CAB D0003683.A1/Final
April 2001

Marine Corps Selection and Classification

William H. Sims • Catherine Hiatt

CNA
4825 Mark Center Drive • Alexandria, Virginia 22311-1850

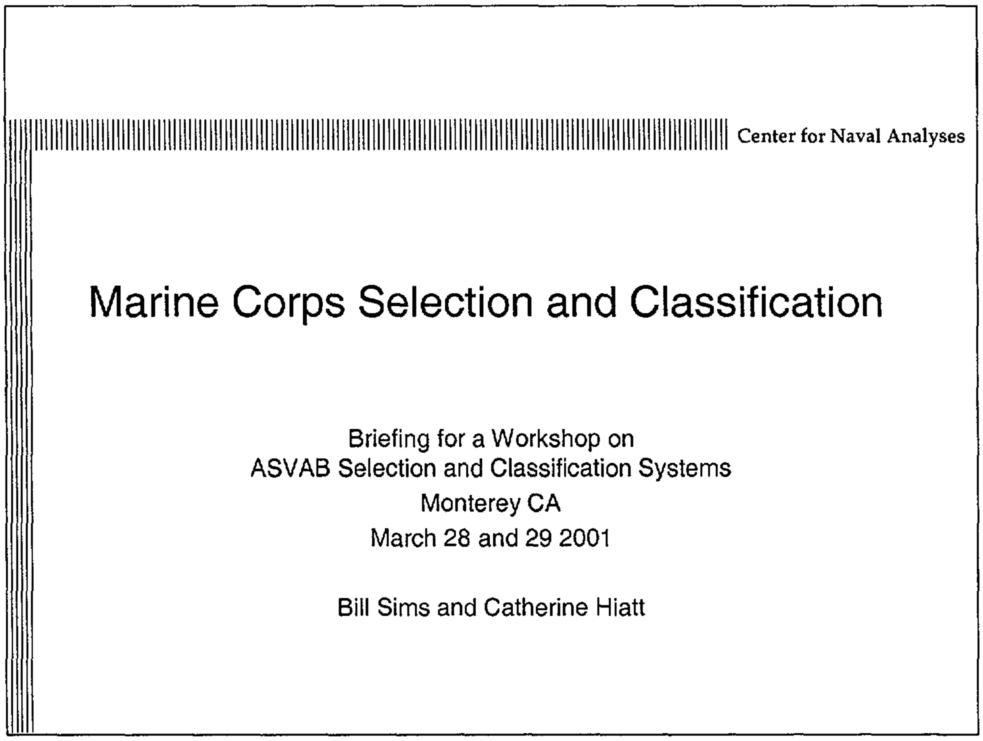
Approved for distribution: April 2001

Donald J. Cymrot

Donald J. Cymrot, Director
Workforce, Education and Training Team
Resource Analysis Division

This document represents the best opinion of CNA at the time of issue.
It does not necessarily represent the opinion of the Department of the Navy.

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED
For copies of this document, call the CNA Document Control and Distribution Section (703) 824-2943



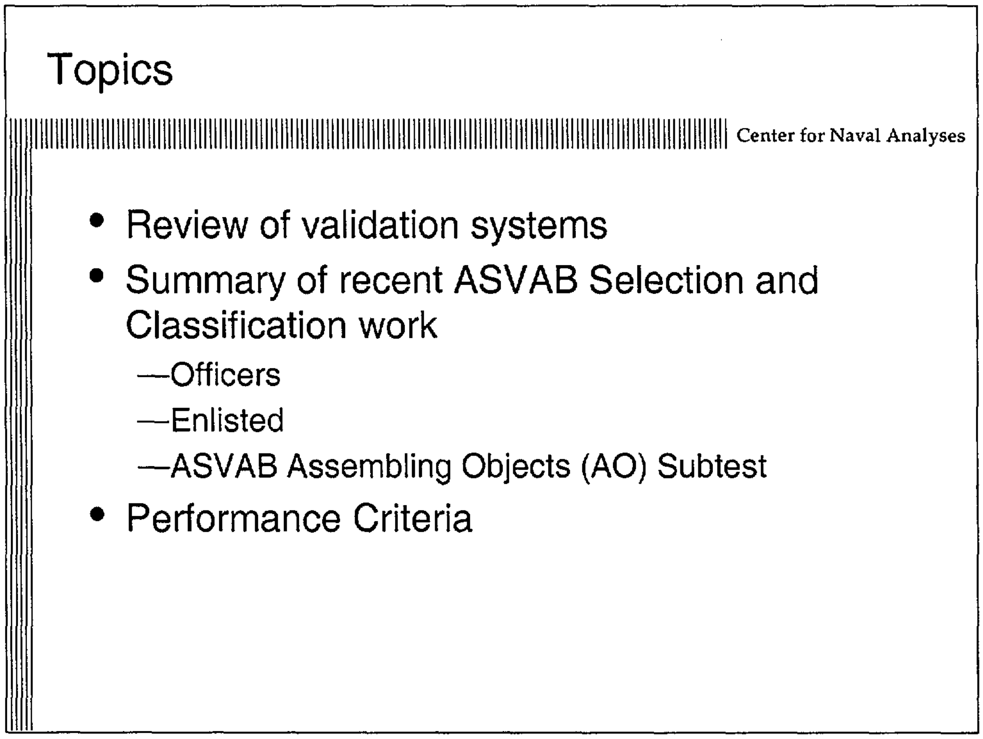
Center for Naval Analyses

Marine Corps Selection and Classification

Briefing for a Workshop on
ASVAB Selection and Classification Systems
Monterey CA
March 28 and 29 2001

Bill Sims and Catherine Hiatt

This report documents a series of briefings on Marine Corps personnel selection and classification issues presented at a workshop held March 28 and 29 2001.



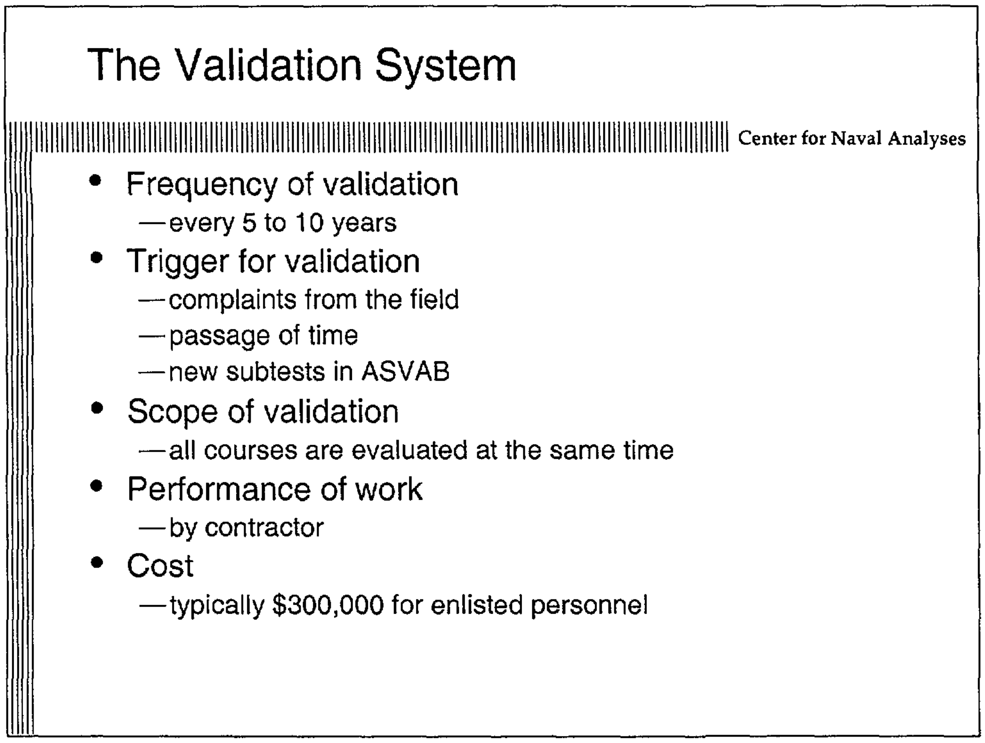
Topics

Center for Naval Analyses

- Review of validation systems
- Summary of recent ASVAB Selection and Classification work
  - —Officers
  - —Enlisted
  - —ASVAB Assembling Objects (AO) Subtest
- Performance Criteria

Separate briefings (combined in this report) covered the following issues:

- Review of validation systems
- Validation of ASVAB for selection and classification of:
  - Officers
  - Enlisted
- Validation of the experimental Assembling Objects subtest
- Performance criteria

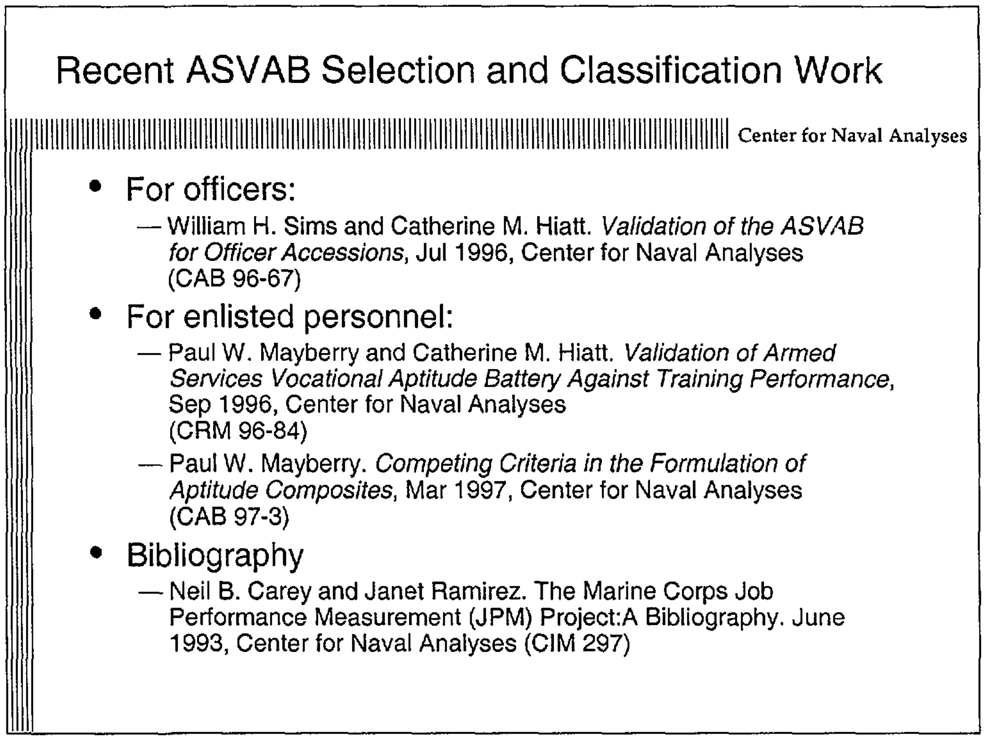


The Validation System

Center for Naval Analyses

- Frequency of validation
  - —every 5 to 10 years
- Trigger for validation
  - —complaints from the field
  - —passage of time
  - —new subtests in ASVAB
- Scope of validation
  - —all courses are evaluated at the same time
- Performance of work
  - —by contractor
- Cost
  - —typically $300,000 for enlisted personnel

Validation of ASVAB is usually done for the Marine Corps every 5 to 10 years. The trigger is usually the addition of new subtests to the ASVAB. All courses are evaluated at the same time. For the last 30 years the Center for Naval Analyses has done this work under contract. A typical analysis covering all enlisted jobs would be expected to cost around $300,000.



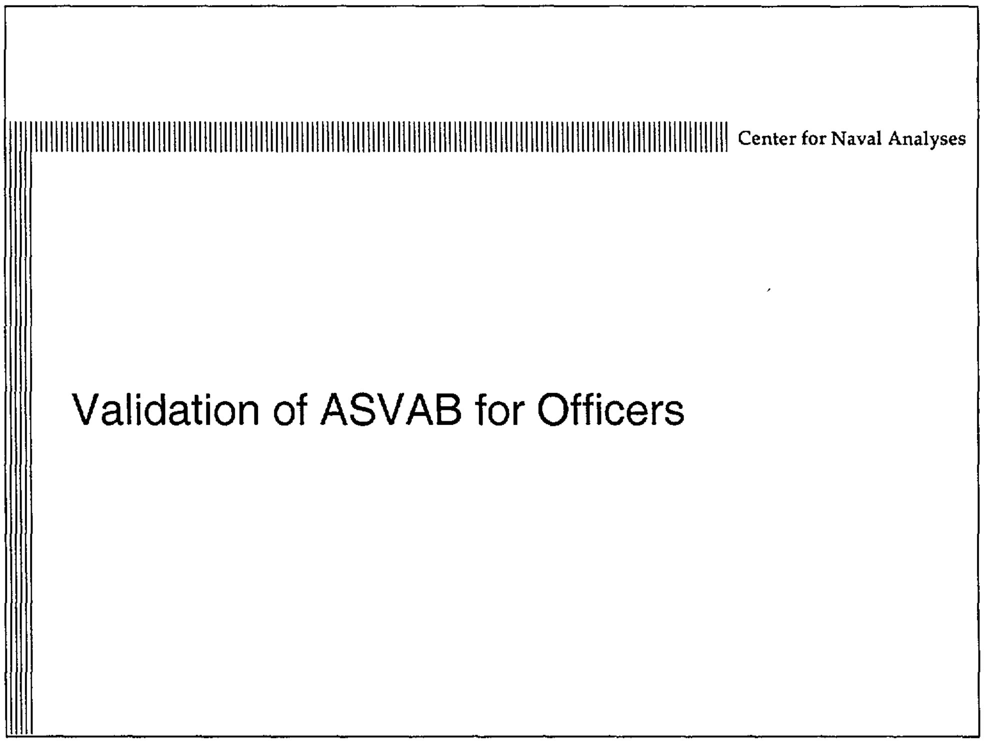
Recent ASVAB Selection and Classification Work

Center for Naval Analyses

- For officers:
  - — William H. Sims and Catherine M. Hiatt. *Validation of the ASVAB for Officer Accessions*, Jul 1996, Center for Naval Analyses (CAB 96-67)
- For enlisted personnel:
  - — Paul W. Mayberry and Catherine M. Hiatt. *Validation of Armed Services Vocational Aptitude Battery Against Training Performance*, Sep 1996, Center for Naval Analyses (CRM 96-84)
  - — Paul W. Mayberry. *Competing Criteria in the Formulation of Aptitude Composites*, Mar 1997, Center for Naval Analyses (CAB 97-3)
- Bibliography
  - — Neil B. Carey and Janet Ramirez. The Marine Corps Job Performance Measurement (JPM) Project:A Bibliography. June 1993, Center for Naval Analyses (CIM 297)

These briefings draw heavily on the reports listed in this slide.

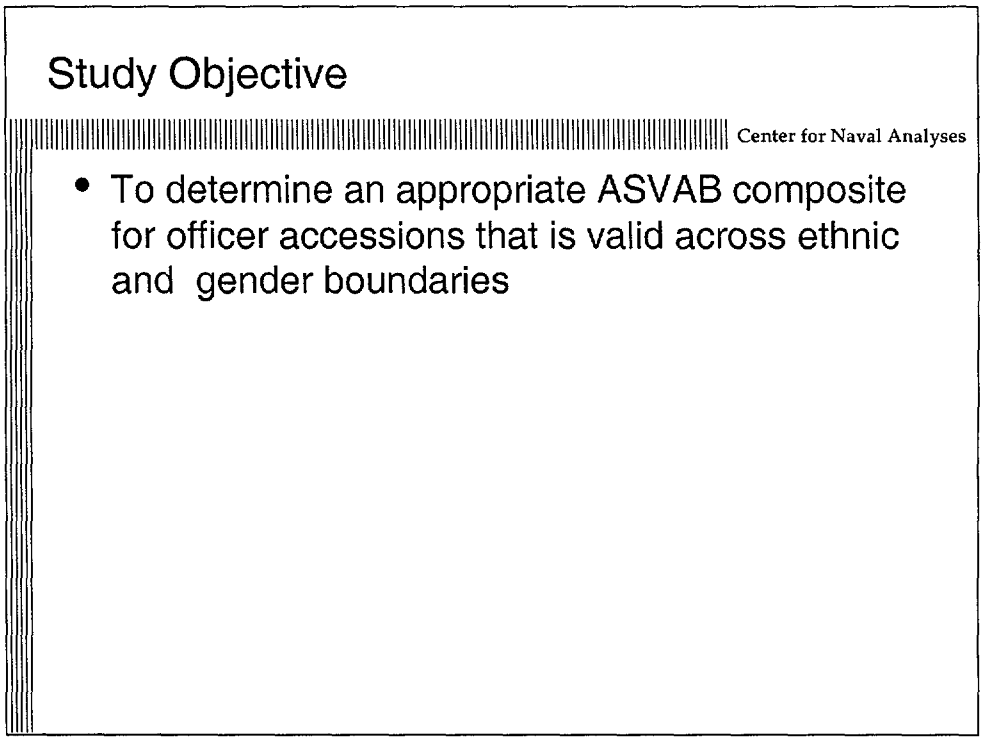
For a more extensive list and short description of each, see the bibliography produced by Carey and Ramirez. This bibliography lists over 50 CNA reports bearing on selection, classification, and performance criteria selection and development.



Center for Naval Analyses

Validation of ASVAB for Officers

This section of the briefing describes some work by myself and Catherine Hiatt on examining the validity of ASVAB for use in the selection of Marine Corps officers.

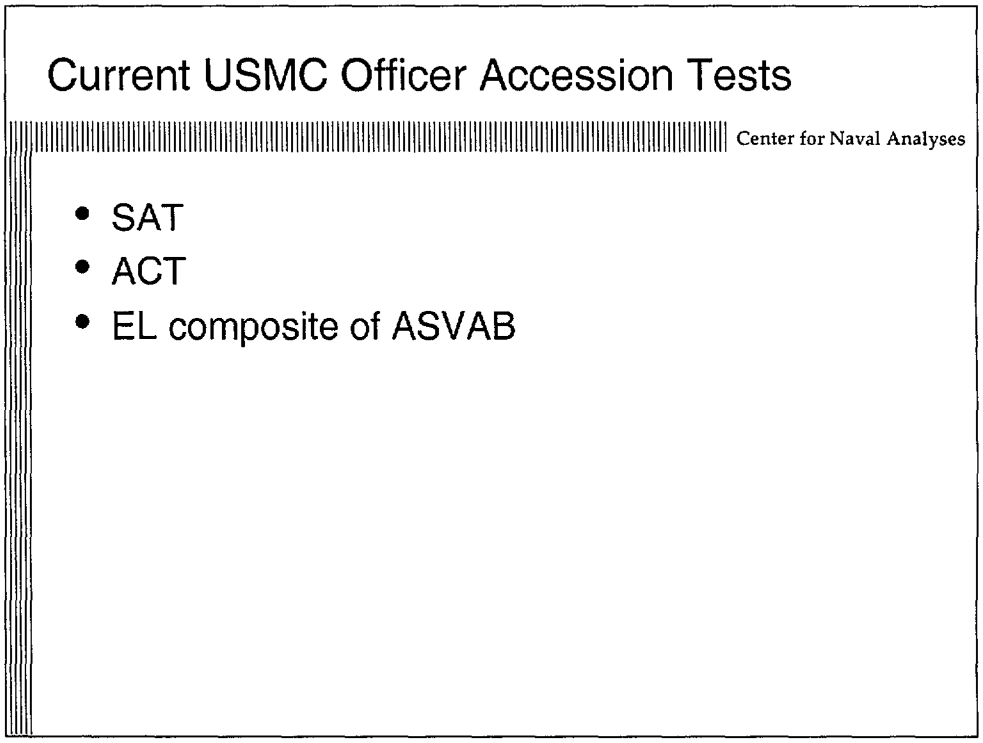


Study Objective

Center for Naval Analyses

- To determine an appropriate ASVAB composite for officer accessions that is valid across ethnic and gender boundaries

The objectives of this study were to determine an appropriate ASVAB composite for officer accessions that is valid across ethnic and gender boundaries.



