							
	ITAB_H		INT_H		WS_H		H^2	
	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
ITAB_H			0.766	(0.003)	-0.562	(0.004)	0.624	(0.005)
STS_H	0.746	(0.002)					0.668	(0.003)
H_TEMP	-0.347	(0.003)					0.147	(0.003)
MCONS_3					0.723	(0.000)	0.627	(0.001)
MCONS_4					0.723	(0.000)	0.627	(0.001)
PSTIME			-0.403	(0.005)	0.254	(0.004)	0.173	(0.005)
GTIME			0.749	(0.002)	0.026	(0.006)	0.702	(0.004)
URAT_03			0.782	(0.001)			0.734	(0.001)
URAT_04			0.782	(0.001)			0.734	(0.001)
H_IPCAP3			-0.701	(0.001)			0.591	(0.002)
H_IPCAP4			-0.701	(0.001)			0.591	(0.002)

Battery Test

The LSE model of the Battery Test (see Figure 1-3) was fitted to the data of five random samples (N = 300) of Navy recruits who had completed both ITAB tests. Table 1-4 presents the means and standard errors of the sampling distributions of each of the model parameters. Note that the factor loadings in Table 1-4 are expressed in standardized form.

The model fits for the Battery Test LSE model were better, but still at a moderate level. This time the correlation between INTELLIGENCE and WORKSTYLE could be constrained to be zero loss of quality of fit.

Pooled estimates of factor score weights

The results of the factor analyses showed remarkable stability across the five samples of Navy recruits. Tables 1-5 and 1-6 show means and standard errors of sampling distributions of the factor score weights for both tests. Here the conclusion is the same; the sampling means were characterized by small standard errors. It was therefore decided to derive pooled estimates of these weights to calculate factor scores. In further analyses the ITAB scores ITAB_H, INT_H,

WS_H and ITAB_B, INT_B, and WS_B were based on those factor score weights. After transformation of these scores to variables with a mean of 50 and standard deviation of 10, the ultimate ITAB scores, viz., ITAB, INT and WS were designed as means over corresponding scores and again transformed into variables with means of 50 and standard deviation of 10. Table 1-7 presents the means and standard deviations in the 5 samples.

Table 1-4: Means and standard errors (SE) of estimated factor loadings and communality (H^2) estimates of the Battery Test LSE model over five samples of Navy recruits (each sample with N = 300)

Indicators	Factors							
	ITAB_B		INT_B		WS_B		H^2	
	Mean	(S.E.)	Mean	(S.E.)	Mean	(S.E.)	Mean	(S.E.)
ITAB_B			0.762	(0.002)	0.365	(0.007)	0.980	(0.001)
STS_B	0.990	(0.000)					0.981	(0.001)
B_TEMP	0.178	(0.003)					0.034	(0.001)
MN_HR01					0.866	(0.002)	0.751	(0.004)
MN_HR02					0.866	(0.002)	0.751	(0.004)
MN_NEM			0.745	(0.001)			0.555	(0.001)
PSUC06			-0.885	(0.001)	-0.313	(0.005)	0.887	(0.001)
H_RED05			-0.335	(0.002)			0.113	(0.001)
MN_NE1			0.792	(0.003)			0.630	(0.005)
REPTAPPL			-0.453	(0.002)	-0.357	(0.001)	0.334	(0.001)
MN_INFEX			0.956	(0.001)			0.914	(0.001)

Table 1-5: Means and standard errors of the estimated factor score weights for three latent variables of the Hidden Target Test (5 samples of Navy recruits of N = 300)

Indicators	Factors					
	ITAB_H		INT		WORKSTYLE	
	Mean	S.E.	Mean	S.E.	Mean	S.E.
ENT_H	0.564	0.001	0.027	0.000	-0.073	0.001
H_TEMP	0.095	0.001	-0.004	0.000	0.012	0.000
MCONS_3	0.026	0.000	0.003	0.000	0.041	0.000
MCONS_4	0.026	0.000	0.003	0.000	0.041	0.000
PSTIME	0.261	0.002	-0.103	0.001	0.225	0.002
GTIME	0.078	0.001	0.128	0.002	0.058	0.002
URAT_03	2.706	0.012	3.318	0.007	0.636	0.012
URAT_04	2.706	0.012	3.318	0.007	0.636	0.012
H_IPCAP3	0.002	0.000	-0.002	0.000	-0.001	0.000
H_IPCAP4	0.002	0.000	-0.002	0.000	-0.001	0.000

Table 1-6: Means and standard errors of the estimated factor score weights for three latent variables of the Battery Test (5 samples of Navy recruits of N = 300)

Indicators	Factors					
	ITAB_B		WORKSTYLE		INT	
	Mean	S.E.	Mean	S.E.	Mean	S.E.
B_TEMP	0.018	0.000	0.002	0.000	0.003	0.000
MN_HR01	0.182	0.002	1.017	0.005	-0.261	0.002
MN_HR02	0.182	0.002	1.017	0.005	-0.261	0.002
MN_NEM	0.054	0.000	-0.069	0.001	0.067	0.000
PSUC06	-2.698	0.022	-0.817	0.012	-1.607	0.010
H_RED05	-0.003	0.000	0.003	0.000	-0.003	0.000
MN_NE1	0.068	0.000	-0.092	0.001	0.087	0.001
REPTAPPL	-0.221	0.002	-0.350	0.004	-0.031	0.001
MN_INFEX	0.146	0.002	-0.171	0.001	0.174	0.001

Table 1-7: Transformed factor scores with mean = 50, sd = 10 in 5 samples of Navy recruits (N = 300)

	ITAB_H		WS_H		INT_H		ITAB_B		WS_B		INT_B	
	m	s.d.	m	s.d.	m	s.d.	m	s.d.	m	s.d.	m	s.d.
HBSAMPLE1:	50.4	9.3	50.5	9.7	49.8	9.8	50.3	9.1	50.1	10.1	50.3	9.3
HBSAMPLE2:	50.5	8.1	49.7	9.6	50.9	9.9	49.7	10.8	49.7	11.2	49.9	10.8
HBSAMPLE3:	49.0	11.6	49.4	10.2	49.6	10.5	49.2	11.1	49.8	11.1	49.3	10.6
HBSAMPLE4:	50.5	11.0	50.5	10.4	49.8	9.8	50.4	9.4	50.4	7.3	50.3	10.1
HBSAMPLE5:	49.7	10.1	49.9	10.2	50.0	10.1	50.3	9.6	50.0	10.4	50.4	9.2
Mean:	**50.0**	**10.0**	**50.0**	**10.0**	**50.0**	**10.0**	**50.0**	**10.0**	**50.0**	**10.0**	**50.0**	**10.0**
Standard error	0.7	1.4	0.5	0.4	0.5	0.3	0.5	0.9	0.3	1.6	0.5	0.7

Reliability of the basic variables and factor scores

Tables 1-8 and 1-9 present the reliabilities of the basic measurements that feed the diagnostic structure of the HTT and BT, respectively. The primary tool to assess the reliability of the basic variables in Tables 1-8 and 1-9 is Cronbach's alpha. Cronbach's alpha gives the lower bound of

Table 1-8: Reliability coefficients of the basic variables of the HTT diagnostic structure estimated in 5 different samples of Navy recruits (N = 300 per sample)

Indicators	Reliability						
	Cronbach's Alpha		Split Half				Omega
			F/S[1]		O/E[2]		
	Mean	(S.E.)	Mean	(S.E.)	Mean	(S.E.)	
ENT_H	0.831	0.000	0.686	0.002	0.886	0.001	
H_TEMP	0.961	0.000	0.919	0.001	0.972	0.000	
MCONS_3	0.714	0.001	0.751	0.000	0.857	0.000	
MCONS_4	0.697	0.000					
URAT_03	0.894	0.000	0.901	0.000	0.951	0.000	
URAT_04	0.900	0.000					
H_IPCAP3	0.831	0.001	0.837	0.001	0.907	0.000	
H_IPCAP4	0.826	0.000					
PSTIME							0.173
GTIME							0.702

[1]) F / S = first half / second half [2]) O / E = odd items / even items

the homogeneity of a collection of items designed to measure the same concept. The coefficients presented in the tables are the means over coefficients calculated in five samples of 300 Navy recruits. Most basic variables in the ITAB diagnostic systems are aggregates of observations during a trial. Each test has 20 trials. As a result most scales are based on 20 measurements. Some scales have identical names, only different suffixes, for example, MCONS_3 and MCONS_4. Such scales were designed as (strict) parallel tests to be marker variables to a latent variable (in this case WORKSTYLE: WS_B). The scales of those parallel tests are based on 10 variables instead of 20. Consequently, the reliability estimates of a single marker variable may be slightly lower than for the scales based on 20 items, but they always come in pairs. In those cases the split-half (O / E) coefficient gives a more accurate impression of the reliability by which the characteristic is being measured.

Table 1-9: Reliability coefficients of the basic variables of the BT diagnostic structure estimated in 5 different samples of Navy recruits (N = 300 per sample)

	Reliability						
	Cronbach's Alpha		Split Half				Omega
			F/S		O/E		
	Mean	S.E.	Mean	S.E.	Mean	S.E.	
ENT_B	0.834	0.000	0.778	0.001	0.833	0.000	
B_TEMP	0.956	0.000	0.911	0.000	0.962	0.000	
MN_NEM	0.704	0.001	0.653	0.001	0.699	0.001	
MN_NE1	0.867	0.001	0.820	0.001	0.872	0.000	
MN_INFEX	0.728	0.001	0.666	0.001	0.739	0.001	
MN_HR01	0.739	0.002	0.782	0.002	0.860	0.001	
MN_HR02	0.731	0.002					
PSUC6							0.887
H_RED05							0.113
REPTAPPL							0.334

[1]) F / S = first half / second half [2]) O / E = odd items / even items

A few variables, which were derived from Markov model analyses, were one-shot measures. Their reliabilities estimated based on the LSE models of the diagnostic structures in which they have a place. They can be found under the columns "omega" in each table.

Measurement always influences the individuals being measured. This becomes a problem for measurement when individuals are affected in different degrees. A quick and easy impression of this phenomenon can be obtained by comparing two split-half coefficients, viz., the S-H (O / E) and the S-H (F / S). Experts (e.g., Nunnally and Bernstein, 1994) suggest that one should reckon with this problem if the S-H (F / S) is 20 decimal points or more lower than S-H (O / E).

The reliabilities (alpha coefficients of Tables 1-8 and 1-9 were used to calculate the reliability of the factor score with the formula (Nunnally, 1967):

$$r_{yy} = 1 - (\sum b_k^2 \sigma_k^2 - \sum b_k^2 \sigma_k^2 r_{kk}) / \sigma_y^2 \qquad [1\text{-}1]$$

where r_{yy} designates the reliability of the linear combination (i.e., factor score); b_k = factor score weights k of the variables in the linear combination, k = 1, …, m; m equals the number of variables in the linear combination, σ_k^2 = variance of variable k; r_{kk} = reliability of variable k; σ_y^2 = variance of the linear combination. Table 1-10 summarizes the results.

Table 1-10: Reliabilities of the ITAB (factor) scores

Test	Est. R.
ITAB	
ITAB	0.919
INT	0.942
WS	0.865
Hidden Target Test (HTT)	
ITAB_H	0.905
INT_H	0.960
WS_H	0.874
Battery Test (BT)	
ITAB_B	0.889
INT_B	0.849
WS_B	0.916

Correlations between (factor) scores

Table 1-11 displays the correlations between the ITAB scores. The scores directly associated with the tests are factor scores. The combination scores are linear combination of two corresponding and equally weighted factor scores (weights: ½). Bold face printed correlations show the correspondence and demonstrates that the equal representation of the test scores in the combined scores worked out fairly well.

Table 1-11: Correlations between ITAB scores. Pooled estimates over 5 samples of Navy recruits (N = 300 per sample)

	Combined ITAB tests			Hidden Target Test			Battery Test		
	ITAB	**INT**	**WS**	**ITA-H**	**INT-H**	**WS-H**	**ITAB-B**	**INT-B**	**WS-B**
Combined ITAB tests									
ITAB									
INT	0.831								
WS	0.392	-0.112							
Hidden Target Test									
ITAB-H	**0.786**	0.607	0.260						
INT-H	0.489	**0.731**	-0.278	0.675					
WS-H	0.233	-0.247	**0.718**	0.221	-0.511				
Battery Test									
ITAB-B	**0.791**	0.707	0.355	0.249	0.104	0.144			
INT-B	0.725	**0.730**	0.115	0.214	0.071	0.149	0.933		
WS-B	0.328	0.084	**0.715**	0.156	0.112	0.032	0.359	0.012	

Conclusions

1. The variables that constitute the empirical basis of the ITAB diagnostic systems proofed to be of a sufficient level of homogeneity and measuring stable attributes of human performance. Only one coefficient of homogeneity was below 0.70. That was the case of a shorter parallel form of measurement (MCONS_4). In combination with the other parallel form (MCONS_3), it measured the target attribute with reliability of 0.857 (S-H (O/E)). Comparison of split-half reliabilities gave little cause for worries about the stability during the process of measurement of the attribute being measured (Tables 1-8 and 1-9).

2. Three one-shot empirical indices, which were all resulting from Markov analyses showed extreme low reliability estimates. Notice that these estimates were derived from the LSE models of the diagnostic structure and depended strongly on the measure of common variance of those variables with the other empirical indices of the models (Tables 1-8 and 1-9).
3. The ITAB scores, the combination scores (ITAB, INT, WS) as well as the scores of separate tests (ITAB_H, INT_H, and WS_H; ITAB_B, INT_B, and WS_B) proofed highly reliable (Table 1-10)..
4. The empirical fit of the LSE model representations of the diagnostic structures of the ITAB tests were moderate, but the factor pattern and resulting factor weight coefficients demonstrated a very high degree of stability over 5 samples of Navy recruits (Appendix and Tables 1-3 and 1-4).
5. All factor scores were transformed into scores with a mean of 50 and standard deviation of 10, which prevented dominance of one test over the other in the combined scores (Table 1-11).