Current USMC Officer Accession Tests

Center for Naval Analyses

- SAT
- ACT
- EL composite of ASVAB

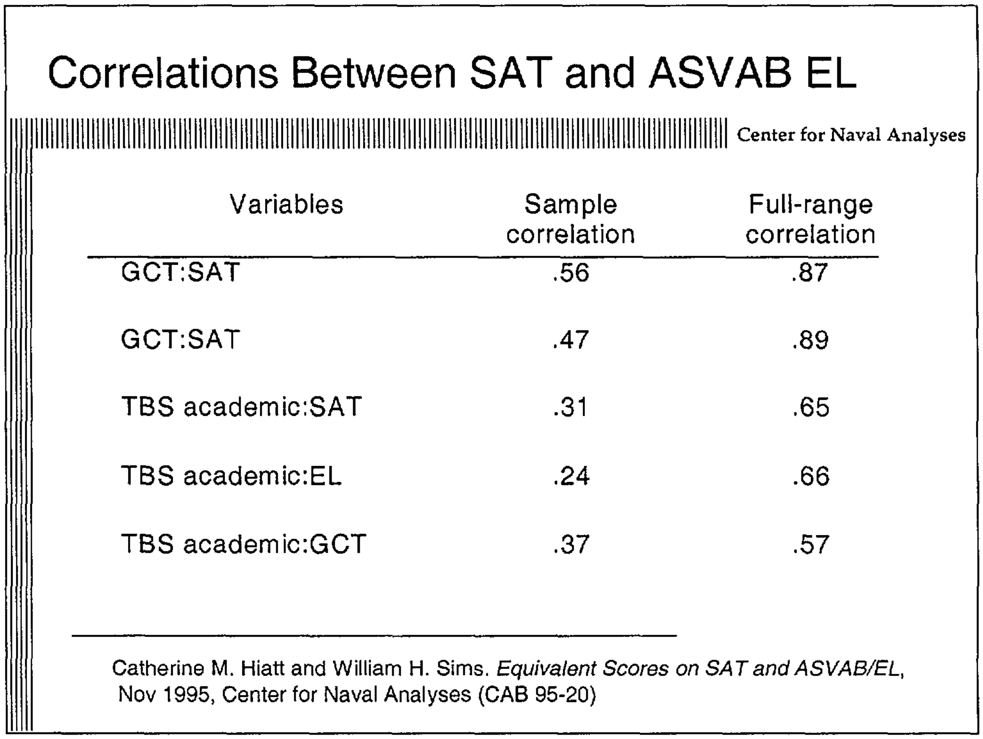
Currently Marine Corps uses the following tests in the officer selection process:

•SAT

•ACT

•EL composite of ASVAB

The EL composite from ASVAB was initially chosen by the Marine Corps as a matter of convenience and because it has very strong mathematics and verbal components.



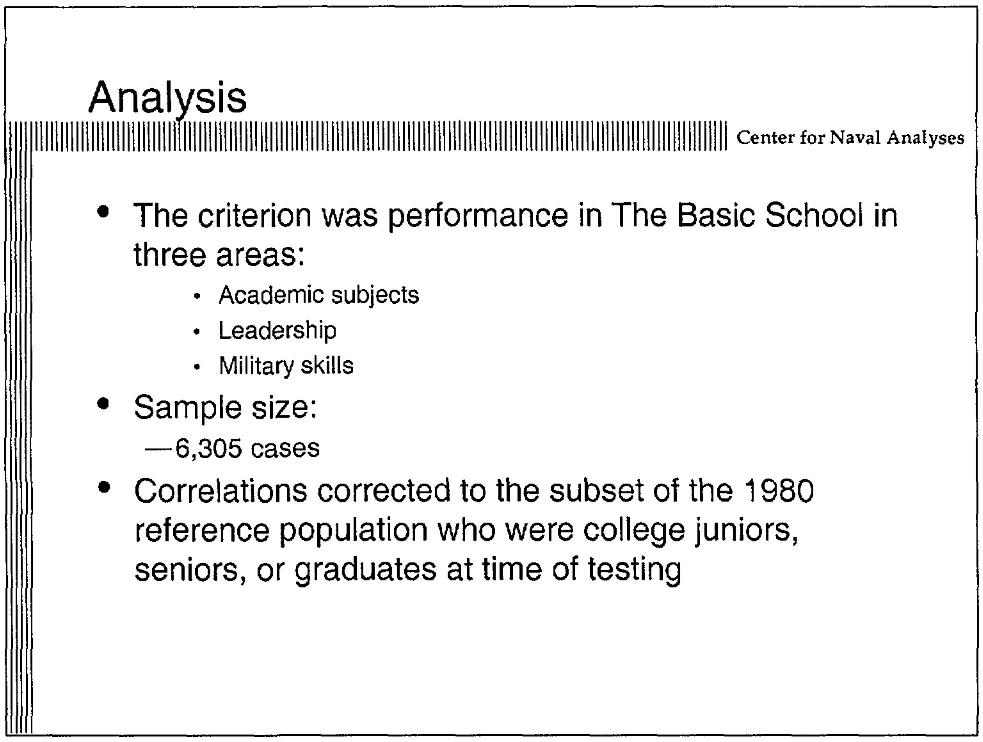
Correlations Between SAT and ASVAB EL

Center for Naval Analyses

Variables	Sample correlation	Full-range correlation
GCT:SAT	.56	.87
GCT:SAT	.47	.89
TBS academic:SAT	.31	.65
TBS academic:EL	.24	.66
TBS academic:GCT	.37	.57

Catherine M. Hiatt and William H. Sims. *Equivalent Scores on SAT and ASVAB/EL*, Nov 1995, Center for Naval Analyses (CAB 95-20)

The ASVAB is a good predictor of academic performance in The Basic School (TBS). TBS is an extensive course of training in academics, leadership, and military skills. The correlation between TBS academic performance and ASVAB EL is 0.66 and compares favorably with the 0.65 observed for officers accessed on the SAT.

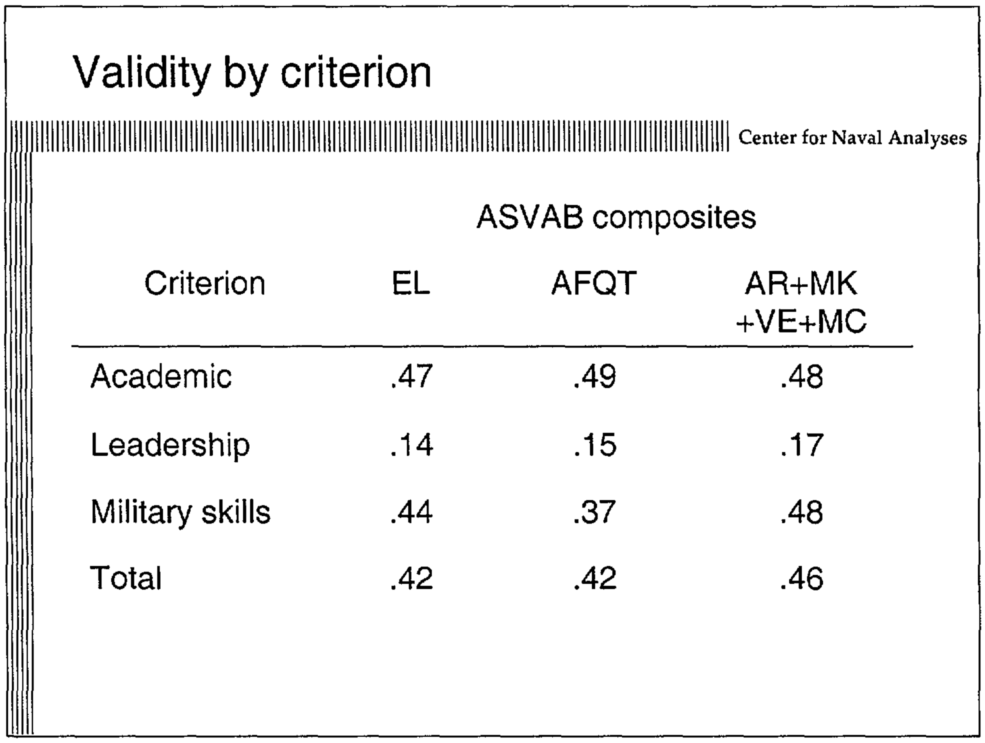


Analysis

Center for Naval Analyses

- The criterion was performance in The Basic School in three areas:
  - Academic subjects
  - Leadership
  - Military skills
- Sample size:
  —6,305 cases
- Correlations corrected to the subset of the 1980 reference population who were college juniors, seniors, or graduates at time of testing

The criterion was performance in academic subjects, leadership, and military skills. All correlations were corrected to the subset of the 1980 reference population who were college juniors, seniors, or graduates at the time of testing.



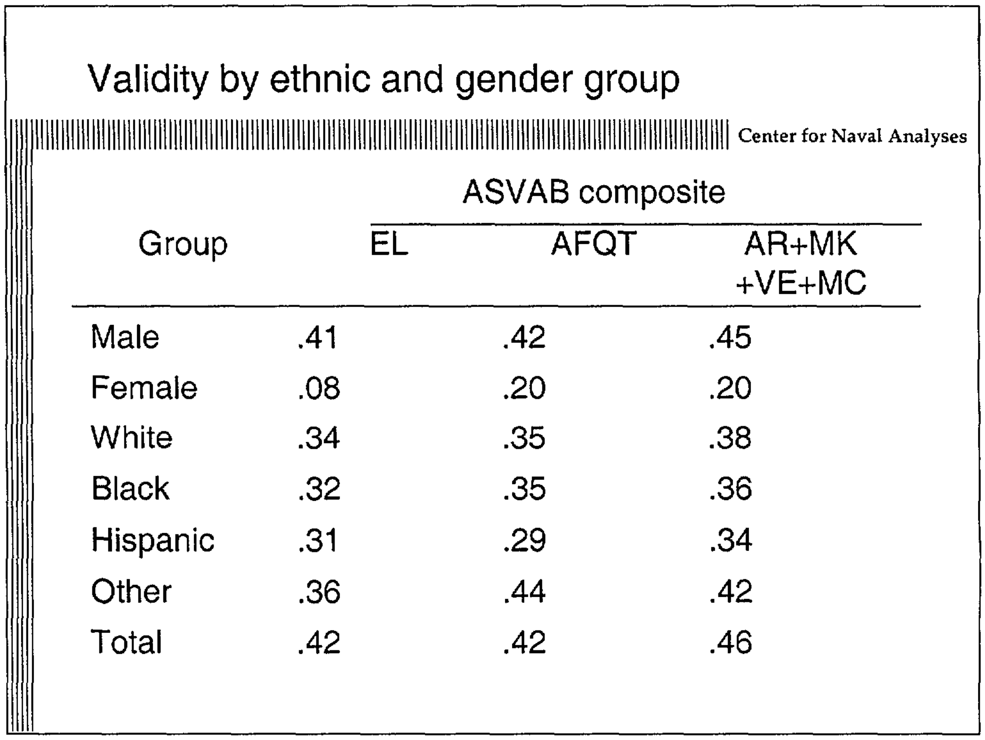
Validity by criterion

Center for Naval Analyses

ASVAB composites

Criterion	EL	AFQT	AR+MK +VE+MC
Academic	.47	.49	.48
Leadership	.14	.15	.17
Military skills	.44	.37	.48
Total	.42	.42	.46

The best predictive composite was found to be AR+MK+VE+MC. This combines strong subtests in math, verbal, and mechanical comprehension areas. This composite is slightly superior to alternatives such as the current EL or to the AFQT.

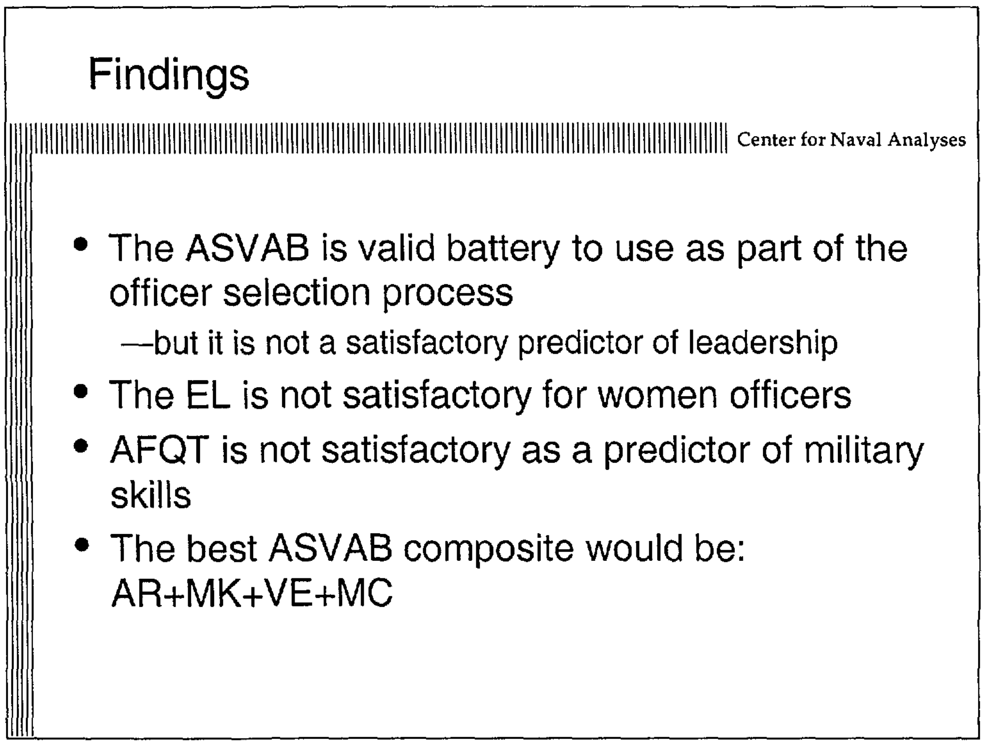


Validity by ethnic and gender group

Center for Naval Analyses

Group	ASVAB composite EL	 AFQT	 AR+MK +VE+MC
Male	.41	.42	.45
Female	.08	.20	.20
White	.34	.35	.38
Black	.32	.35	.36
Hispanic	.31	.29	.34
Other	.36	.44	.42
Total	.42	.42	.46

The recommended composite has somewhat higher validity than alternatives for most ethnic and gender subgroups.



Findings

Center for Naval Analyses

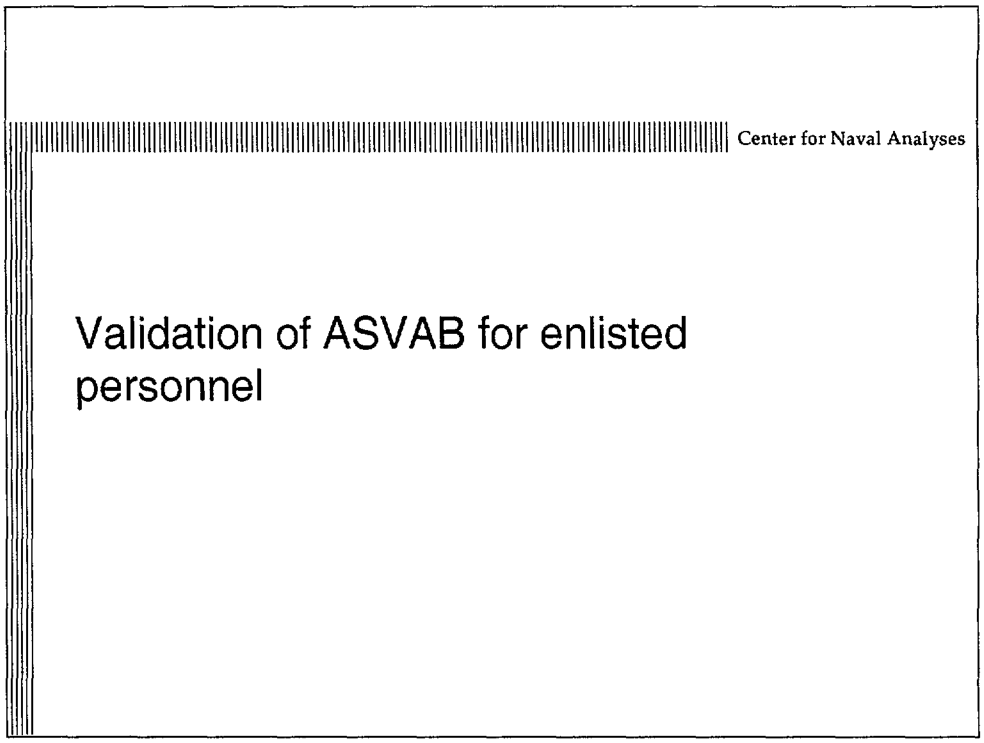
- The ASVAB is valid battery to use as part of the officer selection process
  —but it is not a satisfactory predictor of leadership
- The EL is not satisfactory for women officers
- AFQT is not satisfactory as a predictor of military skills
- The best ASVAB composite would be: AR+MK+VE+MC

We conclude that the ASVAB is a valid aptitude battery to use a part of the officer selection process. However, as expected, it is not a good predictor of leadership.

The ASVAB EL composite does a poor job of predicting performance of female officers.

The AFQT (used for general selector for enlisted personnel) is not a satisfactory predictor of officer military skills.

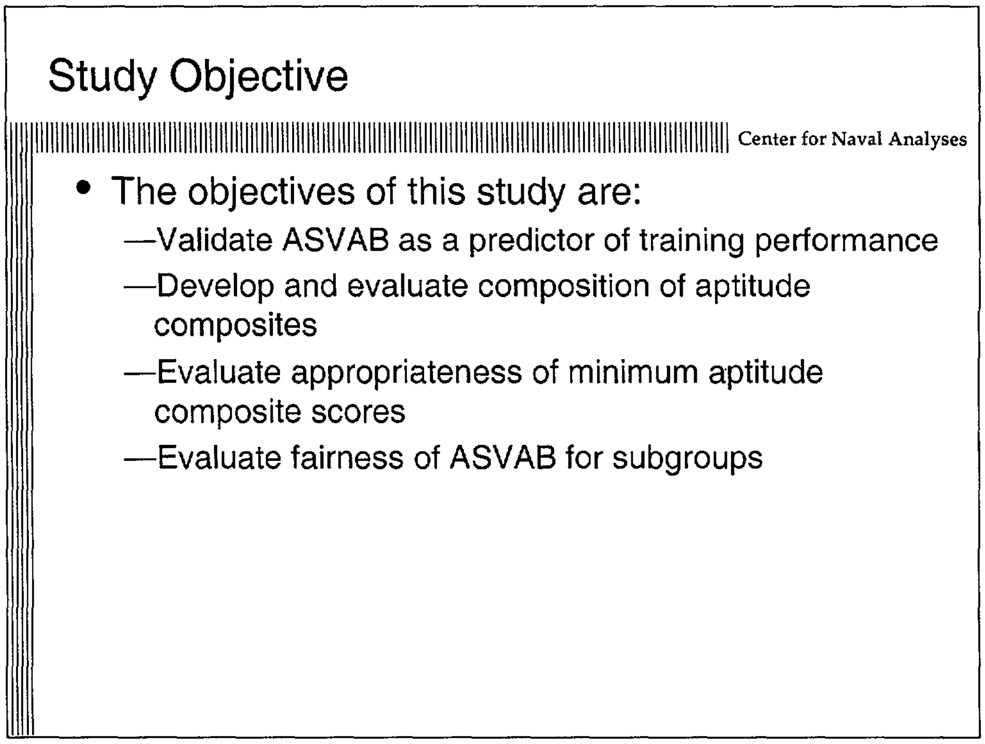
The best ASVAB composite is AR+MK+VE+MC.



Center for Naval Analyses

Validation of ASVAB for enlisted personnel

This section of the briefing describes some recent work by Paul Mayberry and Catherine Hiatt on the validation of ASVAB for the selection and classification of enlisted personnel.