Chapter 2
EVIDENCE OF THE OPERATION OF FLUID INTELLIGENCE

A new measure of fluid intelligence

Traditionally, fluid intelligence has been measured using *tests of inductive reasoning*, for example, Raven's Progressive Matrices test. Another example of an inductive reasoning test is Number Series, a test in which a series of numbers is generated according to a rule and the next number is to be identified.

This tradition gave fluid intelligence a flavor of abstract thinking. Reality is different. Fluid intelligence comes into play whenever a human being is confronted with a new situation and has to figure out how to deal with it. This is not necessarily an abstract activity. It can be as concrete as the activity of a toddler learning to catch a ball rebounding from the wall, or a naval navigator figuring out how a new GPS device works. If anything, fluid intelligence often manifests itself as *practical intelligence*. A test of fluid intelligence should reflect the fundamental processes underlying these activities. That is what the tests of the I.T. Aptitude Battery (ITAB) do.

The design of the ITAB tests is based on a research paradigm that first has been proposed by Newel and Simon in their seminal work on "Human Problem Solving" (1972). The core assumption of this paradigm is the *problem space hypothesis*—all human cognition involves search through some problem space (see also Newell, 1981, 1990).[2] This research paradigm makes a few minimal assumptions about the control structure underlying human task performance - a task control structure consisting of two elements: (1) a *problem space* as the internal representation of the task-environment, and (2) *search control knowledge*, which helps solving a problem by finding a path from the initial state to the goal state. This search can be formally represented as the repeated application of a search control cycle - observe the current state, provide some action input, and observe the effect.

In the conception of fluid intelligence underlying the ITAB design this control cycle is the engine

[2] Artificial Intelligence (AI) researchers as well as cognition psychologists (often) model thinking as searching through a state space – a finite set of states, including an initial state and one or more goal states and a finite set of operators to transform one state into another. In psychology this state space is referred to as a problem space (e.g., Holland, Holyoak, Nisbett and Thagard, 1988; Newell and Simon, 1972; Newell, 1990, 1991; VanLehn, 1990, 1991).

of the operation of fluid intelligence: observe the current state of the object (e.g., a ball), provide some action input (e.g., throw it against the wall) and observe the effect. The ITAB tests were designed to measure the effectiveness of this control cycle.

Fluid intelligence and procedural skill learning

In this report the ITAB tests will be sometimes referred to as tests of fluid intelligence and sometimes as tests of the aptitude to learn procedural skills. This may sound erroneous, but it is not. The ITAB test design is based on a different paradigm for understanding the relationship of problem solving and procedural skill learning than conventional tests. This different approach to understanding the relationship between problem solving and cognitive skill learning is formulated by cognition scientists in the form of a computational theory (e.g., Anderson, 1983, 1987, 1989, 1993; Newell and Simon, 1972; Newell, 1990, 1991; VanLehn, 1990).[3] In this approach certain problem solving techniques (i.e., the core of fluid intelligence) are considered basic elements of procedural learning (Ippel & Zaal, 2004).

ITAB data streams

Each of the ITAB tests generates an interactive task environment (see Chapter 1). The interaction between examinee and test system can be described as a cycle of actions of the examinee (user actions), which cause changes in the environment (system actions) and from which the examinee receives feedback (system output). In response to this feedback the examinee produces a follow-up action. In this way, user actions, system actions, system output and their temporal aspects form a continuous data stream. A specific feature of the ITAB tests is that the data stream is recoded into two separate codes. The first is a so-called Markov Chain code (analysis of action sequences) and the second is a so-called Artificial Intelligence code (analysis of "intelligence" of each step).

In difference with traditional (psychometric) tests of intelligence the ITAB tests are their own laboratory, because of the richness of the data they generate. Until now, however, these data have not been analyzed. Ippel and Zaal (1984) provided a theoretical analysis of what constitutes "intelligence" for the control cycle. Now that the ITAB tests have been used in various contexts and sufficient data have been collected, a research line to investigate our conception of fluid intelligence will be developed.

---

[3] A computational theory is formulated as an effective procedure (Johnson-Laird, 1983), the feasibility of which is tested by computer modeling and its credibility by experimental studies.

Chapter 3
ITAB AND ASVAB: EXTERNAL RELATIONS AS EVIDENCE FOR ITAB MEASUREMENT CLAIMS

Most contemporary intelligence researchers consider general intelligence as an abstraction based on the common variance among cognitive ability tests. One of the most prominent theories is Carroll's (1993) three-stratum theory, which specifies three levels of abstraction based on the common variance among lower-order factors. At the bottom are rather specific factors such as Verbal Ability, Spatial Ability and Numerical Ability. Common variance among those 'primary factors' gives rise to second stratum factors, or group factors, the most important of which are fluid intelligence and crystallized intelligence. In turn, common variance among the secondary factors constitutes general intelligence (factor g).

According to Kline (1998) primary factors with median loadings of fluid intelligence (Gf) of at least 0.60 are Induction, Visualization, Quantitative Reasoning and Ideational Fluency. Similar loadings of crystallized intelligence (Gc) were found for Verbal Ability, Language Development, Reading Comprehension, Sequential Reasoning. Correlations between factors subsumed under different second stratum factors at a more moderate level, but are usually significantly different from zero.

This theoretical framework provides a background to evaluate the empirical findings of recent ITAB research. This chapter reviews the results of a study that investigated the relationship between the Armed Services Vocational Aptitude Battery (ASVAB) and the ITAB. The ASVAB does cover an array of abilities, but it is not a comprehensive basic ability measure. Russell, Peterson, Rosse, Hatten, McHenry and Houston (2001) point out that the ASVAB lacks tests for measuring Memory, Fluency, Perception and Spatial Ability.[4] However, the ASVAB structure is characterized by a general factor that has been related to general intelligence.

ASVAB and AFQT

The Armed Services Vocational Aptitude Battery (ASVAB) is a test battery consisting of nine tests, which in different configurations are used in recruitment, selection and classification of the

[4] Since 2002, the ASVAB also includes a spatial-visualization test: an Assembly of Objects test (AO).

Armed Forces. Table 3-1 lists the ASVAB tests and their measurement claims. The ASVAB does not deliver a comprehensive score other than the AFQT score, which is basically a measure of scholastic ability based on a subset of tests WK, PC, AR and MK. Other composites scores (e.g., the Navy's Minimum ASVAB Selection Composite scores) are also based on linear combinations of subsets of tests and are used to assign new recruits to military occupations (Sellman, 2004). The AFQT was designed as a measure of trainability for jobs in the Armed Forces. It has a five category system analogue to traditional IQ tests, ranging from very bright to very dull. In fact, the AFQT has loadings on general intelligence that are equal to or higher than most traditional intelligence tests. The AFQT is to a large extent a measure of past learning (i.e., crystallized intelligence).