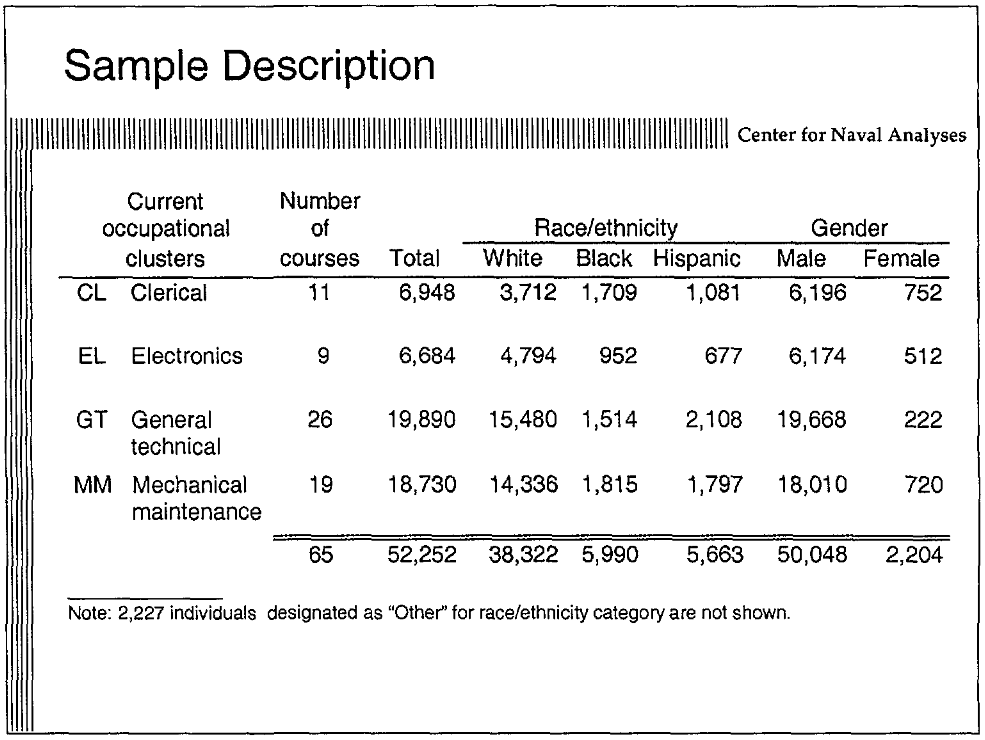
Study Objective

Center for Naval Analyses

- The objectives of this study are:
  - —Validate ASVAB as a predictor of training performance
  - —Develop and evaluate composition of aptitude composites
  - —Evaluate appropriateness of minimum aptitude composite scores
  - —Evaluate fairness of ASVAB for subgroups

The objectives of the study were:

- •Validate ASVAB as a predictor of training performance
- •Develop and evaluate composition of aptitude composites
- •Evaluate appropriateness of minimum aptitude composite scores
- •Evaluate fairness of ASVAB for subgroups



Sample Description

Center for Naval Analyses

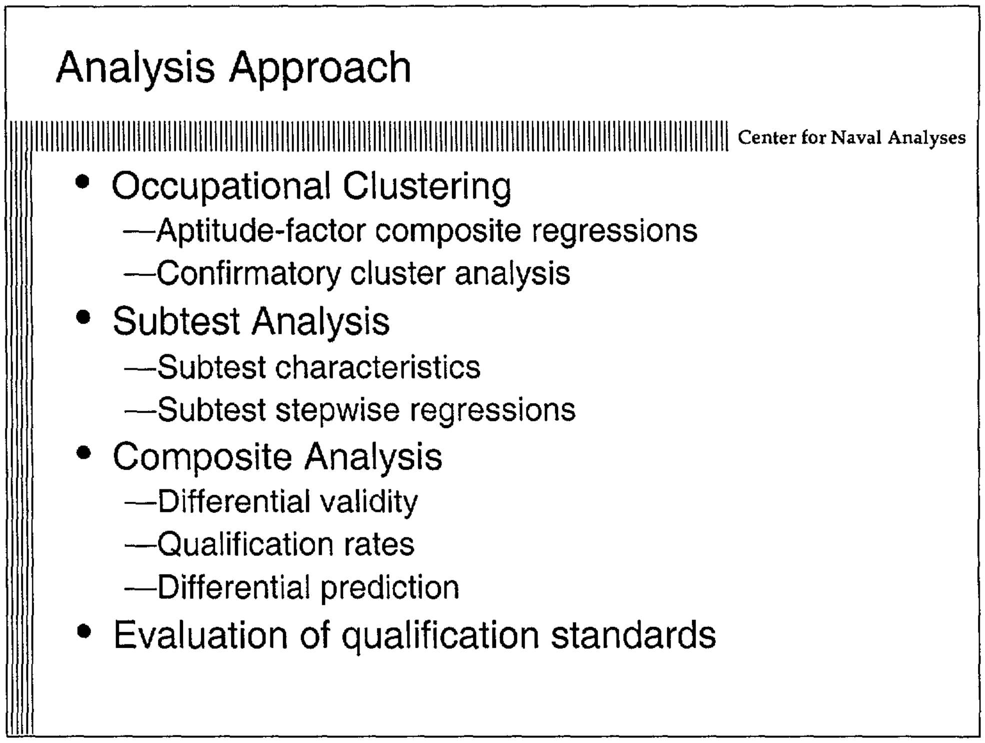
	Current occupational clusters	Number of courses	Total	Race/ethnicity			Gender	
				White	Black	Hispanic	Male	Female
CL	Clerical	11	6,948	3,712	1,709	1,081	6,196	752
EL	Electronics	9	6,684	4,794	952	677	6,174	512
GT	General technical	26	19,890	15,480	1,514	2,108	19,668	222
MM	Mechanical maintenance	19	18,730	14,336	1,815	1,797	18,010	720
		65	52,252	38,322	5,990	5,663	50,048	2,204

Note: 2,227 individuals designated as "Other" for race/ethnicity category are not shown.

We attempted to collect data on all Marines who completed entry-level training within the last three fiscal years. Data were required for both active duty and reserves.

We conducted a variety of data quality analyses. The data were edited to exclude cases that had multiple training records (only the first attempt was included), individuals who were lateral moves into a different occupational specialty, or persons who had been dropped from training for nonacademic reasons. We examined the consistency of training grades over time to ensure proper interpretation. We aggregated courses with small samples based on similarity of selection requirements and functional requirements in training and on the job. We also conducted residual analyses to identify possible outliers and to verify that basic regression assumptions were satisfied.

This slide shows the number of courses within each of the current Marine Corps occupational clusters and the associated sample sizes that were used for analysis after completion of the data quality checks and data editing. We collected complete training and aptitude information for more than 52,000 students.



Analysis Approach

Center for Naval Analyses

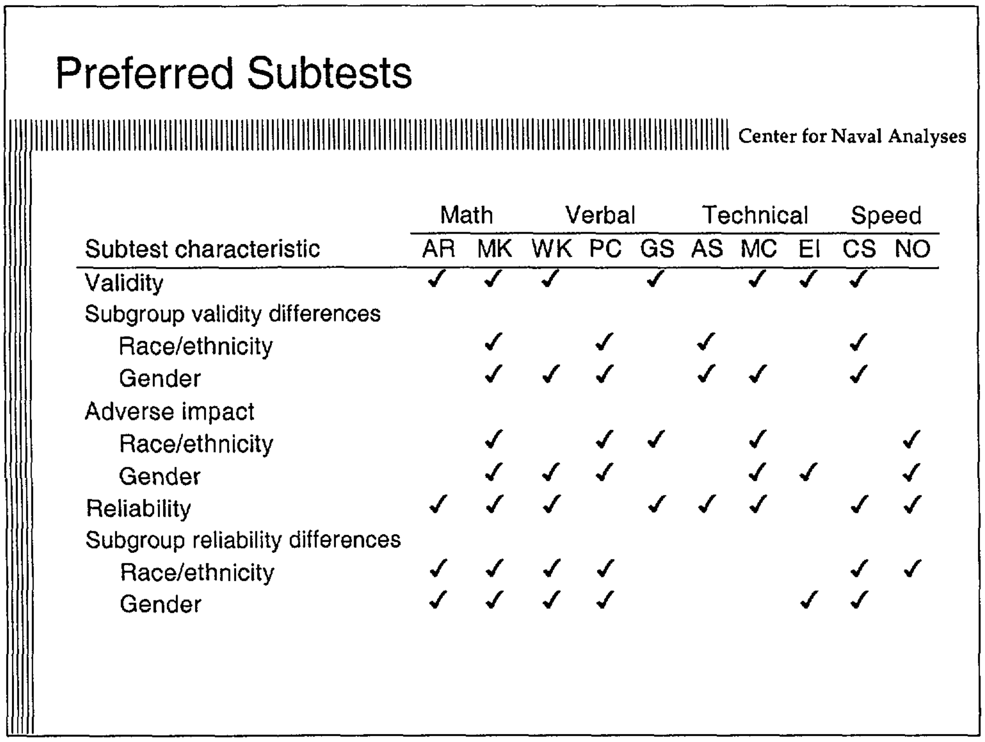
- Occupational Clustering
  - —Aptitude-factor composite regressions
  - —Confirmatory cluster analysis
- Subtest Analysis
  - —Subtest characteristics
  - —Subtest stepwise regressions
- Composite Analysis
  - —Differential validity
  - —Qualification rates
  - —Differential prediction
- Evaluation of qualification standards

This slide details our analysis approach. First, we sought to cluster training courses based on similarity of their prerequisite aptitude requirements. We conducted aptitude-factor regressions to determine which aptitudes were necessary for successful performance in each course. We used cluster analysis to confirm our groupings of occupations.

Second, we examined a variety of subtest characteristics with the intent of forming composites that had high validity, minimal differences in validity across subgroups, limited adverse impact, and high reliability. Often these characteristics were inversely related, which required specific tradeoff decisions. Based on these characteristics and subtest stepwise regressions, we proposed a large number of alternative composite definitions that required further analysis.

Third, for each of the alternative composite definitions, we examined absolute validity, differential validity, minority qualification rates and adverse impact, as well as differential prediction. From these analyses, we proposed a new set of composites for the Marine Corps.

Finally, for the newly proposed aptitude composites, we evaluated qualification standards by considering the aptitude/training-grade relationship and quantifying current Marine Corps policy concerning performance requirements. We conducted a variety of sensitivity analyses of relevant variables to assess their impact on the final model solutions.



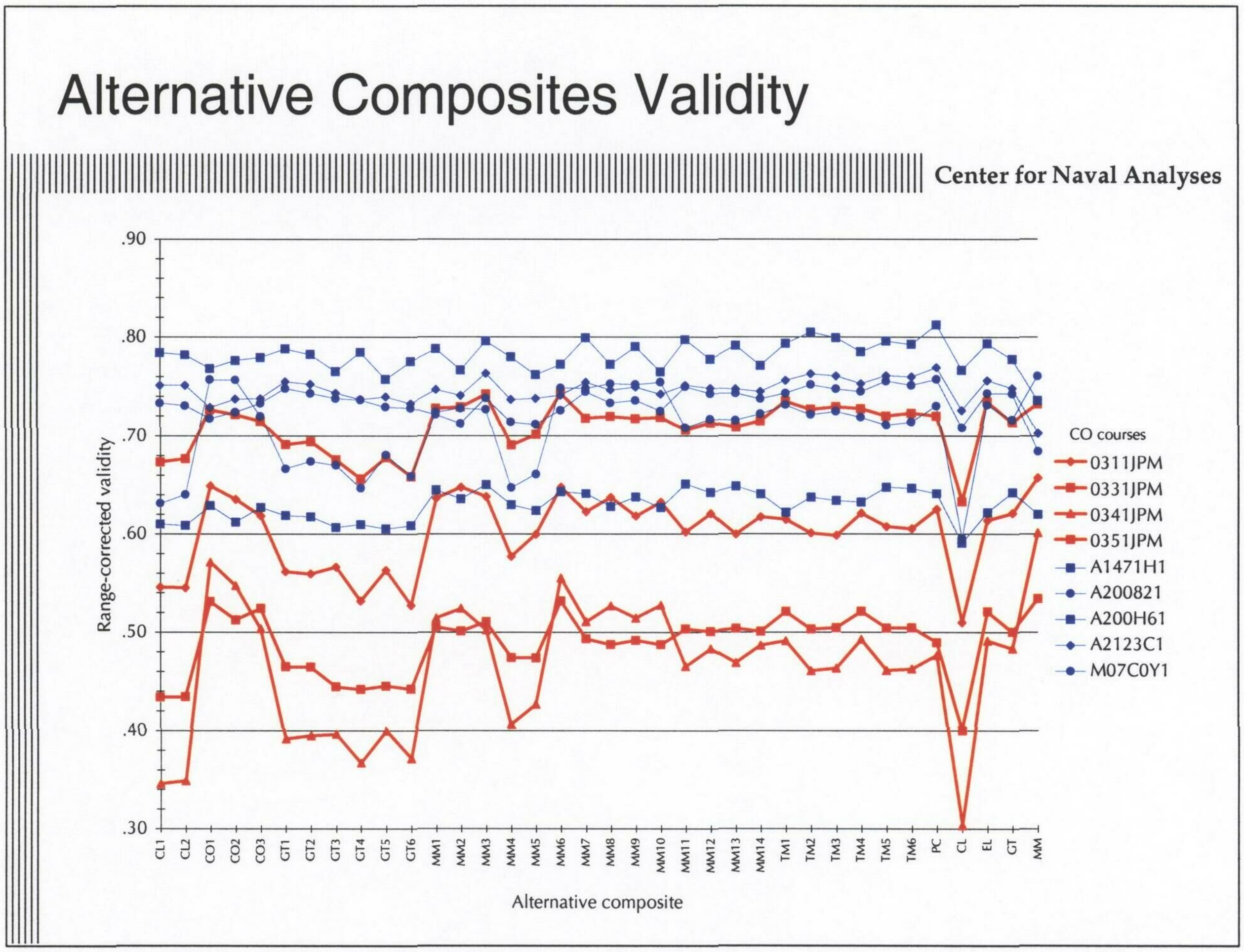
Preferred Subtests

Center for Naval Analyses

	Math		Verbal			Technical			Speed	
Subtest characteristic	AR	MK	WK	PC	GS	AS	MC	EI	CS	NO
Validity	✓	✓	✓		✓		✓	✓	✓	
Subgroup validity differences										
Race/ethnicity		✓		✓		✓			✓	
Gender		✓	✓	✓		✓	✓		✓	
Adverse impact										
Race/ethnicity		✓		✓	✓		✓			✓
Gender		✓	✓	✓			✓	✓		✓
Reliability	✓	✓	✓		✓	✓	✓		✓	✓
Subgroup reliability differences										
Race/ethnicity	✓	✓	✓	✓					✓	✓
Gender	✓	✓	✓	✓				✓	✓	

Determining the specific subtests to be used in the formation of a composite involves multiple considerations. In many cases, these considerations are diametrically opposed. This slide summarizes the general trends of preferred subtests noted in the previous discussions.

The specification of "preferred" is based on differences in characteristics across subtests within each aptitude factor. In general, we wanted to maximize subtest validity and reliability while minimizing these quantities with respect to subgroup differences. Our development of composite alternatives was primarily based on these findings. We considered 31 alternative composite definitions as well as the current Marine Corps composites.



Alternative Composites Validity

Center for Naval Analyses

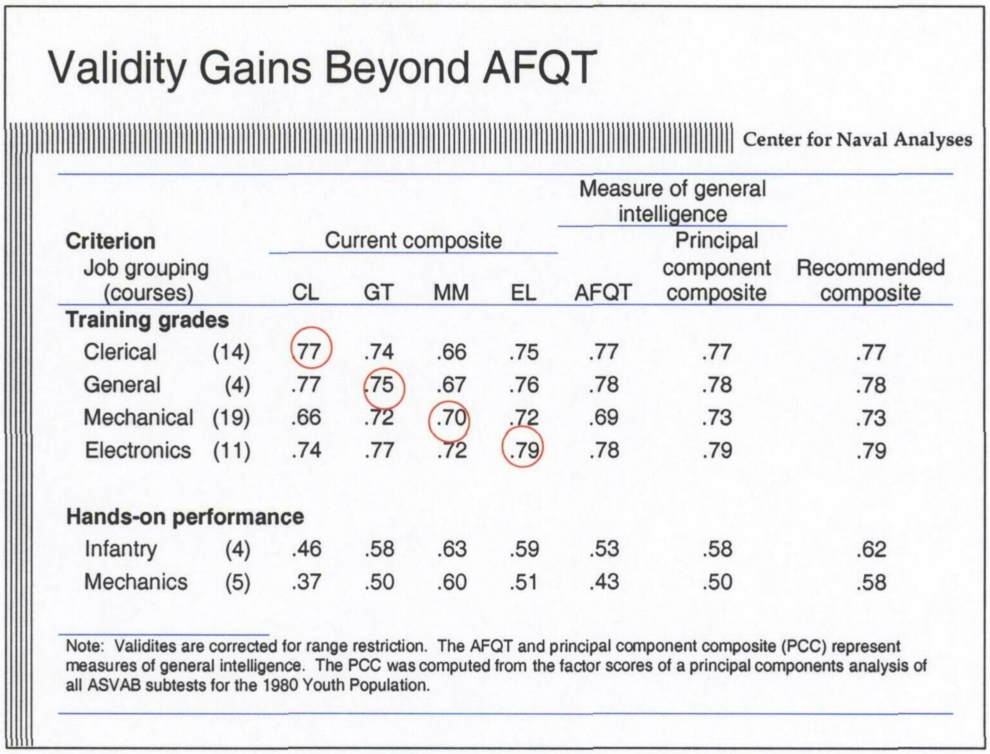
Based on the subtest characteristic information presented earlier, we developed 31 alternative composite definitions as noted on the horizontal axis in this slide. We also used the current Marine Corps composites as a baseline for comparison.

For each occupational grouping of courses, we examined the validities both within and across aptitude areas. For example, this slide notes the findings for the combat (CO) courses. Within the CO composite alternatives, we sought to determine the definition that was consistently the highest. We also sought to determine the most valid definition other than one of the CO alternatives. In this way, we confirmed the correct categorization of courses into their respective aptitude area and also whether other composite definitions could achieve comparable validity.

The magnitude of the range-corrected validities was consistently high. The lower validities in this slide are for the infantry courses where hands-on performance measures were substituted for training grades.

The most striking finding was the lack of differences in validity despite the composite definition. The largest differences generally involved either the clerical courses or the clerical composite definitions. Against the criterion of training grades, there was limited differential validity across the composite definitions.

For the infantry courses with hands-on performance measures as the criterion, however, there is substantially more scatter in the range of validities. While changes continue to be made in the instruction and assessment of training outcomes, success in most schools is still very academically oriented. Therefore, it is difficult for the ASVAB to demonstrate substantial differential validity against such a constrained criterion.



Validity Gains Beyond AFQT

Center for Naval Analyses

						Measure of general intelligence		
Criterion		Current composite						
Job grouping (courses)		CL	GT	MM	EL	AFQT	Principal component composite	Recommended composite
Training grades								
Clerical	(14)	.77	.74	.66	.75	.77	.77	.77
General	(4)	.77	.75	.67	.76	.78	.78	.78
Mechanical	(19)	.66	.72	.70	.72	.69	.73	.73
Electronics	(11)	.74	.77	.72	.79	.78	.79	.79
Hands-on performance								
Infantry	(4)	.46	.58	.63	.59	.53	.58	.62
Mechanics	(5)	.37	.50	.60	.51	.43	.50	.58

Note: Validites are corrected for range restriction. The AFQT and principal component composite (PCC) represent measures of general intelligence. The PCC was computed from the factor scores of a principal components analysis of all ASVAB subtests for the 1980 Youth Population.

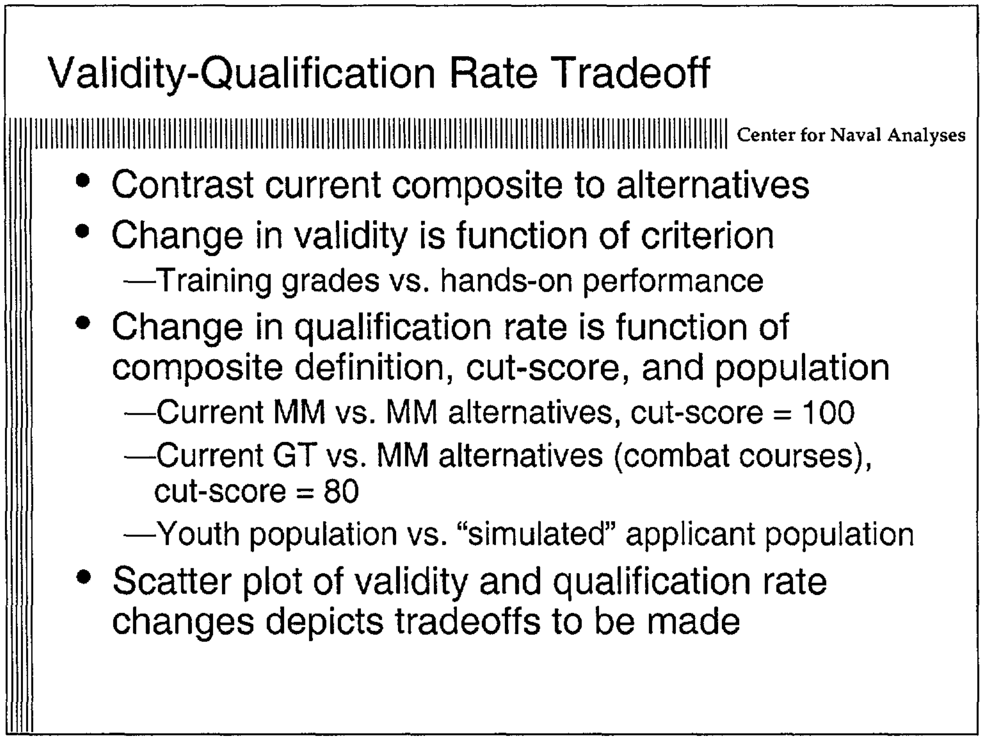
Some researchers have argued that the Armed Forces Qualification Test (AFQT) is a sufficient predictor of military performance; that is, the formation of other composites does not result in improved predictions beyond what AFQT is able to achieve. AFQT is considered by some to be a measure of general intelligence ("G"). Because AFQT is not based on all of the ASVAB subtests, we calculated an alternative G measure: the principal component composite (based on the 1980 Youth Population data). We contrasted the validity findings for this principal component composite to the AFQT and current Marine Corps composites.

The table above shows that the recommended composites of this study will result in improved or equal validities against training grades. However, against a hands-on criterion, the recommended composites were noted to have slightly lower validities than the appropriate current composite. Again, the selection of the criterion makes a difference in terms of composite formation.

There is only conditional evidence for the argument that AFQT is a sufficient predictor. Alternative composites for the less technical courses (clerical and general) cannot incrementally improve the relationship between AFQT and training grades. However, moderate improvements can be made in the prediction of training grades by the recommended composites for the more technical courses (mechanical maintenance and electronics repair).

Likewise, the results show that the performance criterion makes a considerable difference in the validity gain outcomes. For those specialties that had hands-on performance information, neither the AFQT nor the principal component composite were able to achieve the validity levels noted for the recommended composites (an improvement of .09 to .15 validity point over AFQT). Unfortunately, the Marine Corps does not have hands-on performance information for specialties in the clerical, general, or electronics areas to determine if such outcomes are a function primarily of the criterion or job type.

If one believes in training grades as the ultimate criterion, it seems that one would prefer the principal component composite over all others because it is as valid as the recommended composites and surely has high reliability. If one believes that hands-on performance is the ultimate criterion, then specific composites are preferable because predictions are substantially improved over any measure of general intelligence.



Validity-Qualification Rate Tradeoff

Center for Naval Analyses

- Contrast current composite to alternatives
- Change in validity is function of criterion
  - —Training grades vs. hands-on performance
- Change in qualification rate is function of composite definition, cut-score, and population
  - —Current MM vs. MM alternatives, cut-score = 100
  - —Current GT vs. MM alternatives (combat courses), cut-score = 80
  - —Youth population vs. "simulated" applicant population
- Scatter plot of validity and qualification rate changes depicts tradeoffs to be made

Qualification rates are also an important consideration in the formation of aptitude composites. Historically, the quest for validity improvements has been the sole criterion for determining the subtest composition of aptitude composites. In this study, we have examined the tradeoffs between validity gains versus changes in qualification for a range of new composite alternatives. All new composites must be contrasted to the current composites used. One of the recommendations of our study is to use a mechanical maintenance (MM) composite for infantry courses. These courses currently use the General Technical (GT) composite, so we will also compare the new mechanical maintenance alternatives to the historical base of GT for combat jobs.

As noted earlier, validity varies considerably depending on the performance measure. We will contrast changes in validity relative to both training grades and hands-on performance.

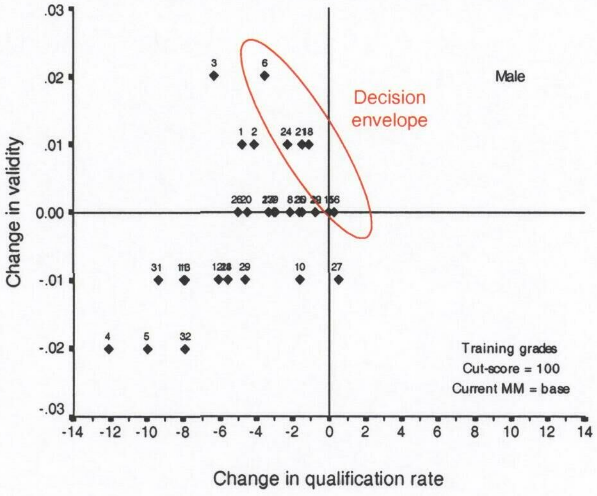
Changes in qualification rates are determined completely independently of any validity information. Qualification rates are completely a function of the composite subtest definition, the cut-score that determines who is and is not qualified, and the population on which these calculations are made. For the mechanical jobs, we will compare the qualification rates for the MM alternatives to the current MM at a cut-score of 100 (the typical qualification standard for most mechanical jobs). For the combat courses, we will examine the qualification rates for the MM alternatives relative to the current GT composite at a cut-score of 80 (again, the typical qualification standard for infantry jobs).

With respect to the population, we calculated qualification rates in the 1980 Youth Population for six subgroups: the cross of three racial groups (white, black, and Hispanic) and gender (male and female). Using these six subgroups as building block and applying appropriate weights, we can determine qualification rates for aggregated subgroups (e.g., males vs. females) in the 1980 Youth Population or other simulated populations for which we know the proportions of these subgroups (e.g., a Marine Corps applicant population).

Finally, we prepared scatter plots to illustrate the empirical relationship between changes in validity and changes in qualification rates and, thereby, to show the tradeoffs involved.

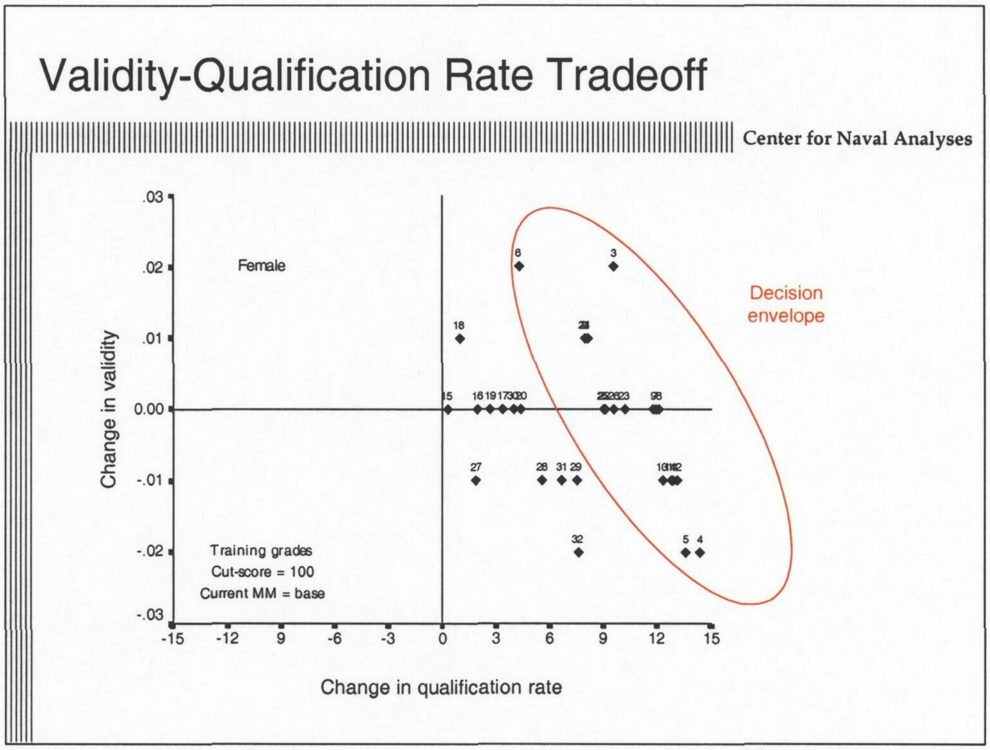
Validity-Qualification Rate Tradeoff

Center for Naval Analyses



This scatter plot represents the results for men in the 1980 Youth Population. Validity calculations were based on training grades, and the current MM composite was the base for all changes. Negative values, on either axis, imply that the current composite is better than the alternative composite. The points on the plot represent the 32 alternative composites being considered as potential replacements for the current MM composite. Appendix A lists the subtests defining each alternative. Changes in validity (based on training grades) have a range of +/- .02 validity point. Similarly, changes in qualification rates range from -12 to +1. The cross-hairs of the axes represent the values for the current MM composite.

Although the overall relationship between change in validity and change in qualification rate is positive, there is an envelope of alternatives representing the feasible solutions. We developed these envelopes based on the following considerations: (1) focusing on the upper right quadrant, while excluding all outcomes in the lower left quadrant, (2) desiring no excessive validity losses, (3) desiring no excessive losses in qualification rates, and (4) applying a modal approach—shooting any major gaps between composite alternatives. Note that the slope of the decision envelope (range of feasible solutions) is negative. This will be a recurring theme.

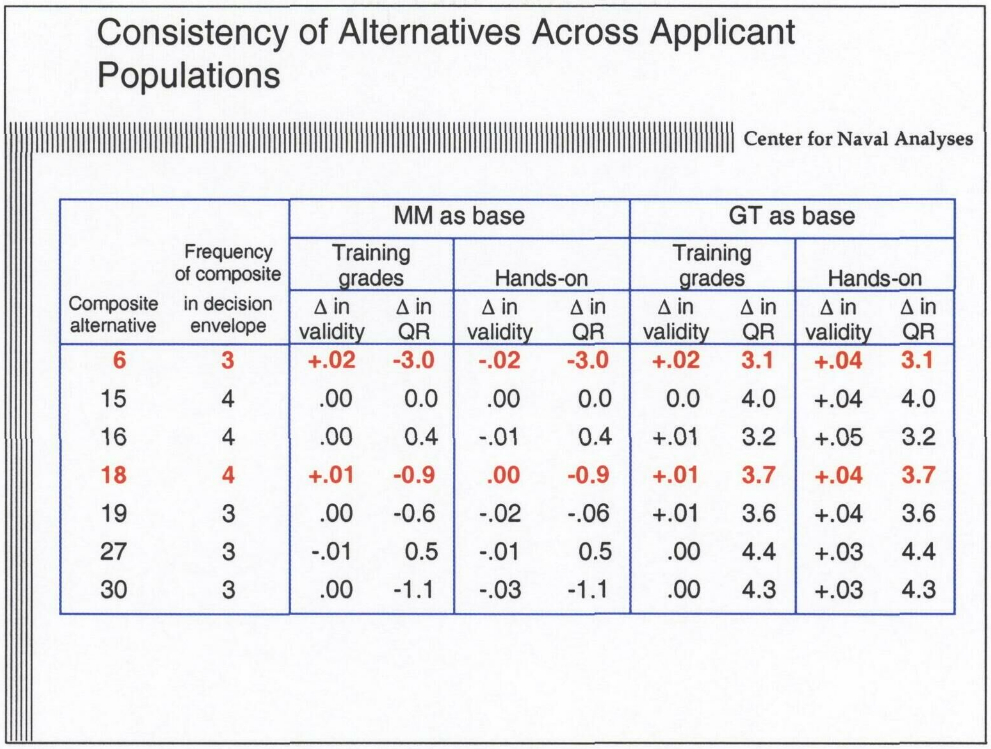


Validity-Qualification Rate Tradeoff

Center for Naval Analyses

The same scatter plot for women shows an entirely different picture. While essentially all of the alternative composites reduced qualification rates for men, the same alternatives net substantial gains in qualification for women. The current MM composite contains one math subtest and all three technical subtests. Any deviations from this highly technical composite result in qualification improvements for women.

These diametrically opposed results (what's good for women is not good for men) present real dilemmas for policy-makers. It also follows that changes made to benefit the minority group may have adverse effects on the majority and thereby increase the recruiting effort required to maintain the same number of accessions.



Consistency of Alternatives Across Applicant Populations

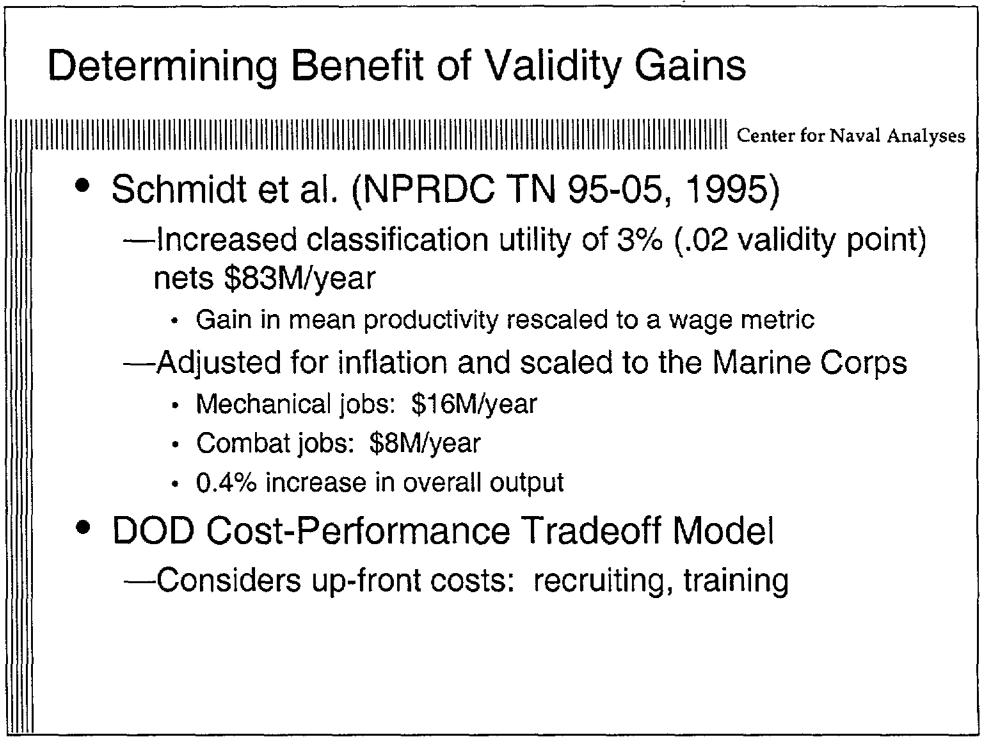
Center for Naval Analyses

		MM as base				GT as base			
	Frequency of composite	Training grades		Hands-on		Training grades		Hands-on	
Composite alternative	in decision envelope	Δ in validity	Δ in QR	Δ in validity	Δ in QR	Δ in validity	Δ in QR	Δ in validity	Δ in QR
6	**3**	**+.02**	**-3.0**	**-.02**	**-3.0**	**+.02**	**3.1**	**+.04**	**3.1**
15	4	.00	0.0	.00	0.0	0.0	4.0	+.04	4.0
16	4	.00	0.4	-.01	0.4	+.01	3.2	+.05	3.2
18	**4**	**+.01**	**-0.9**	**.00**	**-0.9**	**+.01**	**3.7**	**+.04**	**3.7**
19	3	.00	-0.6	-.02	-.06	+.01	3.6	+.04	3.6
27	3	-.01	0.5	-.01	0.5	.00	4.4	+.03	4.4
30	3	.00	-1.1	-.03	-1.1	.00	4.3	+.03	4.3

This table summarizes the consistency of findings for the composite alternatives across the applicant populations. Only those composites that occurred in three or more of the possible four decision envelopes are listed. The magnitudes of both the validity changes and qualification rate changes are noted.

Only alternatives 6 and 18 appear to offer consistent improvements beyond the current MM composite.

The next question involves trying to determine the value, or benefit, associated with these validity gains and the impact of changes in qualification rates.



Determining Benefit of Validity Gains

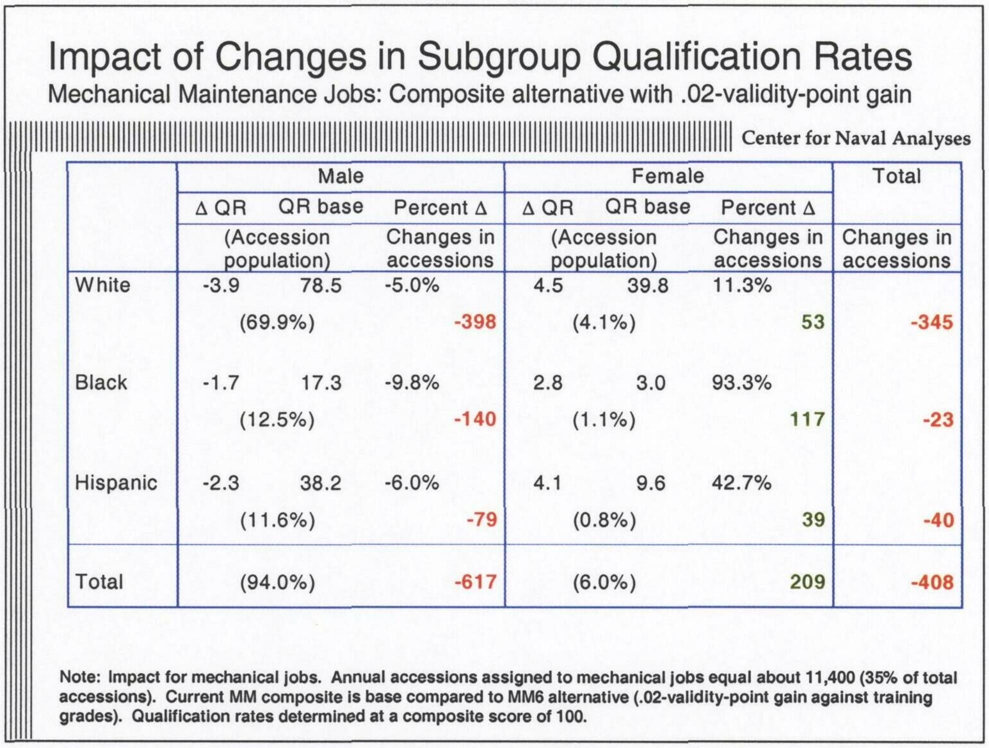
Center for Naval Analyses

- Schmidt et al. (NPRDC TN 95-05, 1995)
  - —Increased classification utility of 3% (.02 validity point) nets $83M/year
    - Gain in mean productivity rescaled to a wage metric
  - —Adjusted for inflation and scaled to the Marine Corps
    - Mechanical jobs: $16M/year
    - Combat jobs: $8M/year
    - 0.4% increase in overall output
- DOD Cost-Performance Tradeoff Model
  - —Considers up-front costs: recruiting, training

In research conducted in support of the Enhanced Computerized Adaptive Testing program, Frank Schmidt and others sought to quantify in dollars the potential benefit associated with validity gains of new predictor tests supplementing the current ASVAB. Based on generalizations from other research and various assumptions about job complexity and wage levels, Schmidt translated gains in mean productivity to a dollar metric. The findings were that a 3-percent increase in classification utility (which reflected a .02 validity gain) would result in approximately an $83-million benefit to the Navy on an annual basis.