Table 3-1: ASVAB tests and 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.

A Holzinger-Spearman Bi-Factor Model for the ASVAB

The factor structure of the ASVAB has been investigated over the years by a number of researchers. Two features characterize the ASVAB factor structure: (1) it has a dominant general factor accounting for approximately 60 percent of the variance (Kass, Mitchell, Grafton, & Wing, 1983; Welsh, Watson, & Ree, 1990), (2) in a number of studies group factors have been identified and could be replicated, viz., verbal ability (Verbal), quantitative ability (Quantitative), Speed, and Technical Knowledge (Kass et al., 1983; Welsh, Kucinkas, & Curran, 1990). The results of these

factor analytic studies supported a Holzinger-Spearman Bi-Factor Model for the ASVAB tests, that is, each ASVAB test loads on a general factor and a group factor with the notable exception of the split loadings of General Science (GS) on the Verbal and Technical Knowledge factors (Kass et al., 1983). The present version of the ASVAB no longer comprises tests measuring the factor Speed.

While the above studies were done with exploratory factor analysis techniques, Ippel (2006), Ippel and Watson (2008) used a confirmatory technique to determine the structure of the ASVAB. The last model standing was characterized by a general factor loaded by all ASVAB tests and three group factors, viz., Verbal, Quantitative and Technical Knowledge. As in the Kass et al. (1983) study General Science loaded on the Verbal and Technical Knowledge factors. Note, however, that the general factor in this model explained less variance than was claimed in the literature (i.e., 43.67% of the total variance).

In the sequel of this chapter this ASVAB bi-factor model was fitted to ASVAB test data and ITAB test data, both the ITAB tests combined into one score (ITAB) and for the separate tests (ITAB_B and ITAB_H).

METHOD

Goal of the Study

This study investigates the place of the ITAB within the ASVAB bi-factor structure. The ITAB measures fluent intelligence in a way that is not dependent on domain-specific knowledge. It is therefore expected that the ITAB will exclusively show an ASVAB general factor loading. However, since the ASVAB general factor is an abstraction of mostly domain-specific knowledge (i.e., more typical for crystallized intelligence or even "general knowledge" (e.g., Furnham, Camorro-Premuzic, 2006; Rolfhus and Ackerman, 1999) measured with he ASVAB tests, the loading of the general factor on the ITAB will be typically not in the range of 0.60s.

Notice that the impact of this study concerns the *meaning* of the ITAB, not its contribution to the predictive utility of the ASVAB. The selection and classification system of the Armed Forces, as far as it utilizes the ASVAB does not work with factor scores, but uses selection composites based on subsets of ASVAB tests. The predictive utility of the ITAB as a possible addition to the ASVAB is the central topic of Chapter 4.

Procedure

Participants were 2609 Navy recruits from schools with a focus on technical ratings in the Engineering and Combat Systems areas at the Great Lakes Navy training facility near Chicago. The recruits were

administered two tests of the Information Technology Aptitude Battery (ITAB), viz., the Hidden Target Test and the Battery Test. A total of 1420 recruits completed both tests. The Naval Personnel Development Command provided ASVAB test data for 1394 of the participants. [5]

INSTRUMENTS

I.T. Aptitude Battery (ITAB)

The I.T. Aptitude Battery consists of two tests, viz., the Hidden Target Test (HTT) and the Battery Test (BT). The tests are described in detail in Chapter 1.

Armed Forces Vocational Aptitude Battery (ASVAB)

The Armed Forces Vocational Aptitude Battery (ASVAB) consists of nine tests, which are listed in Table 3-1 in Chapter 3.

SAMPLING

Participants were 2609 U.S. Navy recruits from schools with a focus on technical ratings in the Engineering and Combat Systems areas of the Great Lakes Navy training facility near Chicago, IL. The recruits were administered two tests of the I.T. Aptitude Battery (ITAB), viz., the Hidden Target Test (HTT) and the Battery Test (BT). Due to initial technical problems related to the firewall at the Chicago's Great Lakes Recruit Training Command not all recruits completed both tests. We were able to collect complete test protocols of both tests of a total of 1420 recruits. In this analysis 5 independent and random samples of 300 recruits each were drawn from this grand sample. The samples used in this study were identical to the samples used in Chapter 1.

ANALYSIS

Model testing

The ASVAB bi-factor model (Ippel, 2006; Ippel and Watson, 2008) was fitted to the data of one sample. Subsequently, the ITAB scores were entered into the analysis and the effects were evaluated. Then the data were cross-validated against the remaining samples of Navy recruits (N = 300 each).

Indices for "goodness of fit"

In several analyses of this study a variety of fit indices will be reported to evaluate the goodness of fit in Linear Structural Equations (LSE) model tests. The indices can be divided into three cat-

[5] It turned out that, due to miscommunications with the proctors, in the beginning of the data collection process many recruits had taken only one of the two ITAB tests.

egories: (1) indices based on discrepancy between model and sample data, viz., chi-square (χ^2) or CMIN, and CMIN/DF; (2) indices based on a population discrepancy function, viz., the root mean square error of approximation (RMSEA); PCLOSE, the p-value for testing the null hypothesis that the population RMSEA is no greater than 0.05. (3) Indices based on comparisons with base-line models, viz., the normed-fit index (NFI), the comparative-fit index (CFI).

The RMSEA index for an exact fit equals 0.00. RMSEA values of approximately 0.08 or less usually are considered a reasonable error of approximation. The NFI and CFI values representing a close fit are well over 0.9. Only one of the indices, χ^2, is a statistical test in the strict sense, but of limited value since the power of this test to detect a discrepancy between model and data is largely controlled by the size of the sample. In this study it will be mainly used to signal the significance of a difference between models in the process of further constraining these models.

RESULTS AND DISCUSSION

Figure 3-1 is a graphical representation of the ASVAB bi-factor model used in the multi-group study of Ippel (2006), which was used as a starting point in this analysis. Two differences should be noted. First, the factor loading of Verbal on AO was eliminated and second, all latent variables were standardized (i.e., means were set equal to zero and variances equal to one), which kept the model

Table 3-2: A sequence of model tests of relaxing regression constraints group factor loadings on the ITAB on the model fit in a random sample of 300 Navy recruits

model	df	par	X2	X2 diff	sign.	CMIN/DF	RMSEA	PCLOSE	NFI	CFI
1. Ippel (2006) model	17	37	27.56			1.62	0.045	0.559	0.974	0.990
2. m#1+ ITAB	25	40	35.80			1.43	0.038	0.748	0.990	0.968
3. m#2 & (TK→ ITAB)	21	44	35.47	0.33	n.s.	1.47	0.040	0.705	0.969	0.989
4. m#2 & (V → ITAB)	24	41	35.20	0.60	n.s.	1.86	0.040	0.705	0.969	0.990
5. m#2 & (Q → ITAB)	24	41	35.70	0.10	n.s.	1.49	0.040	0.697	0.968	0.989
6. m#1 & ITAB_H/B	34	43	71.90			2.12	0.061	0.165	0.940	0.967

identified without fixing a regression parameter to one for each latent variable. The constraints on the error variances were necessary to prevent negative values.