Using this value as an estimate of the benefit associated with a .02-point validity gain, we adjusted the figure for inflation and rescaled it to reflect the size of the Marine Corps. Additional proportional adjustments were made to scale this figure to be applicable to mechanical and combat jobs. The estimated benefit was determined to be $16 million and $8 million annually for mechanical and combat jobs, respectively. Both of these figures represent a .4-percent increase in overall output.

While the exact value of these figures may be argued, they do provide a gross estimate of the benefit associated with gains in validity. Similar work being conducted by OASD to update its cost-performance tradeoff model may shed additional light on such costs and benefits.



Impact of Changes in Subgroup Qualification Rates

Mechanical Maintenance Jobs: Composite alternative with .02-validity-point gain

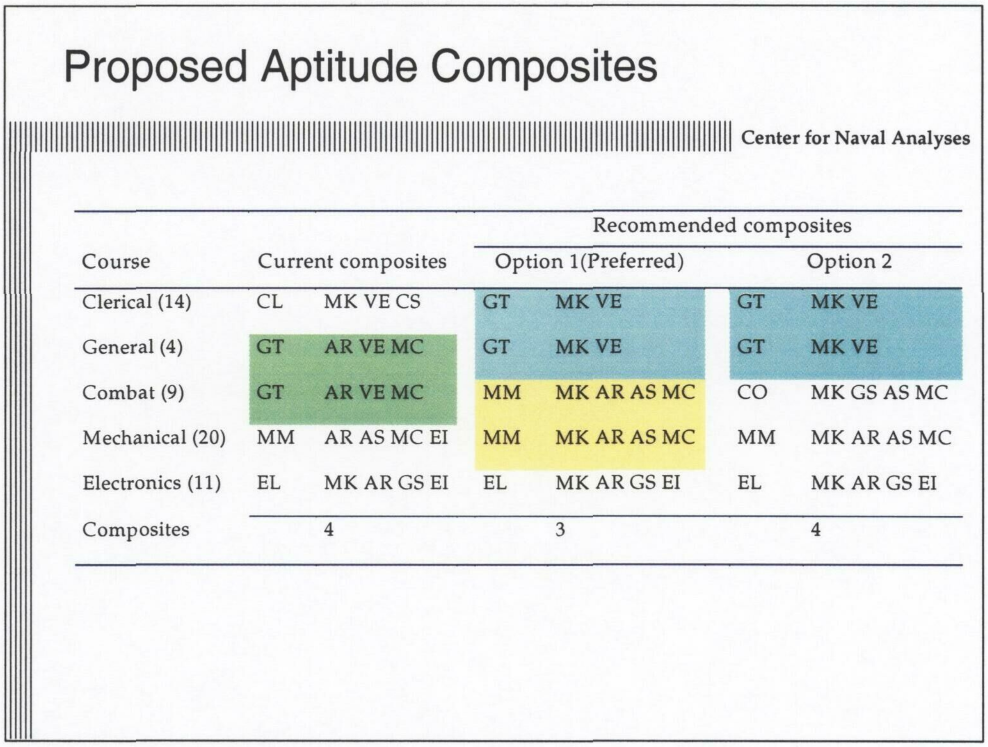
Center for Naval Analyses

	Male			Female			Total
	Δ QR	QR base	Percent Δ	Δ QR	QR base	Percent Δ	
	(Accession population)		Changes in accessions	(Accession population)		Changes in accessions	Changes in accessions
White	-3.9	78.5	-5.0%	4.5	39.8	11.3%	
	(69.9%)		-398	(4.1%)		53	-345
Black	-1.7	17.3	-9.8%	2.8	3.0	93.3%	
	(12.5%)		-140	(1.1%)		117	-23
Hispanic	-2.3	38.2	-6.0%	4.1	9.6	42.7%	
	(11.6%)		-79	(0.8%)		39	-40
Total	(94.0%)		-617	(6.0%)		209	-408

Note: Impact for mechanical jobs. Annual accessions assigned to mechanical jobs equal about 11,400 (35% of total accessions). Current MM composite is base compared to MM6 alternative (.02-validity-point gain against training grades). Qualification rates determined at a composite score of 100.

What is the impact of changes in qualification rates? This slide provides a lot of information but essentially shows the number of accessions gained or lost simply because of the change in composite definition. These calculations are based on the MM6 composite, which was shown to result in a .02-validity-point gain.

This slide shows that larger numbers of women will qualify based on changing from the current MM composite to MM6. However, such a change will also result in fewer men qualifying. The gains in women do not exceed the loss of men, so additional recruiting effort would be required to achieve the same accession goals. These gains and losses should be interpreted relative to the 11,400 annual accessions required for the specialties within the mechanical maintenance occupational fields.

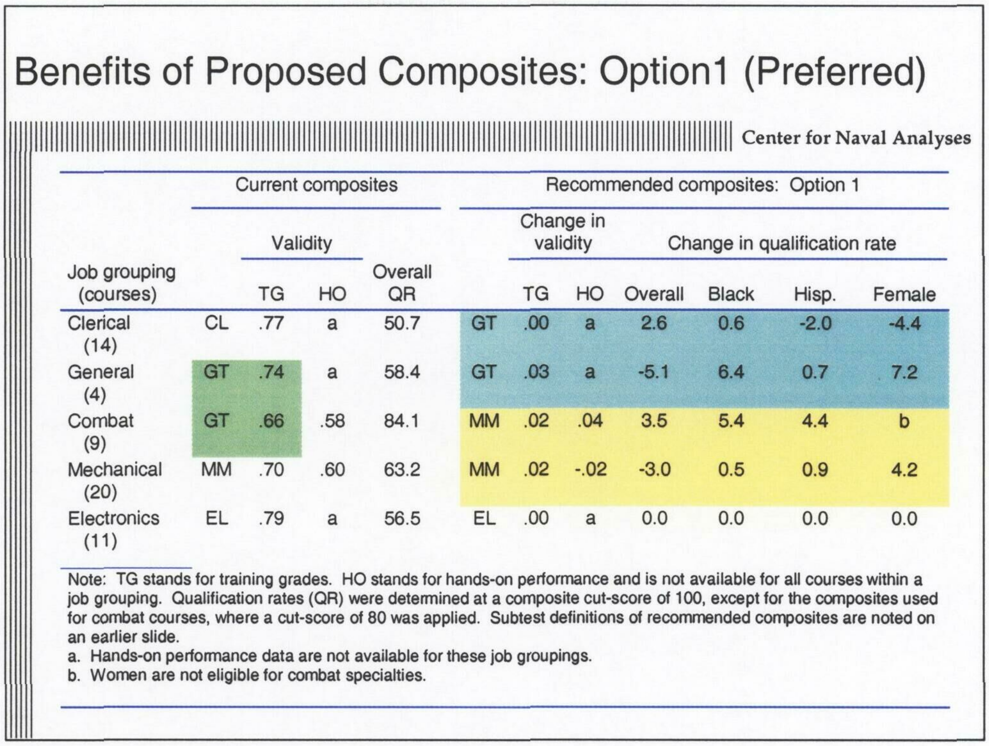


Proposed Aptitude Composites

Center for Naval Analyses

			Recommended composites			
Course	Current composites		Option 1(Preferred)		Option 2	
Clerical (14)	CL	MK VE CS	GT	MK VE	GT	MK VE
General (4)	GT	AR VE MC	GT	MK VE	GT	MK VE
Combat (9)	GT	AR VE MC	MM	MK AR AS MC	CO	MK GS AS MC
Mechanical (20)	MM	AR AS MC EI	MM	MK AR AS MC	MM	MK AR AS MC
Electronics (11)	EL	MK AR GS EI	EL	MK AR GS EI	EL	MK AR GS EI
Composites		4		3		4

Given the previous considerations, we propose the following options for recommended aptitude composites. First, the GT composite would be modified and would be used for clerical as well as general technical courses. Second, no changes would be made to the the EL composite. Third, there are two options for the mechanical and combat courses. The first option is our preferred choice that uses the same MM composite for both the mechanical and combat courses. An alternative would be to develop a specific combat (CO) composite while using the revised MM composite for MM courses.



Benefits of Proposed Composites: Option1 (Preferred)

Center for Naval Analyses

	Current composites				Recommended composites: Option 1						
		Validity				Change in validity		Change in qualification rate			
Job grouping (courses)		TG	HO	Overall QR		TG	HO	Overall	Black	Hisp.	Female
Clerical (14)	CL	.77	a	50.7	GT	.00	a	2.6	0.6	-2.0	-4.4
General (4)	GT	.74	a	58.4	GT	.03	a	-5.1	6.4	0.7	7.2
Combat (9)	GT	.66	.58	84.1	MM	.02	.04	3.5	5.4	4.4	b
Mechanical (20)	MM	.70	.60	63.2	MM	.02	-.02	-3.0	0.5	0.9	4.2
Electronics (11)	EL	.79	a	56.5	EL	.00	a	0.0	0.0	0.0	0.0

Note: TG stands for training grades. HO stands for hands-on performance and is not available for all courses within a job grouping. Qualification rates (QR) were determined at a composite cut-score of 100, except for the composites used for combat courses, where a cut-score of 80 was applied. Subtest definitions of recommended composites are noted on an earlier slide.

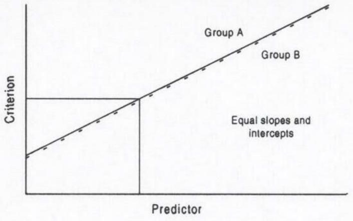
a. Hands-on performance data are not available for these job groupings.

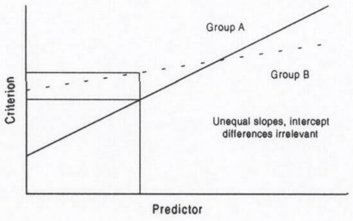
b. Women are not eligible for combat specialties.

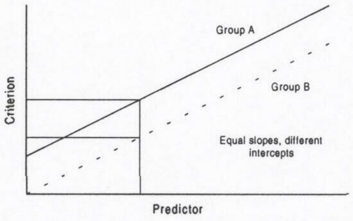
In general, changes from the current composites to the recommended composites result in improved predictions of training success—by .02 to .03 validity point. Large improvements in qualification rates can also be made for women (+4.2 points in mechanical maintenance jobs and +7.2 points in general technical jobs), blacks (+6.4 points in general technical jobs and +5.1 points in combat jobs), and Hispanics (+4.4 points in combat jobs). The "cost" associated with these benefits is that fewer white men will qualify in both general technical and mechanical maintenance jobs. Given the preponderance of white men in any accession population, these changes in composite definitions may require more recruiting effort to achieve the same overall level of total accessions.

Differential Prediction

Center for Naval Analyses





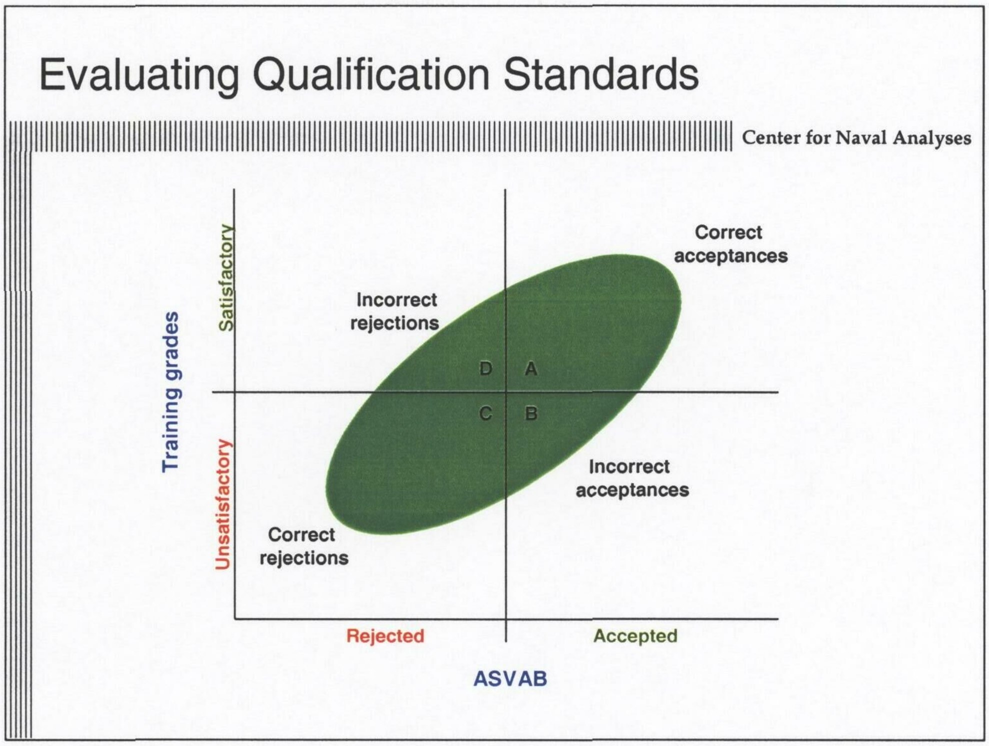


- Race/ethnicity comparisons (N=25)
  - — 17 equal predictions
  - — 2 unequal slopes
  - — 6 unequal intercepts
    - No consistent under prediction
    - Effect size less than 1/3 sd
- Gender comparisons (N=20)
  - — 15 equal prediction
  - — 1 unequal slope
  - — 4 unequal intercepts
    - No consistent under prediction
    - Effect size less than 1/4 sd

We restricted our analysis of differential prediction to those courses that had sufficient sample sizes (i.e., subgroup samples greater than 50). We then sequentially tested the commonality of regression slopes and intercepts.

For race/ethnicity comparisons, we found equal slopes and intercepts for over two-thirds of the courses. For two courses, we found differences in slopes. Further analysis determined that the current qualification standard for these two courses fell in a region of significant intercept differences. For these two courses and the others with significant intercept differences, no subgroups were found to be consistently over predicted. The intercept differences represented less than a third of a standard deviation in training grades. The level of under prediction translated into less than 5 aptitude composite points.

Similar results were noted for the gender comparisons, with over 70 percent of the courses demonstrating equality of slopes and intercepts. The noted differences in intercepts were typically in the range of a quarter of a standard deviation for training grades, and less than 5 aptitude composite points.



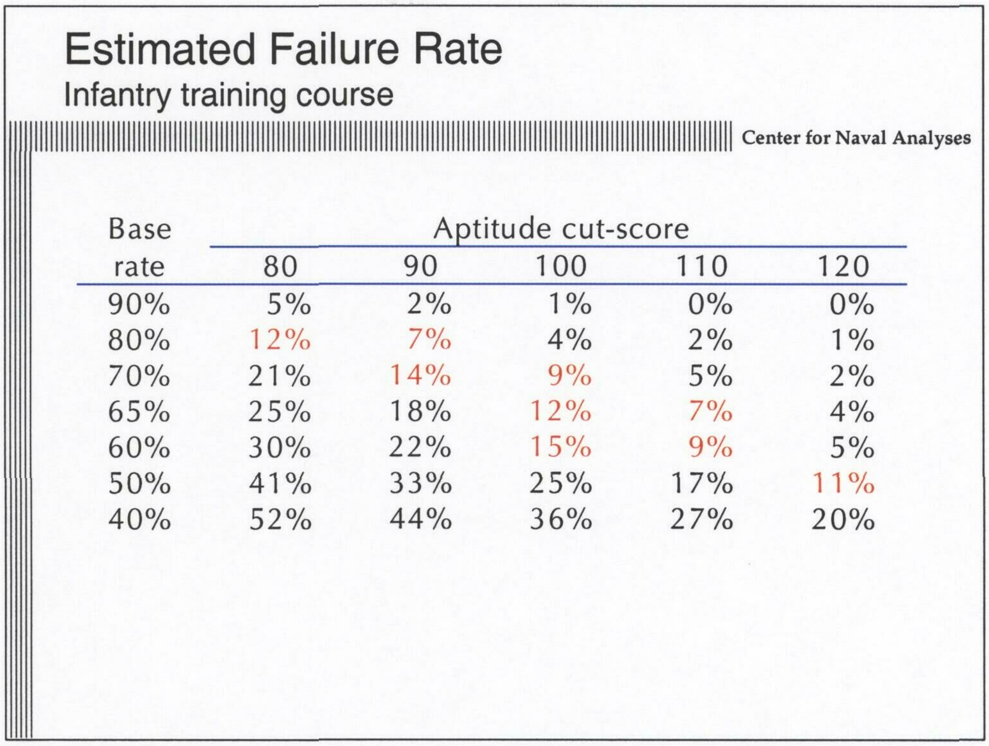
Evaluating Qualification Standards

Center for Naval Analyses

Our validity findings provide strong empirical evidence for the appropriateness of using the ASVAB to predict training success. However, the results simply imply that more is better: more aptitude results in better training performance. Validation research provides only limited information about the utility of selection tests in the decision-making process of determining who is qualified for service and who is not. Therefore, we conducted a series of analyses focused on illustrating the impact of aptitude cut-scores on training outcomes.

Our approach is an extension of earlier work in decision theory and utility analysis by Taylor-Russell, who tabulated the relationships between selection ratios, base rates, and test validity. As opposed to making strong assumptions about the distributions of performance and aptitude, we simply computed predicted training grades for each member of the 1980 Youth Population for each training course and then applied varying performance requirements to determine the outcomes associated with various levels of aptitude qualification standards. This slide illustrates the decision outcomes associated with any selection and classification process.

The initial step in this model is to specify a performance requirement that distinguishes satisfactory from unsatisfactory performers. Akin to most training courses, there is some passing score, above which persons can demonstrate appropriate proficiency, and below which persons require remedial training or are reassigned. One means of establishing this performance requirement is to specify the percentage of the applicant population that would perform successfully if no aptitude test were used to select them. This percentage is commonly referred to as the base rate. In earlier research based on World War II information on the Army General Classification Test for civilians entering military service, base rates for infantrymen were about 80 percent, for automotive mechanics about 65 percent, and for radio repairmen about 40 to 50 percent. Using these three specialties as general markers, base rates for other courses may be estimated.



Estimated Failure Rate

Infantry training course

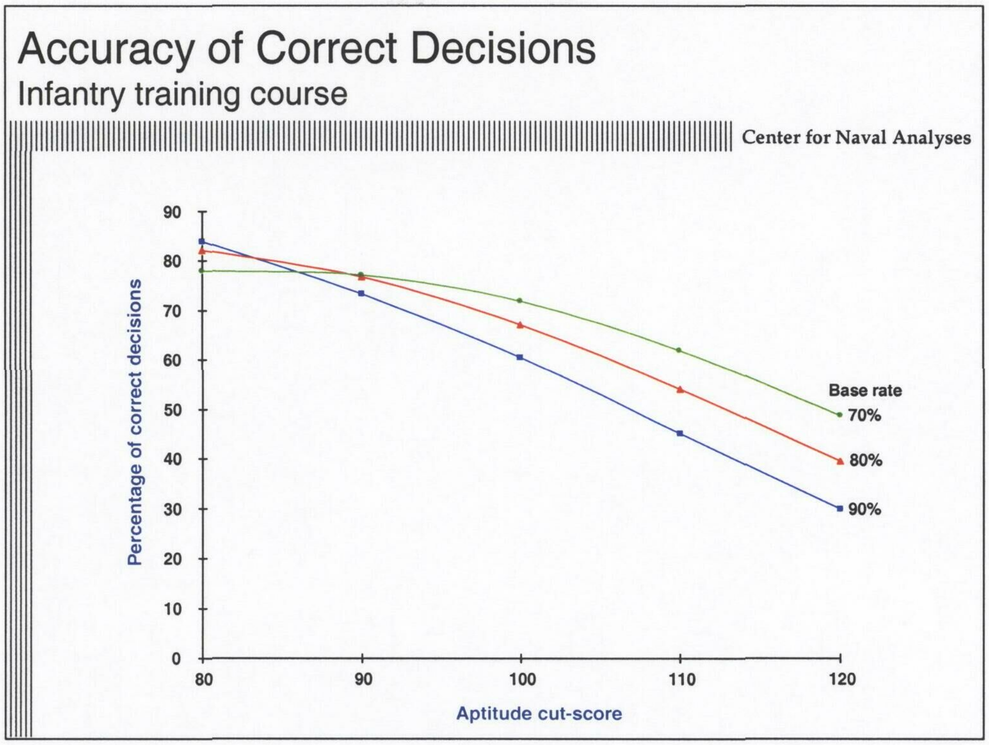
Center for Naval Analyses

Base	Aptitude cut-score				
rate	80	90	100	110	120
90%	5%	2%	1%	0%	0%
80%	12%	7%	4%	2%	1%
70%	21%	14%	9%	5%	2%
65%	25%	18%	12%	7%	4%
60%	30%	22%	15%	9%	5%
50%	41%	33%	25%	17%	11%
40%	52%	44%	36%	27%	20%

The Marine Corps' policy concerning failure rates has been that academic attrition should not exceed 10 percent. Although there has been some flexibility in the application of this rule of thumb, schools have noted that undue academic problems, recycles, and eventual reduced job performance resulted when this failure rate was consistently surpassed. If schools have been able to demonstrate that failure rates are greater than 10 percent, the qualification standards for those courses have received specific attention of Headquarters, Marine Corps.

This slide shows estimated failure rates for the infantry course across a variety of base rates and for varying levels of aptitude. As noted earlier, an approximate base rate for infantrymen is 80 percent, which implies a failure rate of 20 percent. According to Marine Corps policy, a failure rate of 20 percent is too high. Given the positive relationship between ASVAB and training success, this failure rate can be reduced by imposing aptitude minimums. Successively higher aptitude standards will result in correspondingly lower failure rates. Such aptitude standards are progressively raised until a failure rate close to 10 percent is achieved.

In this example for infantry courses, using a base rate of 80 percent results in an aptitude standard between 80 and 90. Alternative base rates can also be considered to determine the impact on the resulting standard.



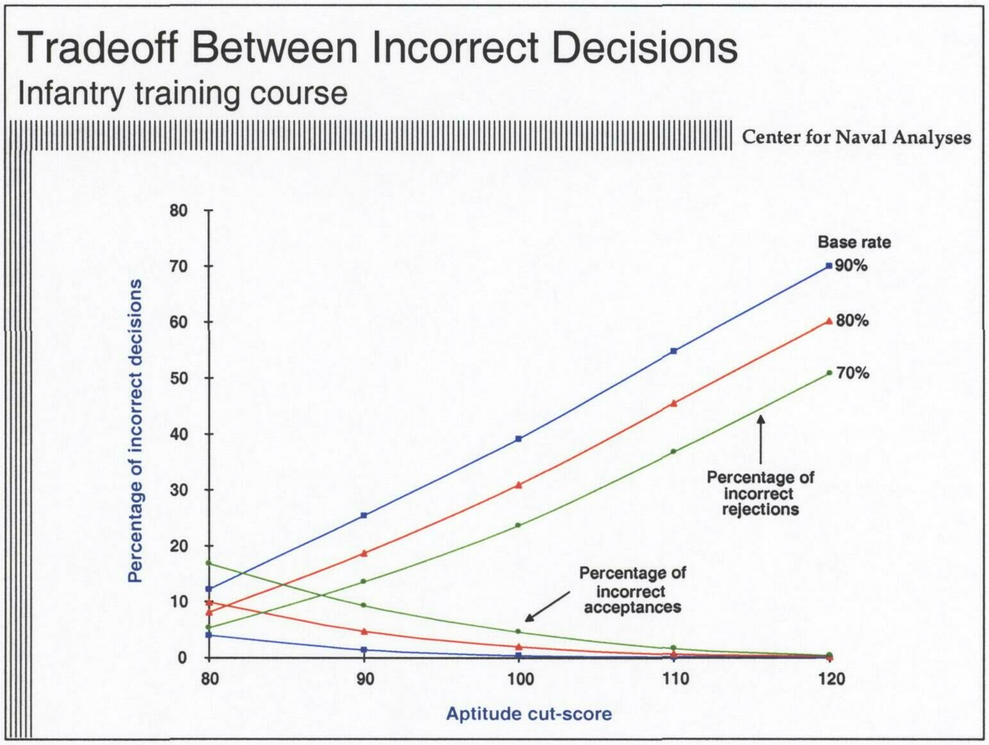
Accuracy of Correct Decisions

Infantry training course

Center for Naval Analyses

A second desired outcome is to maximize the number of correct decisions made in the selection process. For a given base rate, as the aptitude cut-score is changed, the number of correct decisions (i.e., those individuals who are accepted and are successful performers and those who are rejected and would have been unsuccessful performers) also changes. As the aptitude cut-score increases to reduce failure rate, the number of erroneous decisions in terms of incorrect rejections also increases. Usually, a plot of these changes will show an invert U function with the asymptote of this curve being the point of maximum correct decisions.

This plot shows that at a base rate of 80 percent, the percentage of correct decisions begins to drop sharply after an aptitude score of 90. This implies that the qualification standard should not exceed 90 if the desired effect is to maximize the number of correct selection decisions.



Tradeoff Between Incorrect Decisions

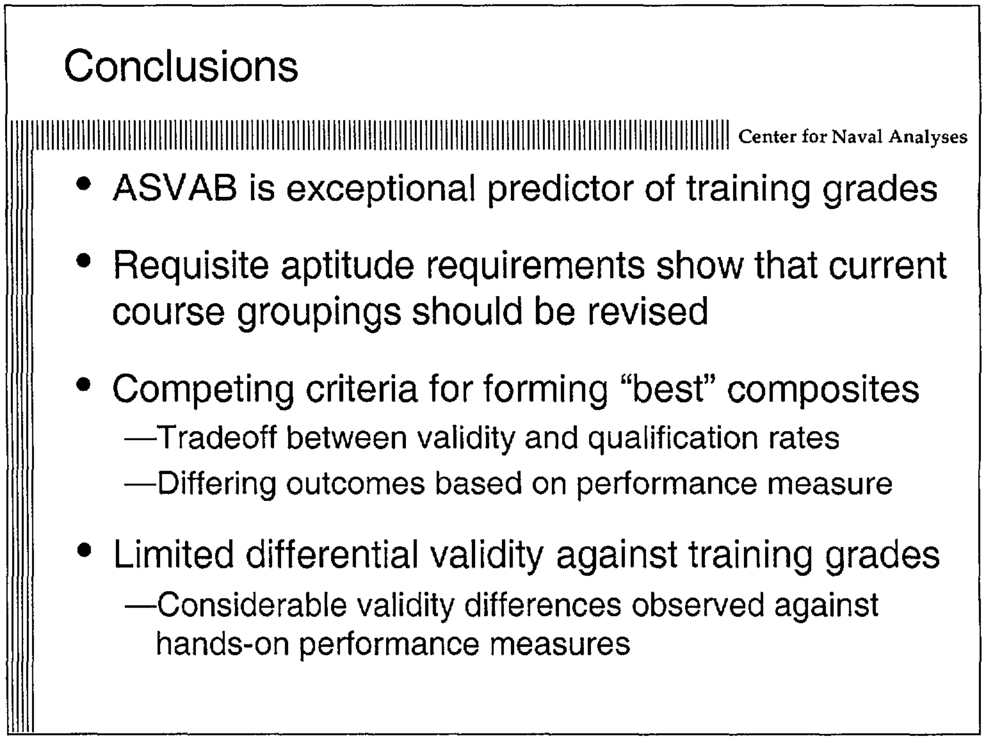
Infantry training course

Center for Naval Analyses

As aptitude increases, the number of accepted persons who are unsatisfactory performers decreases. Conversely, the aptitude standards are raised, the number of persons who would have been successful performers, but were not accepted for service, increases. The policy issue is how to reconcile these two different types of incorrect decisions.

One approach is that the incorrect decisions are equally valued (or avoided). The solution to this approach is shown in this graph as the intersection of the two corresponding lines for a given base rate. For the infantry, this represents an aptitude standard of slightly more than 80.

However, if these two types of incorrect decisions are not equally acceptable, the aptitude standards would need to be adjusted accordingly. The classic example involves pilots, where significant consequences are associated with having failures, in terms of training costs, accomplishment of military mission, and even human life. In these types of cases, policymakers would tend to want to minimize the number of incorrect acceptances at the expense of having considerably more individuals being incorrectly rejected. Such explicit valuations would result in higher aptitude standards.



Conclusions

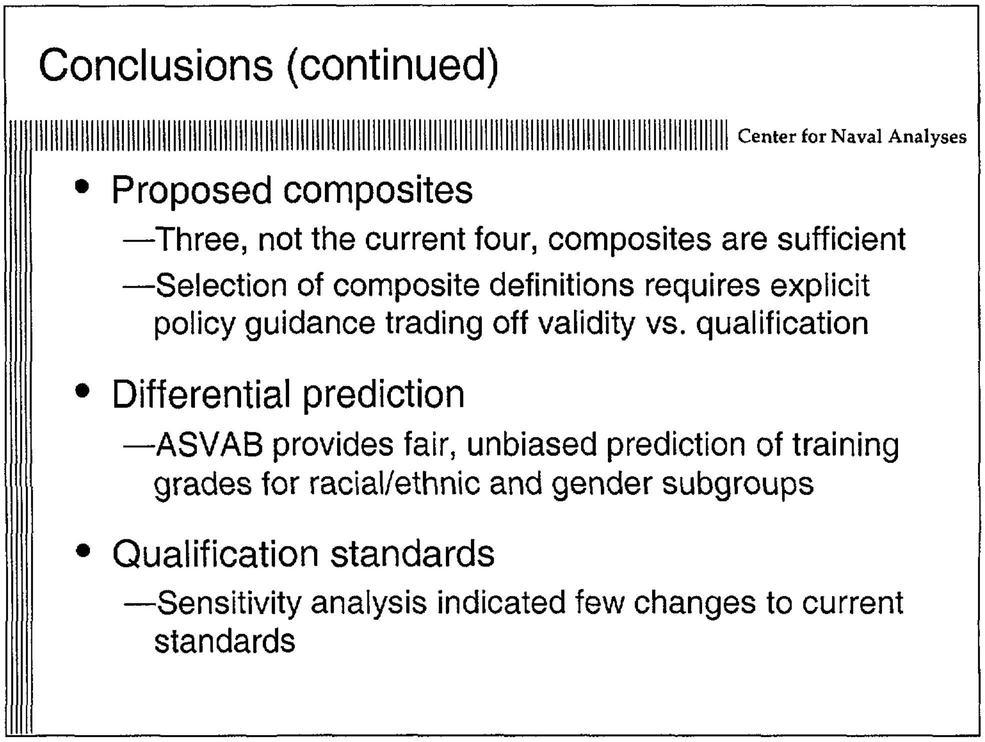
Center for Naval Analyses

- ASVAB is exceptional predictor of training grades
- Requisite aptitude requirements show that current course groupings should be revised
- Competing criteria for forming "best" composites
  - —Tradeoff between validity and qualification rates
  - —Differing outcomes based on performance measure
- Limited differential validity against training grades
  - —Considerable validity differences observed against hands-on performance measures

We found the ASVAB to be an excellent predictor of training grades for Marine Corps entry-level courses. Range-corrected validities averaged in the low- to mid-70s across all courses for their respective aptitude composites. The validity findings were consistently high with the exception of most infantry instruction. These courses generally suffered from limited distributions of training grades reflected in strong ceiling effects. No data transformations could be made to correct these distributional problems; rather, we substituted hands-on performance test scores collected as part of the Marine Corps Job Performance Measurement (JPM) Project. Given the exceptionally high magnitude and consistency of these validity results, there is little room for any new aptitude measures to improve the prediction of training grades.

Courses are usually clustered based on varying criteria so that a single aptitude composite can be used to determine eligibility for all courses within that cluster. We grouped the entry-level courses based on their requisite aptitude requirements (i.e., math, verbal, technical, and speed factors). Our results showed that several courses should be grouped differently and use another composite for determining course qualification. We also found that the math factor was required for successful performance in every training course and, therefore, should be included in the formation of all composites.

Historically, subtest validity results have been the primary driver in determining the subtest definitions of selection composites. Subtest validity should be one of several considerations. We also examined differences in validities by subgroups, adverse impact, and reliability. Although there was variation in subtest validities by important subgroups (white, black, and Hispanic; male and female), such differences were generally not significant. However, no ASVAB subtest satisfies the definition for acceptable levels of adverse impact for racial/ethnic subgroups—even at the lowest levels of selection. Male-female differences in qualification rates are also significant on technical subtests at mid- to upper-levels of selection. Subtest reliabilities tend to be much lower on the technical subtests for blacks and women.



Conclusions (continued)

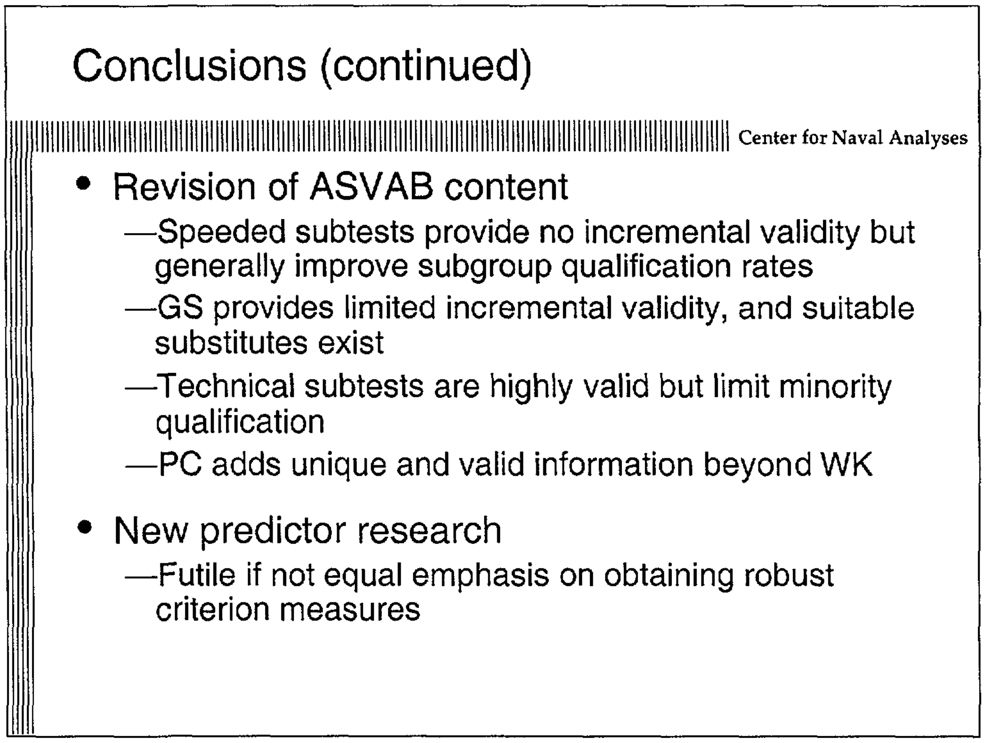
Center for Naval Analyses

- Proposed composites
  - —Three, not the current four, composites are sufficient
  - —Selection of composite definitions requires explicit policy guidance trading off validity vs. qualification
- Differential prediction
  - —ASVAB provides fair, unbiased prediction of training grades for racial/ethnic and gender subgroups
- Qualification standards
  - —Sensitivity analysis indicated few changes to current standards

Considering the competing criteria discussed earlier, we proposed composite definitions that differ somewhat from the current Marine Corps definitions. Three, not the current four, aptitude composites are sufficient to properly assign recruits to the appropriate military specialty. The proposed alternatives result in validity gains of about .02 point over the current composites. The number of blacks and women qualifying under the proposed composites would also increase for both mechanical maintenance and general technical courses. However, qualification rates for women would be lower for the proposed composite in clerical courses. The new composite structure would also require some realignment of courses with composites. For example, most clerical courses would now use the new general technical composite; infantry courses would now use the new mechanical maintenance composite.

We examined the differential prediction of each of the proposed composites for those courses that had sufficient minority sample sizes (we established a minimum of 50 students). There were 25 courses that satisfied this criterion for the racial/ethnic comparisons and 20 courses for gender comparisons. For more than two-thirds of the courses, we found equal predictions of training grades for whites, blacks, and Hispanics. For the remaining courses, we noted over prediction of training performance, but no subgroup consistently was over predicted. In general, the intercept differences represented less than a third of a standard deviation in training grades (this translated into less than 5 aptitude points). We noted similar results for the predictions by gender subgroups. Again, 70 percent of the courses had equal prediction, and no subgroup was consistently over predicted. We concluded that the ASVAB provides fair and equitable prediction of training grades for minority subgroups.

Based on a model that considers the relationship between aptitude and training grades, various definitions of successful performance, and policy guidance on an acceptable training failure rate, we determined the current qualification standards to be appropriate. We did identify four courses that potentially need to have their aptitude cutoffs adjusted. We also conducted sensitivity analyses on relevant variables to assess their impact on the final solution for each course.



Conclusions (continued)

Center for Naval Analyses

- Revision of ASVAB content
  - —Speeded subtests provide no incremental validity but generally improve subgroup qualification rates
  - —GS provides limited incremental validity, and suitable substitutes exist
  - —Technical subtests are highly valid but limit minority qualification
  - —PC adds unique and valid information beyond WK
- New predictor research
  - —Futile if not equal emphasis on obtaining robust criterion measures

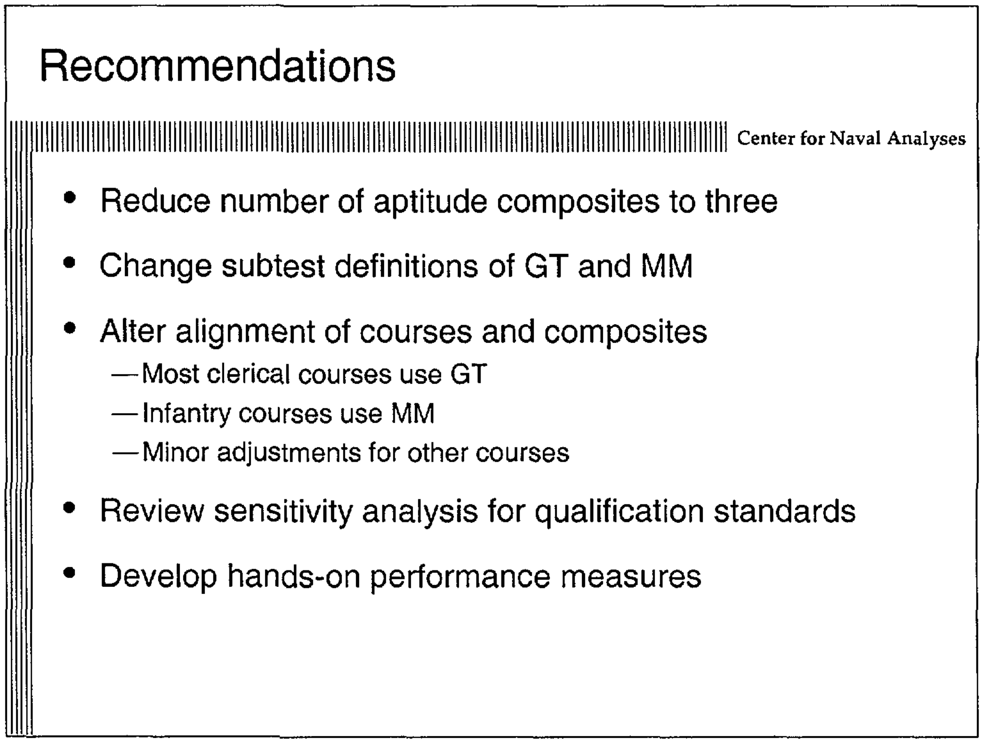
This research also has implications for the potential revision of ASVAB content. Contrasting composites with and without the speeded subtests showed that these two measures added little, if any, incremental validity and that, if necessary, suitable subtests can be substituted to recover any lost validity resulting from their deletion. However, a drawback to their exclusion from the ASVAB would be that fewer numbers of blacks and women would qualify for enlistment. The composites we are proposing in this study for the Marine Corps to implement do not include either the Coding Speed or Numerical Operations subtests.