The first row of Table 3-2 specifies the goodness of fit indices of the first sample. The fit is adequate by all standards. The next step was to include the ITAB into the analysis. This changed the total number of model parameters and degrees of freedom. In fact, it defined a

new model. The model fit indices of the expanded model are specified in row 2 of Table 3-2. Inclusion of the ITAB did not seem to have negatively influenced the overall adequacy of the fit. The next three rows of Table 3-2 describe the effects of relaxing a particular model constraint related to the factor pureness of the ITAB. Model 2 gives the model fit of the test battery with the ITAB exclusively influenced by the general ASVAB factor, not by any of the group factors.

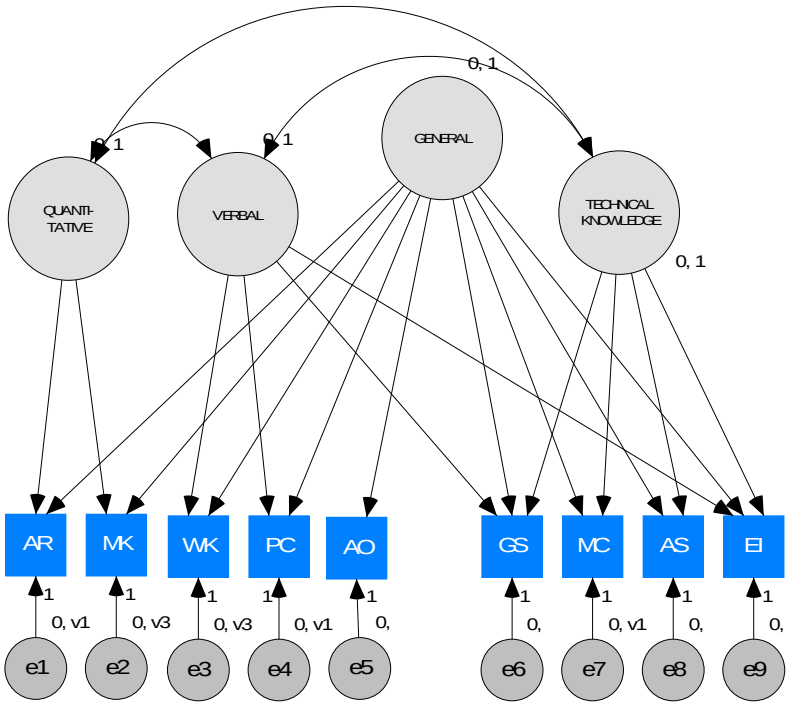

Figure 3-1: LSE model representation of the ASVAB bi-factor model (see: Ippel, 2006)

Relaxation of the relationship between Technical Knowledge and ITAB (model #3) does not result in a significant improvement of model fit, the corresponding factor loading on the ITAB was 0.042 and tested not significantly different from zero. Relaxation of the constraint on the regression of the Verbal factor on the ITAB (model #4) essentially did not change the fit, and again the factor loading tested not significantly different from zero. (B = -0.055, n.s.). Finally, the relaxation of the regression constraint of Quantitative on the ITAB (model #5) was not significant and the regression coefficient tested not different from zero (B = -0.153, n.s.). The final row of Table 3-2 gives the model fit indices of an expanded ASVAB model with the ITAB tests as separate variables. This model also showed a very acceptable fit to the data as well.

Table 3-3: Pattern of factor loadings in same sample (nr. 3) of Navy recruits (N = 300) in three different solutions: A = ASVAB tests only; AI = ASVAB tests and ITAB score; AHB = ASVAB tests, HTT and BT

	G			Verbal			Quantitative			Technical Knowledge			Communality		
	A	AI	AHB	A	AI	AHB	A	AI	AHB	A	AI	AHB	A	AI	AHB
g															
verbal	0	0	0				-0.129	-0.125	-0.076	0.277	-0.320	-0.347			
quant.	0	0	0	-0.129	-0.125	-0.076				-0.393	0.423	0.369			
TK	0	0	0	0.277	-0.32	-0.347	-0.393	-0.320	0.369						
WK	0.453	0.435	0.42	0.815	0.806	0.809							0.869	0.838	0.831
PC	0.480	0.469	0.458	0.369	0.402	0.420							0.367	0.381	0.386
MK	0.577	0.614	0.611				0.69	0.623	0.621				0.81	0.766	0.759
AR	0.682	0.711	0.717				0.182	0.139	0.149				0.498	0.525	0.536
MC	0.721	0.710	0.701							0.386	-0.422	-0.444	0.668	0.682	0.688
EI	0.569	0.539	0.519	0.278	0.296	0.306				0.378	-0.397	-0.408	0.602	0.611	0.617
AS	0.318	0.301	0.287							0.847	-0.823	-0.819	0.819	0.768	0.768
GS	0.639	0.623	0.614	0.429	0.450	0.459				0.185	-0.197	-0.203	0.671	0.686	0.686
AO	0.532	0.553	0.568										0.283	0.305	0.323
ITAB		0.512												0.262	
HTT			0.487												0.238
BT			0.529												0.280

Table 3-3 displays the factor loading pattern of the three solutions specified in Table 3-2 and tested on the same random sample of Navy recruits (N = 300). The general factor in the solutions presented in Table 3-3 binds major portions of the common variance of all tests with the exception of AS, which has larger TK factor loadings, and MK, which has factor loadings of about equal size of G and V. The pattern of factor loadings characterizes this general factor as a mixture of crystallized intelligence and general knowledge. AO is somewhat isolated in this context. AO is a conventional test of intelligence, that is, in difference with the ASVAB tests no domain-specific knowledge is assumed. It is a spatial-visualization test. Visualization (Vz) is sometimes considered a component of fluid intelligence (Gf), sometimes it is considered another second-order factor in its own right next to fluid intelligence (Carroll, 1993). AO and ITAB are conceptually not closely related. Nevertheless, Table 3-3 illustrates the effect of adding not domain-dependent tests (i.e., the ITAB tests in combination or separate) to the analysis – the factor load-

ing of AO increased stepwise. Not only did the factor loading of AO increase, the average of the g-factor loadings of the other tests decreased, viz., from 0.555 (under A), to 0.550 (under AI) to 0.541 (under AHB). Since the average communality of the latter group of tests remains the same (average H^2 = 0.66), a slight redistribution of common variance from the general to the group factors must have taken place. This can be verified by visual inspection of the factor pattern in Table 3-3.