Although the General Science (GS) subtest is included in the electronics composite that we propose, other verbal tests could sufficiently replace it. One justification for our use of GS was an attempt to obtain some diversity among the subtests across all composites.

The technical subtests were among the most valid aptitude measures for the mechanical and electronics courses—especially if the criterion was hands-on performance, but also to a reasonable degree against training grades. These results indicate that the technical subtests not only have absolute validity, but also incremental validity over the other ASVAB subtests. In other words, there are no alternative subtests that can be used in the place of these measures without significant loss of validity. Despite these high validity levels, the technical subtests are more restrictive for minority qualification than the other ASVAB subtests.

Paragraph Comprehension (PC) and Word Knowledge (WK) are used in combination as the verbal composite. Questions have been raised about both the reliability and measurement efficiency of PC (i.e., the amount of information gained relative to testing time required). Our analysis indicates that PC provides unique and valid information beyond that of WK and that the subtest should be retained.

Finally, research efforts are flawed if they focus only on identifying valid new predictor tests for inclusion in the ASVAB and do not place an equal emphasis on obtaining robust criterion measures. Validation research for new aptitude measures that is limited to training grades will tend to restrict both the type and scope of new tests that would be considered for expanding the constructs measured by the selection battery.

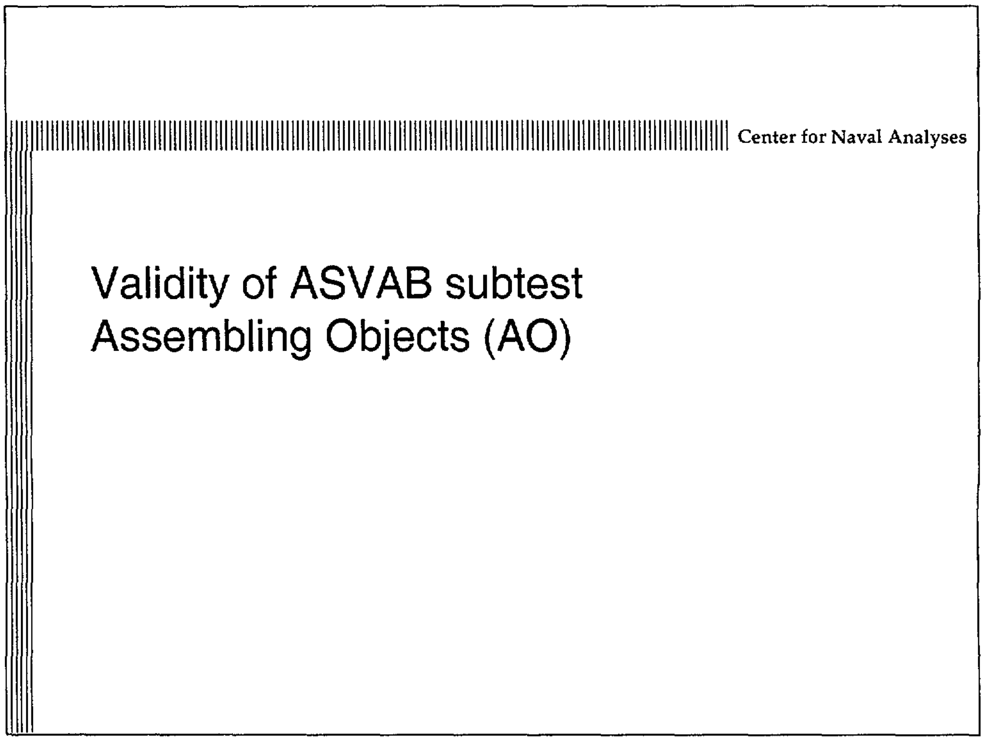


Recommendations

Center for Naval Analyses

- Reduce number of aptitude composites to three
- Change subtest definitions of GT and MM
- Alter alignment of courses and composites
  - —Most clerical courses use GT
  - —Infantry courses use MM
  - —Minor adjustments for other courses
- Review sensitivity analysis for qualification standards
- Develop hands-on performance measures

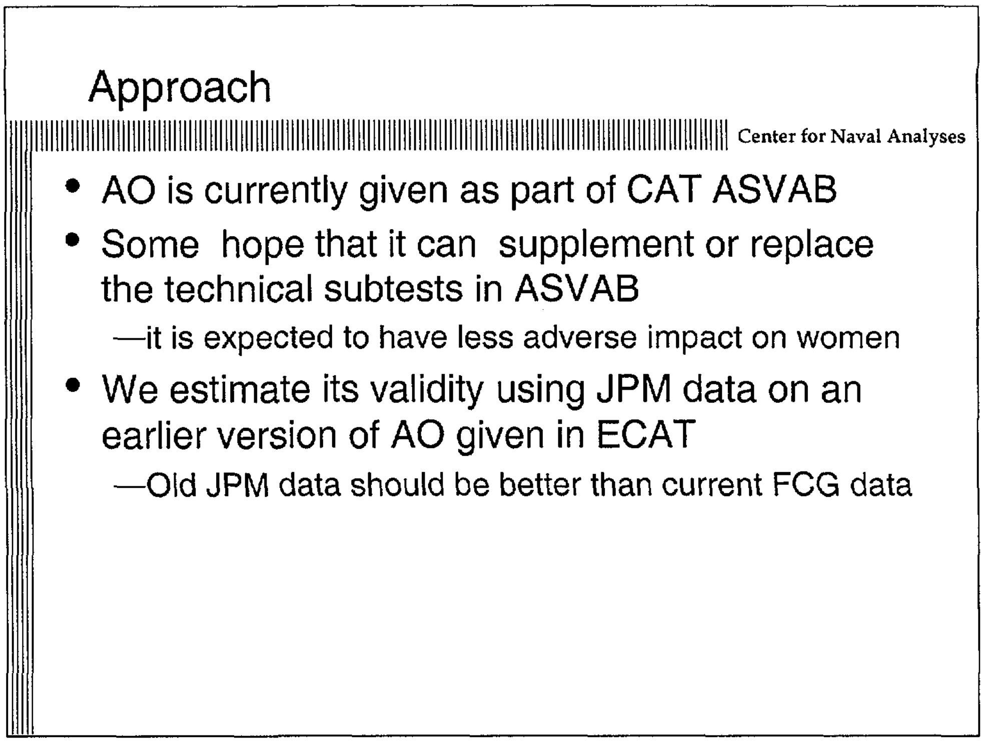
Based on our analysis of the prediction of training performance for entry-level courses, we make the above recommendations to the Marine Corps.



Center for Naval Analyses

Validity of ASVAB subtest Assembling Objects (AO)

Recently an experimental subtest known as Assembling Objects (AO) has been added to the battery. This section of the briefing examines evidence for its validity.



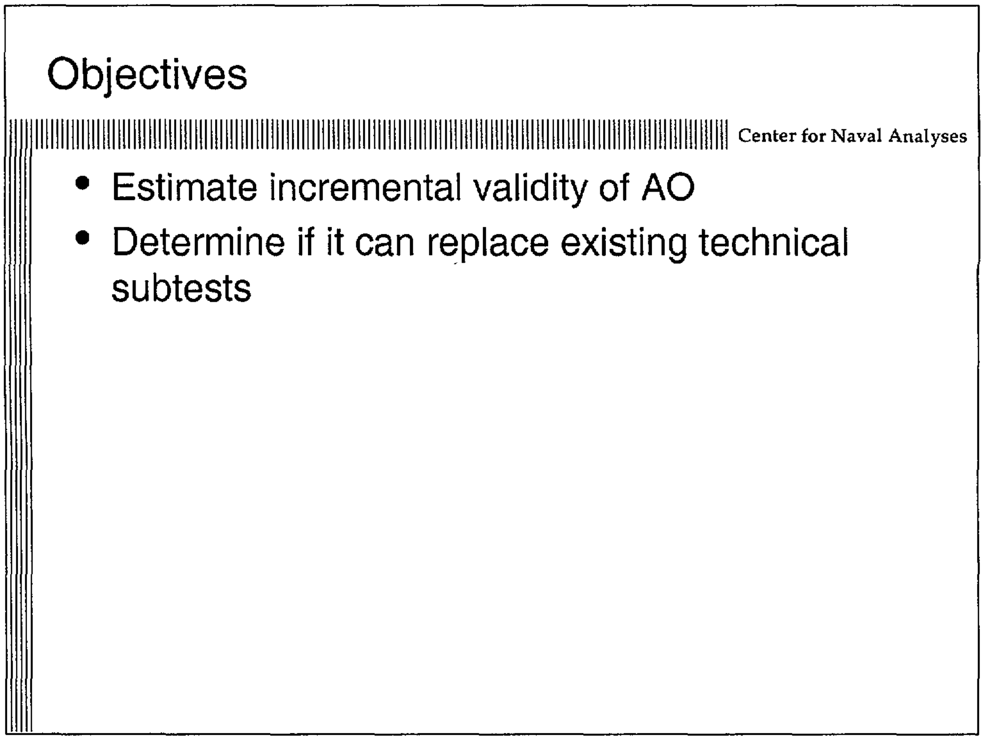
Approach

Center for Naval Analyses

- AO is currently given as part of CAT ASVAB
- Some hope that it can supplement or replace the technical subtests in ASVAB
  - —it is expected to have less adverse impact on women
- We estimate its validity using JPM data on an earlier version of AO given in ECAT
  - —Old JPM data should be better than current FCG data

AO was recently added to the battery as an experimental subtest. The hope was that it might be able to replace some or all of the ASVAB technical subtests (Auto Shop, Mechanical Comprehension, and Electrical Information) that have been criticized as having large adverse impact on the selection of females.

We will estimate its validity using hands-on job performance data collected in the late 1980s and early 1990s. This data (although collected on an earlier, but similar, version of AO) is believed to be far superior to any criterion data now available.

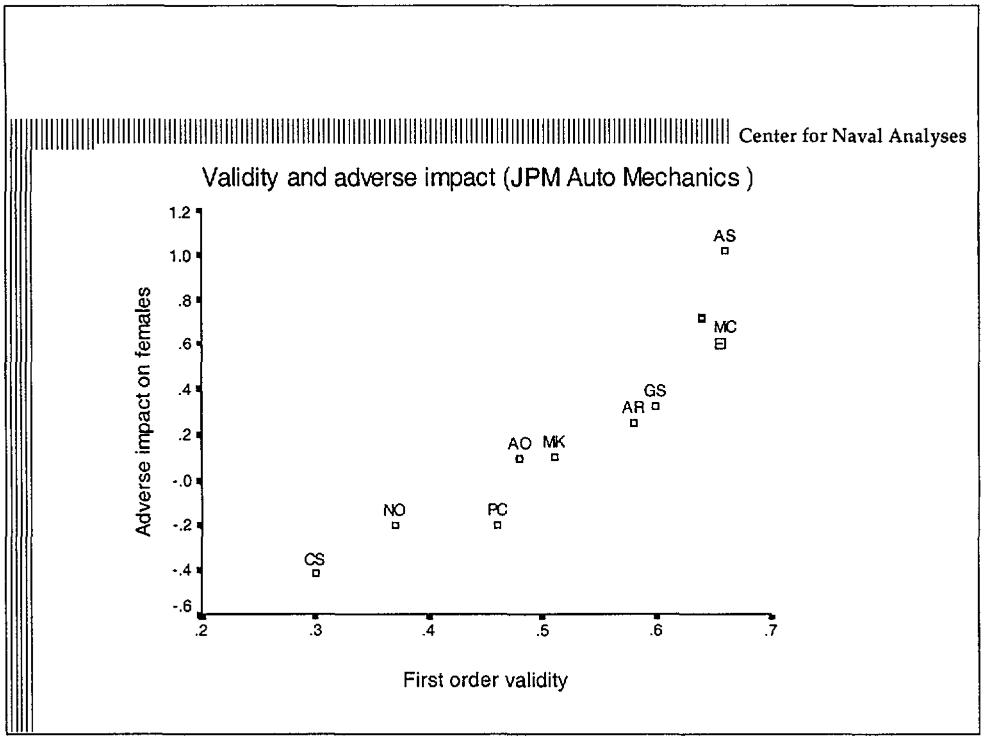


Objectives

Center for Naval Analyses

- Estimate incremental validity of AO
- Determine if it can replace existing technical subtests

The objective is to estimate the incremental validity of AO and to determine if it can replace existing technical subtests.

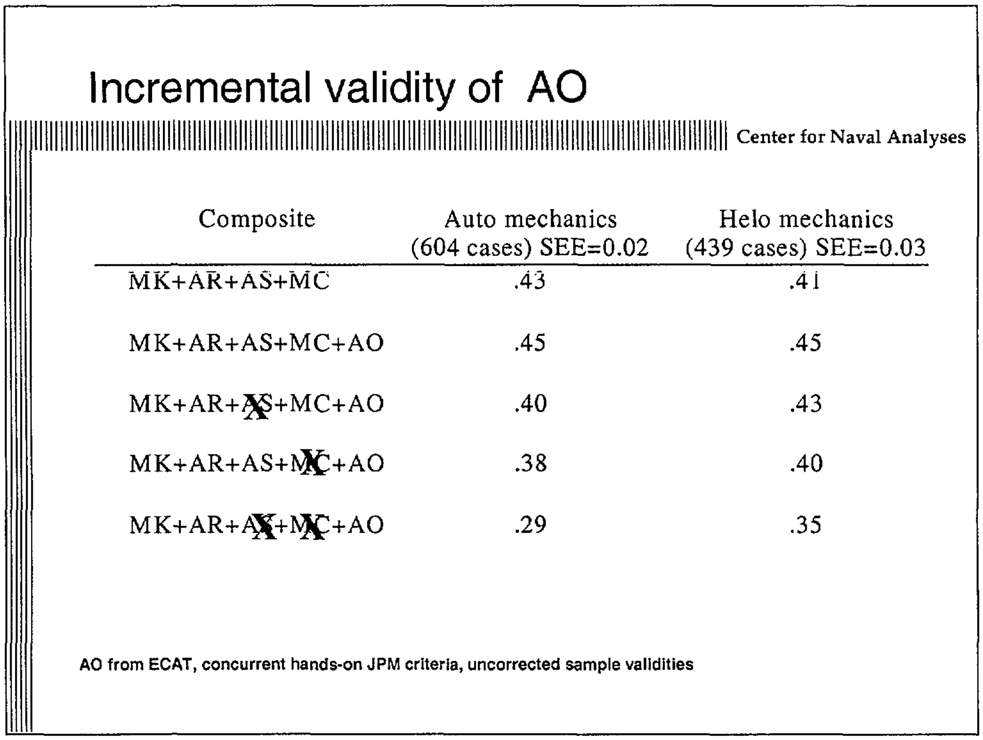


Center for Naval Analyses

Validity and adverse impact (JPM Auto Mechanics)

There are important tradeoffs to be considered in trying to reduce adverse impact on females in mechanical jobs. This slide shows adverse impact on females versus first order validity of males for each subtest. (Due to poor planning we did not over sample females and hence did not have enough cases to plot validity for females.)

The data suggest that subtests with high validity (for males and possibly for females) also have high adverse impact on females, and visa versa. In order to use subtests with lower adverse impact on females it appears necessary to give up validity (for males and possibly females). Note that the AO subtest is located in the middle of the scatter plot. It has very low adverse impact on females but also low validity for mechanical jobs.



Incremental validity of AO

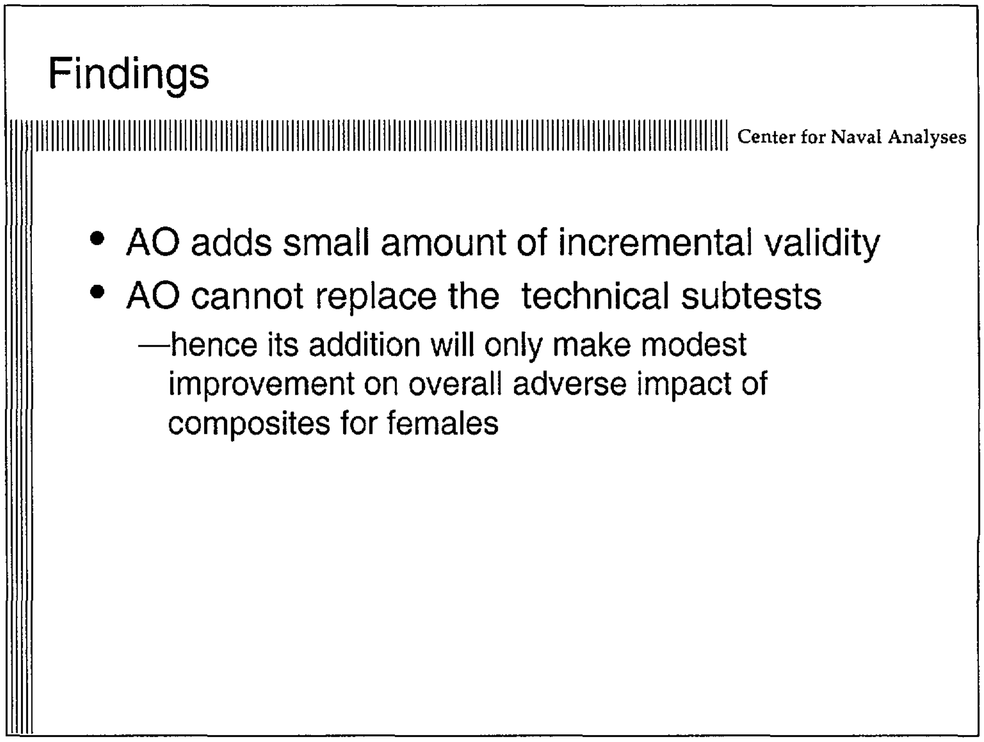
Center for Naval Analyses

Composite	Auto mechanics (604 cases) SEE=0.02	Helo mechanics (439 cases) SEE=0.03
MK+AR+AS+MC	.43	.41
MK+AR+AS+MC+AO	.45	.45
MK+AR+~~AS~~+MC+AO	.40	.43
MK+AR+AS+~~MC~~+AO	.38	.40
MK+AR+~~AS~~+~~MC~~+AO	.29	.35

AO from ECAT, concurrent hands-on JPM criteria, uncorrected sample validities

In this slide we show the incremental validity of AO against a hands-on job performance criteria for auto mechanics and helicopter mechanics. The first line shows the validity for the composite that we recommend for mechanical courses. The second line shows that adding AO increases the validity by .02 to .04. Clearly AO does have some incremental validity.

In the third, forth, and fifth line we remove various combinations of the technical subtests. In each case there is a cost in validity. It is a policy decision as to whether the cost in validity is worth the gain in reducing adverse impact. We have not estimated the reduction in adverse impact for females from removing one of the current subtests and replacing it by AO. However, we would expect that the benefit would be muted because it is clear that either AS or MC must remain in the composite.

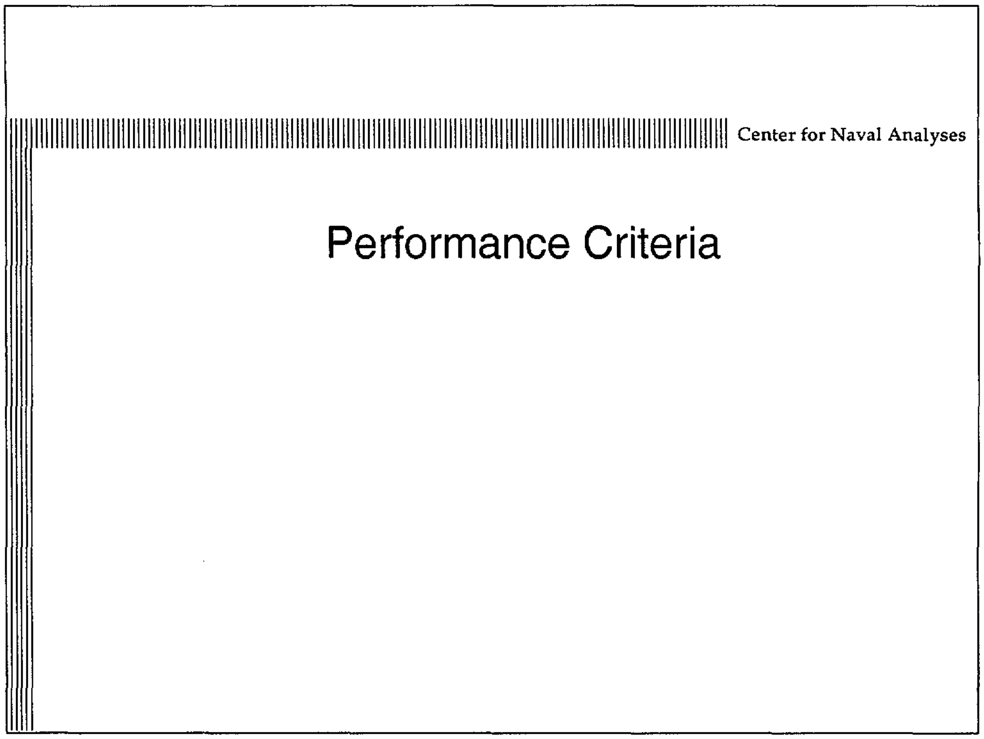


Findings

Center for Naval Analyses

- AO adds small amount of incremental validity
- AO cannot replace the technical subtests
  - —hence its addition will only make modest improvement on overall adverse impact of composites for females

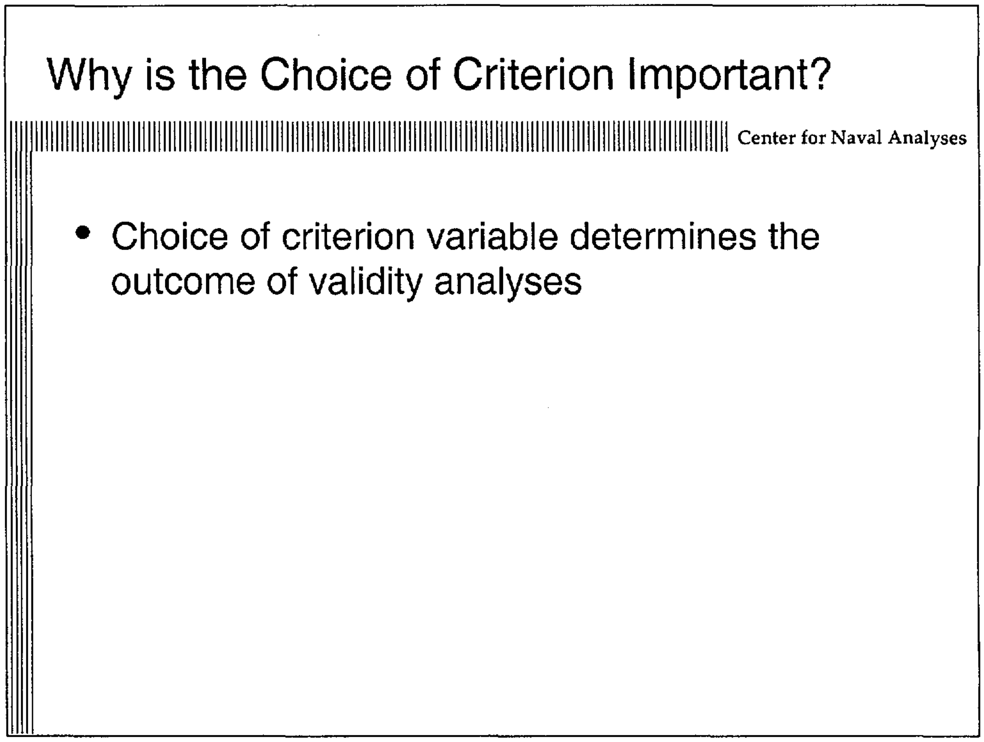
We conclude that AO adds a small amount of incremental validity. We also conclude that it can not totally replace the technical subtests. Its addition to a composite is expected to only make modest improvements in overall adverse impact on females.



Center for Naval Analyses

Performance Criteria

In this section we examine the issue of performance criteria

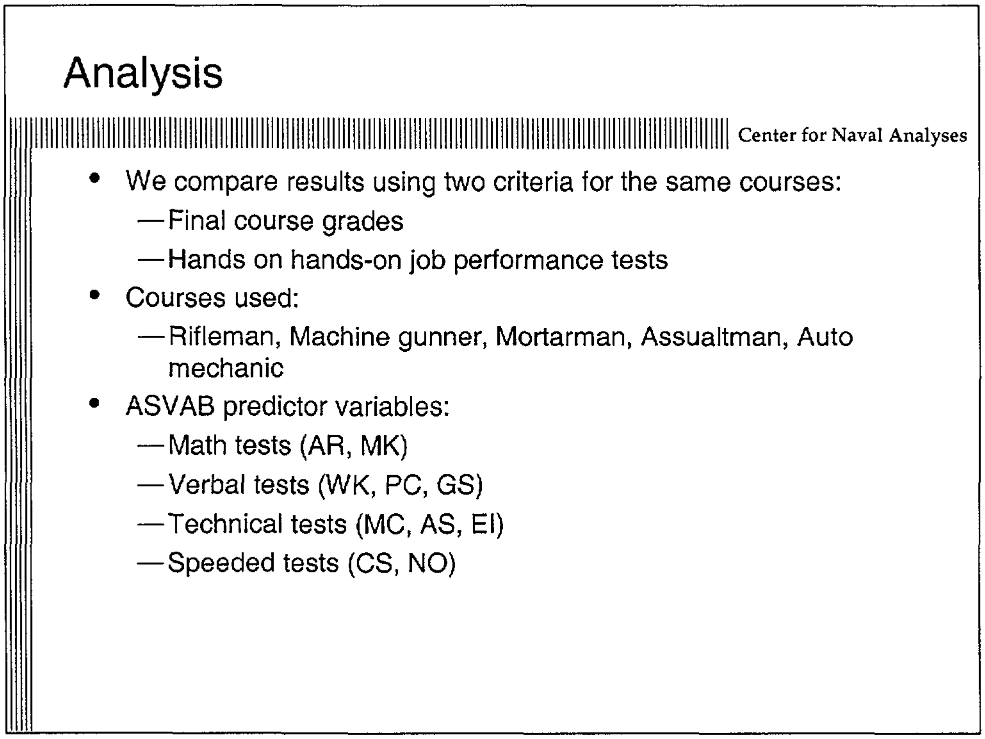


Why is the Choice of Criterion Important?

Center for Naval Analyses

- Choice of criterion variable determines the outcome of validity analyses

The choice of criterion variable is important because this choice determines the outcome of the validity analysis. One choice will lead to "AFQT like" composites, the other will require one or more technical subtests.



Analysis

Center for Naval Analyses

- We compare results using two criteria for the same courses:
  - —Final course grades
  - —Hands on hands-on job performance tests
- Courses used:
  - —Rifleman, Machine gunner, Mortarman, Assualtman, Auto mechanic
- ASVAB predictor variables:
  - —Math tests (AR, MK)
  - —Verbal tests (WK, PC, GS)
  - —Technical tests (MC, AS, EI)
  - —Speeded tests (CS, NO)

We will calculate validities for using the final courses grade (FCG) and hands-on job performance (JPM) criterion measures for the same courses. The courses examined include rifleman, machine gunner, mortarman, assualtman, and auto mechanic.

As ASVAB predictor variables we will use sums of subtests that exemplify the four factors always found in factor analysis of the battery, i.e.,

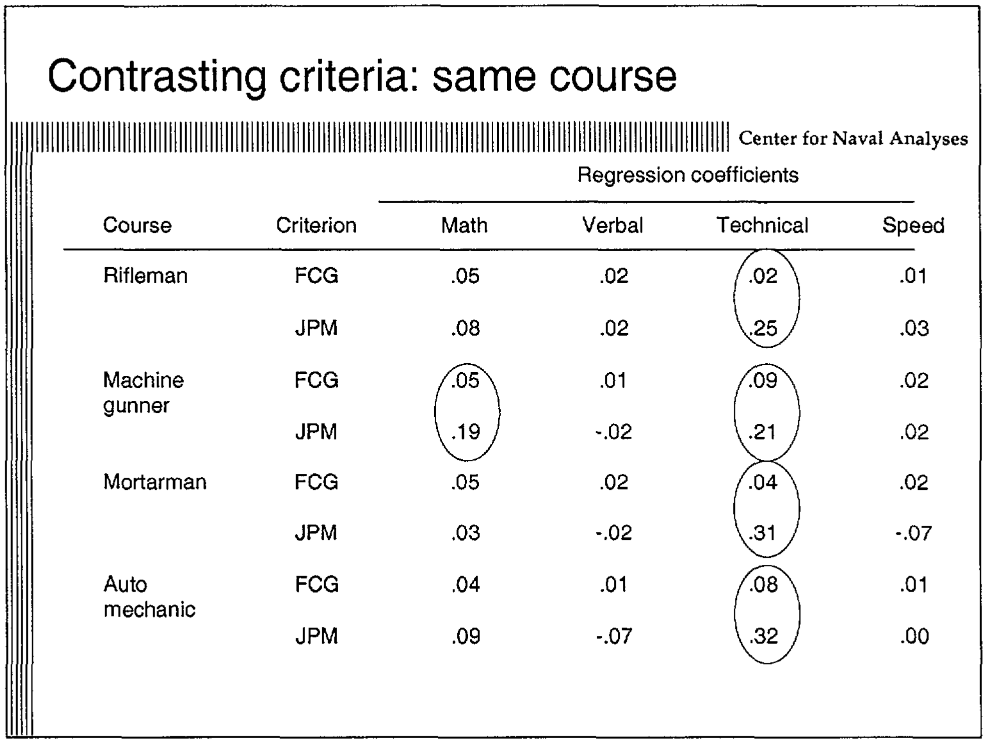
•Math subtests (AR+MK)

•Verbal subtests (WK+PC+GS)

•Technical subtests (MK+AR)

•Speeded subtests (CS+NO)

Note the these sums do not represent "pure" factors, they are merely heavily loaded with the pure factor.



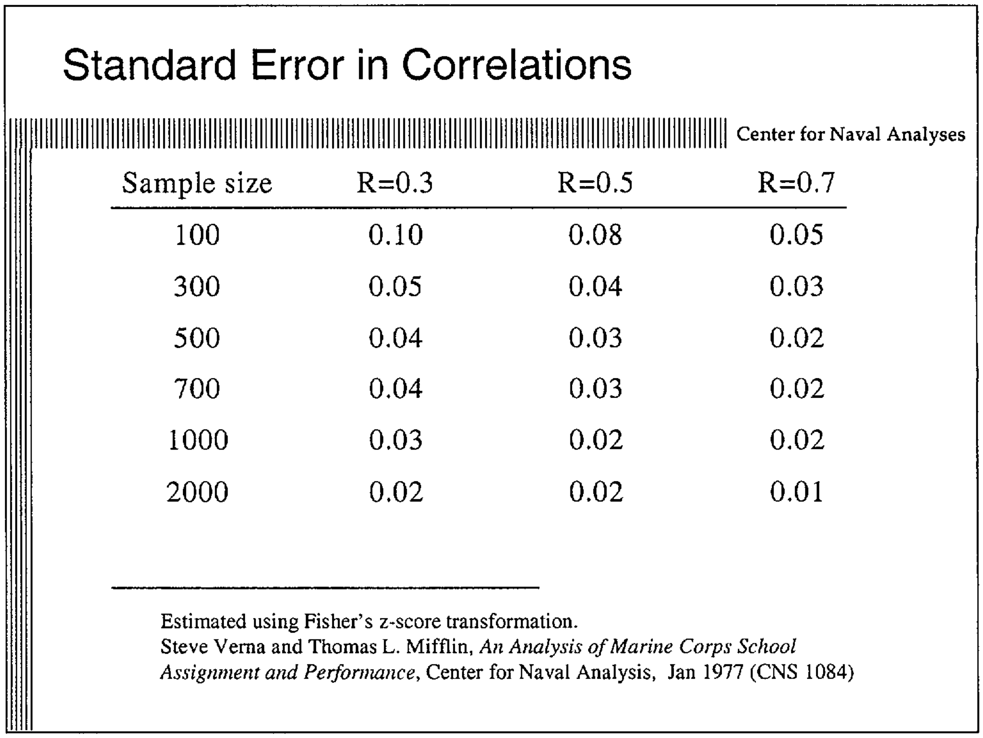
Contrasting criteria: same course

Center for Naval Analyses

		Regression coefficients			
Course	Criterion	Math	Verbal	Technical	Speed
Rifleman	FCG	.05	.02	.02	.01
	JPM	.08	.02	.25	.03
Machine gunner	FCG	.05	.01	.09	.02
	JPM	.19	-.02	.21	.02
Mortarman	FCG	.05	.02	.04	.02
	JPM	.03	-.02	.31	-.07
Auto mechanic	FCG	.04	.01	.08	.01
	JPM	.09	-.07	.32	.00

This slide contrasts the coefficients from a regression of the form:
FCG (or JPM) = A(Math) + B (Verbal) + C (Technical) + D(Speed)

The two performance criteria are seen to lead to totally different requirements for the technical subtests. Using a FCG criteria the technical subtests are not very important. Using a JPM criterion they are critically important. Clearly the choice of criterion drives the results.



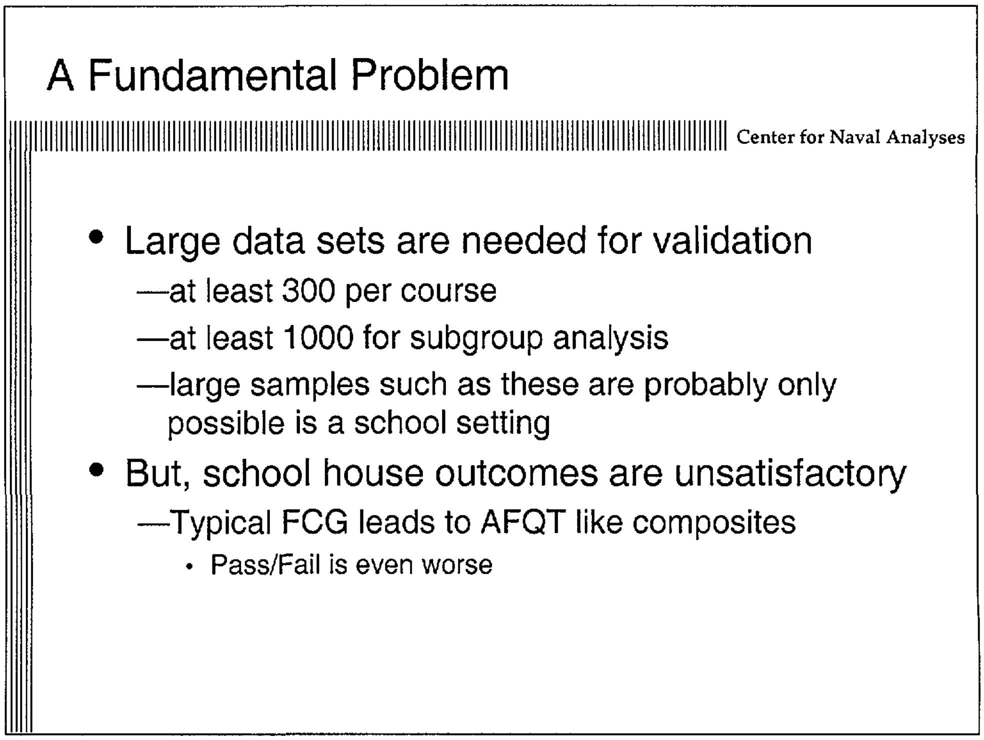
Standard Error in Correlations

Center for Naval Analyses

Sample size	R=0.3	R=0.5	R=0.7
100	0.10	0.08	0.05
300	0.05	0.04	0.03
500	0.04	0.03	0.02
700	0.04	0.03	0.02
1000	0.03	0.02	0.02
2000	0.02	0.02	0.01

Estimated using Fisher's z-score transformation.
Steve Verna and Thomas L. Mifflin, *An Analysis of Marine Corps School Assignment and Performance*, Center for Naval Analysis, Jan 1977 (CNS 1084)

It is well to recognize the limitations of sample size on the standard error of correlations. Estimates of standard error for various sizes of sample and correlation coefficients are given in this slide.

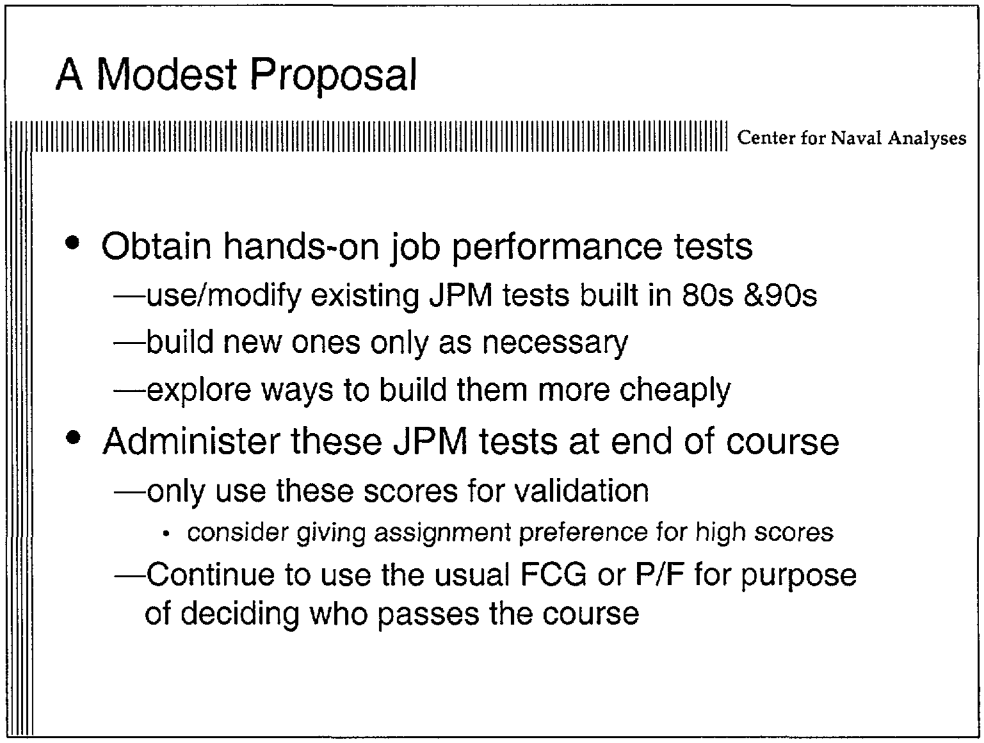


A Fundamental Problem

Center for Naval Analyses

- Large data sets are needed for validation
  - —at least 300 per course
  - —at least 1000 for subgroup analysis
  - —large samples such as these are probably only possible is a school setting
- But, school house outcomes are unsatisfactory
  - —Typical FCG leads to AFQT like composites
    - Pass/Fail is even worse

As we see it we have a fundamental problem in validation. We need large data sets which are likely only available in a school setting. However typical schoolhouse criteria (FCG) are unsatisfactory and pass/fail is even worse.



A Modest Proposal

Center for Naval Analyses

- Obtain hands-on job performance tests
  - —use/modify existing JPM tests built in 80s &90s
  - —build new ones only as necessary
  - —explore ways to build them more cheaply
- Administer these JPM tests at end of course
  - —only use these scores for validation
    - consider giving assignment preference for high scores
  - —Continue to use the usual FCG or P/F for purpose of deciding who passes the course

This problem leads us to a modest proposal.

Hands-on JPM or something very similar must be used to validate ASVAB. Otherwise we might as well use AFQT and be done with it. We recognize that JPM measures are very expensive to develop.

We propose that the ASVAB community obtain hands-on JPM criteria or something very similar for use in validation. This could be done by using or modifying the JPM measures built in the 1980s and 1990s. New JPM tests should only be built if absolutely necessary. We should explore new ways to build them more cheaply.

These JPM tests should be routinely administered at the end of each course and the scores used for validation. Students might be motivated to try hard by some preferential assignment for high scores. Students could continue to be passed on the basis of the current FCG or Pass/Fail criterion.

CNA