The results summarized in Table 3-2 demonstrate that the ITAB tests exclusively loaded on the general factor. This common variance must come from what general intelligence shines through in crystallized intelligence.

Conclusions

1. Within the ASVAB bi-factor structure, the ITAB exclusively loads on the g-factor. The loadings are moderately high, which could be expected if the claim that the ITAB tests measure fluent intelligence is valid. ASVAB tests, with the exception of AO, measure domain-specific knowledge and the ASVAB general factor is the embodiment of the common variance between those tests.
2. As a single not (knowledge) domain-specific test of cognitive functioning AO is somewhat isolated in the ASVAB context. It is a spatial-visualization test. Visualization (Vz) is sometimes considered a component of fluid intelligence (Gf), sometimes it is considered another second-order factor in its own right next to fluid intelligence (Carroll, 1993). Addition of a one extra test outside the domain of crystallized intelligence or general knowledge will not change the factor structure much. Suggestions in that direction (e.g., Waters, Russell, & Sellman, 2007) seem to overlook that the ASVAB is not a general cognitive ability test battery, but a battery measuring a mixture of general knowledge and crystallized intelligence.

Chapter 4

PREDICTING CRITERION PERFORMANCE IN THE NAVY APPRENTICE TECHNICAL TRAINING PROGRAM

METHOD

Goal of the Study

This study investigated the prediction of training performance of Navy recruits participating in the Navy's Apprentice Technical Training (ATT) program at the Great Lakes Recruit Training Command in Chicago, IL. The goal was to determine the relationships among ASVAB tests, several selection composites based on the ASVAB, ITAB and a set of newly developed criterion measures (Watson and Ippel, 2008). In particular, we tested whether the ITAB had incremental validity over certain ASVAB selection composites in predicting success in the ATT program.

Procedure

The ASVAB test scores were obtained with paper-and-pencil enlistment tests, several months before this study was conducted. The ITAB tests were administered shortly before, or during the first sessions of, the 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, 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.

INSTRUMENTS

Criterion Variables

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 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) was determined by Navy experts 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.[6]

The new criterion variables were designed so as to incorporate the absolute standard of minimal competence as well as to certify sufficient individual differences variance on the variables. The new criterion variables measure 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 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 post-test scores as an increasing function of the observed scores.

---

[6] 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).

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,

$$\text{Logit } [\pi(x)] = \log (\pi(x) / (1 - \pi(x))) = \alpha + \beta x \qquad [4\text{-}1]$$

Subsequently, the logit score obtained with Equation [4-1] can be transformed into the estimated probability that X = 1 at a fixed value x of X_O by

$$\text{Est. } \pi(x) = \exp (\alpha + \beta x) / (1 + \exp (\alpha + \beta x)) \qquad [4\text{-}2]$$

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}).[7]

Some of the training modules were designed for all ATT ratings; others were designed for certain specialties.[8] In this study we exclusively focus on general modules, that is, modules that were part of the basic training of all ATT ratings.

Table 4-1.a. and Table 4-1.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 variables with one exception: the modeling of the P-test for module 1 failed. The model fit was low and the resulting

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

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

variable correlated negatively with the other variables. The columns four in each table shows the fits after the outliers were removed from the data.

Table 4-1.a: Model Fits and Reliabilities of the new MCL measures of Declarative knowledge (N = 2773) (from: Watson & Ippel, 2008)

module	model fit			reliability estimates			
	R^2	out-liers	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

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 variables. The final four columns in Tables 1.a. and 1.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 variables 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. 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.

Table 4-1.b: Model fits and reliabilities for the new MCL measures of procedural knowledge (N = 2773) (from: Watson & Ippel, 2008)

module	model fit			reliability estimates			
	R^2	out-liers	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	0.98	0.64	0.46	0.53	0.54

*) no Post Test available

Predictors (1): ASVAB tests and ASVAB selection composites

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. The test battery is described in more detail in Chapter 3. 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).

While the AFQT score, derived from the ASVAB, serves as a screening test for all Services, the Services combine the various ASVAB tests into different "aptitude area" composites, which are used to assign new recruits to military occupations (Sellman, 2004). 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[9] 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

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

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 (ITAB)

The I.T. Aptitude Battery is described in detail in Chapter 1.

SAMPLING

Correlation matrices of ASVAB tests, ITAB, and ATT criterion scores were based on the availability of complete data for those variables. The final sample size was determined by whether the data sets included scores on ASVAB tests, ITAB tests, and the ATT criterion scores. The initial sample for ASVAB scores was 2547; the initial sample for ATT criterion scores was 2773; the initial sample for the ITAB test scores was 1391; the sample of complete ASVAB and ATT criterion scores was 896; the final sample including the ITAB scores was 436.

ANALYSES

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.

$$\Delta R^2 = R^2_{ASVAB\ +ITAB} - R^2_{ASVAB} \qquad [4\text{-}3]$$

$$F_{1,\,N-3} = (N-1)\left(\frac{\Delta R^2}{1 - R^2_{ASVAB\ +\ ITAB}}\right) \qquad [4\text{-}4]$$

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 were 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:

$$R_{xx} = 1 - (s_x^2 / S_x^2)\,(1 - r_{xx}), \qquad [4\text{-}5]$$

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:

$$\rho = 1 - [(N - 1) / (N - p - 1)] (1 - r^2), \qquad [4\text{-}6]$$

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.[10]

Simultaneous versus a sequential hurdle model

The first prediction model in which the incremental validity of the ITAB was tested was a simultaneous predictor model. Both predictors were considered at the same time with the only logical distinction that the ASVAB tests used in the ASVAB Selection Composites were a given and the incremental validity of the ITAB over these selection composites was object of study. An alternative approach is to define cut-off scores with the established selection composites and investigates whether the ITAB did possess incremental validity in such a situation. Guion (1998) refers to this approach as a sequential hurdle model to selection.

As was specified in the Analyses section, an important difference between the two ASVAB Selection Composites is formed by the job clusters for which they define cut-off scores. The first ASVAB Selection Composite, ASC01, is not very selective. It defines the base rate in the population of reference for the intended rating cluster at about 95 percent.[11] ASC02 is much more selective. It defines the base rate of success in the cluster of job ratings with an ASC02 of minimally 222 at 57.5 percent.

In the sequel we will present the results of an incremental validity analysis of the ITAB over the ASC02 in a sample in which the variance had been restricted by applying the ASC02 cut-off score.

RESULTS

Reliability Estimates

Tables 4-1.a and 4-1.b show the original reliabilities as estimated in the initial sample (N = 2773).

---

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

[11] A population base rate is the proportion of individuals in the population of reference that can be expected to be successful in a particular job, or cluster of jobs.

Table 4-2 shows the corrected reliabilities for the same sample and the corrected values for the ratings cluster ASC02 with the cut-off score of 211.

Table 4-2: 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 Measuremen	**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.

The values in Table 4-2 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 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 4-3 and 4-4. Table 4-3 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 in explained variance in all models tested.

Table 4-3: 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 [12]

Selection	modules (C)	N	R predictors		Multiple R			Incremental Validity		
Composite			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

The correlation between the ITAB and ASC01 was 0.364 and the correlation with ASC02 was 0.359. Table 4-4 displays the percentages of improvement in predicting criterion variance after correction for attenuation due to unreliability of the criterion.

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*

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

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 would not have made much of a difference since both the ASVAB tests and the ITAB tests are highly reliable tests.

Table 4-4: 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 [13]

Selection	N	criterion variables		R predictors		(CMR)	Incremental Validity	
Composite		name	Reliability	ASC..	ITAB	R^2	ΔR^2	% 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%

Tables 4-5 and 4-6 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 4-6. 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 oc-

[13] 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.

curred in the prediction of procedural skill criterion scores of the Digital Logic Function (DLF) Module (24-P), albeit less extreme.

Table 4-5: 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	R^2	ΔR^2	% 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%

DISCUSSION

In general, 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 large. Somewhat weaker, but still very large, was the effect with on the prediction of the DLF skill test.

The incremental utility effect of the ITAB appeared 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.

Table 4-6: 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%

Conclusions

1. All tests of incremental validity of the ITAB over two ASVAB selection composites for the prediction of success in technical training at Navy schools were significant. The estimates were made under assumptions more conservative than in comparable studies.
2. The estimates of the percentages of improvement in predictive validity were made with corrections for criterion unreliability. This is a recommended procedure (see footnote 9) and standard practice. The estimates were substantially higher than in comparable studies.
3. The improvement effects of predictive utility were very general. It was shown to occur in all training modules under study. At the same time, it the effect were the largest for modules that require a relatively high level abstract thinking (i.e., Digital Logic Functions).
4. The ASVAB is a very high standard to be held against. Most experts would give any well-designed test low odds on improving the predictive utility of the ASVAB. We believe that ITAB could, because the ITAB was designed to measure the aptitude to learn procedural skills and makes optimal use of information technology to realize that goal. Conventional tests in the domain of cognition, such as the ASVAB, were not designed to measure procedural skill learning. Even when technical knowledge is the measurement objective, it is being treated as declarative knowledge (e.g., ASVAB subtests: Auto/Shop, Mechanical Comprehension, Electronics Information).

Chapter 5

CONCLUSIONS AND RECOMMENDATIONS

Conclusions

Conclusion 1: *The basic variables in the diagnostic structures of the ITAB tests, which are aggregates of data generated in the interaction of the examinee with the task environment generated by the test systems, proofed to be of sufficient level homogeneity and stability to be considered attributes of human performance.*

Evidence for homogeneity was found by applying Cronbach's alpha to all multi-item scales in the diagnostic structures of the ITAB tests. Stability was concluded based on comparison of different split-half coefficients on a single scale (i.e., S-H(O/E) versus S_H(F/S). Different evidence regarding the stability of the estimates came from comparing estimates parameter values over five independent random samples with N = 300 each.

Conclusion 2: *The ITAB scores, viz., the combination scores (ITAB, INT, WS) as well as the scores of separate tests (ITAB_H, INT_H, WS_H; ITAB_B, INT_B, WS_B) proofed highly reliable.*

The estimate reliabilities ranged between 0.865 and 0.960.

Conclusion 3: *Construct validity research on the ITAB is still in its initial stage. However, since the tests produces a continuous data stream while being administered, chances are very good for research to generate insights in the functioning of the control cycle as an engine of the operation of fluid intelligence.*

Conclusion 4: *Within the ASVAB bi-factor structure, the ITAB exclusively loads on the ASVAB g factor. The loadings were moderately high, which should be expected if the claim that the ITAB tests measure fluid intelligence is valid. ASVAB tests with the exception of AO, measure domain-specific knowledge and the ASVAB general factor is the embodiment of the common variance between those tests.*

Conclusion 5: *The ITAB tests proofed to possess incremental validity over several ASVAB selection composites in the prediction of training success in the Apprentice Technical Training program at the Great Lakes Recruit Training Command in Chicago, IL.*

5-1 All tests of incremental validity of the ITAB over two ASVAB selection composites for the prediction of success in technical training at Navy schools were significant and substantial.

5-2 The estimates were made under more conservative assumptions then in comparable effect studies.

5-3 The improvement effects of predictive utility were very general. It was shown to occur under all training modules under study. At the same time the effect were the largest for modules that require abstract thinking (e.g., digital logic circuits).

Recommendation

Recommendation 1: *The criterion scores to be predicted in this study were the result of a recent pilot study (Ippel & Seals, 2008, see also: Watson & Ippel, 2008). A correction of unreliability of the criterion was applied to these scores. These criterion scores deserve a deeper and more thorough investigation and development.*

Correcting for criterion unreliability (correction for attenuation) is standard practice in the evaluation of predictive utility of psychometric tests. This correction takes care of attenuation of the scores according to particular assumption regarding the distribution of error variance and its relation to the true score variance. It does not correct for inaccuracies following from a weak conceptualization of what is being measured. In particular, the way in which procedural skills are being measured in the ATT post tests deserves attention.

References

Anderson, J.R. (1983). *The Architecture of Cognition.* Cambridge, MA: Harvard University Press.

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review, 94*, 192-210.

Anderson, J.R., & Singley, M.K. (1989). *The Transfer of Cognitive Skill*. Cambridge, MA: Harvard University Press.

Anderson, J.R., & Singley, M.K. (1989). *The Transfer of Cognitive Skill*. Cambridge, MA: Harvard University Press.

Ash, Robert, B. (1965). *Information Theory*. New York: Dover Publications.

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

Carlstedt, B. (2001). Differentiation of cognitive abilities as a function of level of general intelligence: A latent variable approach. *Multivariate Behavioral Research, 36,* 589 – 609.

Carroll, J.B. (1993). *Human Cognitive Abilities. A survey of factor-analytic studies.* Cambridge (UK): Cambridge University Press.

Chamorro-Premuzic, T., & Furnham, A. (2006). Intellectual competence and the "intelligent personality": A third way in differential psychology. *Review of General Psychology*. In press.

Dunbar, S. B., & Linn, R. L. (1991). Range restriction adjustments in the prediction of military job performance. In A. K. Wigdor & B. F. Green, Jr. (Eds.), *Performance Assessment for the Workplace, II* (pp. 127-157). Washington, DC: National Academy Press.

Edwards, E. (1964). *Information Transmission*. London, UK: Chapman & Hall.

Evans, M.G. (1999). On the asymmetry of g. *Psychological Reports, 85*, 1059-1069.

Garner, W.R. (1962). *Uncertainty and Structure as Psychological Concepts*. New York: John Wiley and Sons, Ltd.

Guion, R.M. (1998). *Assessment, Measurement, and Prediction for Personnel Decisions.* Mahwah, N.J.: L.E.A.

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

Holland, J.H., Holyoak, K.J., Nisbett, R.E., & Thagard, P.R. (1988). Induction: Processes of inference, learning and discovery. *Journal of Economic Behavior and Organization, 9,* 318-321.

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

Ippel, M.J. (2006). Investigation into the measurement invariance of ASVAB scores using a multi-group confirmatory factor analysis. *CogniMetrics Research Report, Nr. 06-01*. (44 pp).

Ippel, M.J. (2007) Measurement Invariance of ASVAB and ITAB tested in three clusters of Navy Ratings. *CogniMetrics Research Report, Nr. 07-01.* (47 pp).

Ippel, M.J., & Watson, S.E. (October, 2008). ASVAB: E Pluribus Unum? Paper presented at the 50th Annual Conference of the International Military Testing Association., Amsterdam, September 29 – October 3.

Ippel, M.J., & Zaal, J.N. (Oct. 2004*). A New Work-Environment: New Aptitudes Require New Measures.* Paper presented at the 46th Annual Conference of the International Military Testing Association (IMTA), 26-28 Oct. 2004, Brussels, Belgium (http://www.internationalmta.org/Documents/2004/2004033P.pdf).

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).

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.

Kass, R.A., Mitchell, K.J., Grafton, F.C., & Wing, H. (1983). Factorial Validity of the Armed Service Vocational Aptitude Battery (ASVAB), Forms 8, 9, and 10; 1981 Army applicant Sample. *Educational and Psychological Measurement, 43*, 1077-1087.

Kline RB. 1998. Principles and Practice of Structural Equation Modeling. New York:Guilford.

Krippendorff, K. Information Theory. *Structural Models for Qualitative Data*. London, UK: Sage Publications.

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

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

Lovett, H.T. (1977). Criterion-Referenced Reliability Estimated by ANOVA, Educational and Psychological Measurement, 37, 1, 21-29.

McDonald, R.P. (1999). *Test Theory: A Unified Treatment*. Mahwah, NJ: L.E.A.

Newell, H. (1981). Reasoning, Problem Solving and Decision Processes. Problem Space as a Fundamental Category. In: R. *Nickerson (Eds.) Attention and Performance*. Vol. 8. (pp. 693 – 718). Hillsdale, N.J.: L.E.A.

Newell, A. (1990). *Universal Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Simon, H.A. (1972). *Human Problem Solving.* Englewood Cliffs, NJ: Prentice-Hall.

Nunnally J C. *Psychometric theory.* New York: McGraw Hill, 1967, 640 p.

Rolfhus, E.L., & Ackerman, P.L., (1999). Assessing individual differences in knowledge: knowledge, intelligence, and related traits, *Journal of Educational Psychology,* 91, 511-526.

Russell, T. L., Peterson, N. G., Rosse, R.L., Hatten, J.T., McHenry, and Houston, J.S. (2001). The measurement of cognitive, perceptual, and psychomotor abilities. In: J.P. Campbell & D.J. Knapp (Eds.). Exploring the limits of personnel selection and classification. Mahwah (N.J.): Lawrence Erlbaum Associates.

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

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York: MacMillan [Reprinted: New York: AMS Publishers].

Thorndike, R.L. (1949). *Personnel Selection*. New York: Wiley.

VanLehn, K. (1990). *Mind Bugs. The Origins of Procedural Misconceptions.* Cambridge, MA: MIT Press.

VanLehn, K. (Ed) (1991) Architectures for intelligence: The 22nd Carnegie Mellon symposium on cognition. Hillsdale, NJ: Lawrence Erlbaum Assoc.

Waters S.D., Russell T.L., & Sellman S.W. (2007). Review of non-verbal reasoning tests (FR-07-36). Alexandria, VA: Human Resources Research Organization.

Welsh, J.R., Watson, T.W., & Ree, M.J. (1990). *Armed Services Vocational Aptitude Battery (AS-VAB): Predicting military criteria from general and specific abilities* (AFHRL-TR-90-63). Brooks, AFB, TX: U.S. Air Force Human Resources Laboratory.

Welsh, J.R., Kucinkas, S.K., & Curran, L.T. (1990). *Armed Services Vocational Aptitude Battery (ASVAB): Integrative review of validity studies* (AFHRL-TR-90-22). Brooks, AFB, TX: U.S. Air Force Human Resources Laboratory.

Wherry, R.J. (1937). A new formula for predicting the shrinkage of the coefficient of multiple correlations. Annals of Mathematical Statistics, 2, 446-457.

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

Appendix 1

MODEL FITS OF LSE MODELS OF THE ITAB DIAGNOSTIC STRUCTURES IN 5 RANDOM SAMPLES OF 300 NAVY RECRUITS

Indices for "goodness of fit"

In several analyses of this study a variety of fit indices will be reported to evaluate the goodness of fit in Linear Structural Equations (LSE) model tests. The indices can be divided into three categories: (1) indices based on discrepancy between model and sample data, viz., chi-square (χ^2) or CMIN, and CMIN/DF; (2) an index based on a population discrepancy function, viz., the root mean square error of approximation (RMSEA); PCLOSE, the p-value for testing the null hypothesis that the population RMSEA is no greater than 0.05. (3) Indices based on comparisons with base-line models, viz., the normed-fit index (NFI), the comparative-fit index (CFI).

The RMSEA index for an exact fit equals 0.00. RMSEA values of approximately 0.08 or less usually are considered a reasonable error of approximation. The NFI and CFI values representing a close fit are well over 0.9. Only one of the indices, χ^2, is a statistical test in the strict sense, but of limited value since the power of this test to detect a discrepancy between model and data is largely controlled by the size of the sample. In this study it will be mainly used to signal the significance of a difference between models in the process of further constraining these models.

Battery Test

Sample	df	par	X2	sign.	CMIN/DF	RMSEA	PCLOSE	NFI	CFI
HBSample1	35	30	162.547	0.000	4.644	0.011	0.000	0.931	0.944
HBSample2	35	30	264.296	0.000	7.551	0.148	0.000	0.895	0.907
HBSample3	35	30	215.515	0.000	6.158	0.131	0.000	0.901	0.915
HBSample4	35	30	193.953	0.000	5.542	0.123	0.000	0.916	0.930
HBSample5	35	30	202.217	0.000	5.778	0.126	0.000	0.916	0.891

Hidden Target Test

Sample	df	par	X2	sign.	CMIN/DF	RMSEA	PCLOSE	NFI	CFI
HBSample1	40	25	353.238	0.000	8.831	0.162	0.000	0.873	0.886
HBSample2	40	25	578.951	0.000	14.474	0.212	0.000	0.804	0.815
HBSample3	40	25	654.458	0.000	16.362	0.227	0.000	0.816	0.771
HBSample4	40	25	761.793	0.000	19.045	0.246	0.000	0.747	0.757
HBSample5	40	25	465.876	0.000	11.647	0.189	0.000	0.819	0.832