NPRST

Navy Personnel Research, Studies, and Technology
5720 Integrity Drive • Millington, Tennessee 38055-1000 • www.nprst.navy.mil

research at work

NPRST-TR-15-1 October 2014

Introductory Guide for Conducting ASVAB Validation/Standards Studies in the U.S. Navy

Janet D. Held, M.S.
Navy Personnel Research, Studies, and Technology
Sarah A. Hezlett, Ph.D. and Jeff W. Johnson, Ph.D.
Personnel Decisions Research Institute, Inc.
Rodney A. McCloy, Ph.D.
Human Resources Research Organization
Fritz Drasgow, Ph.D.
University of Illinois at Urbana-Champaign
Eduardo Salas, Ph.D.
University of Central Florida

NPRST-TR-15-1
October 2014

Introductory Guide for Conducting ASVAB Validation/Standards Studies in the U.S. Navy

Janet D. Held, M.S.
Navy Personnel Research, Studies, and Technology
Sarah A. Hezlett, Ph.D.
Jeff W. Johnson, Ph.D.
Personnel Decisions Research Institute, Inc.
Rodney A. McCloy, Ph.D.
Human Resources Research Organization
Fritz Drasgow, Ph.D.
University of Illinois Urbana-Champaign
Eduardo Salas, Ph.D.
University of Central Florida

Edited by
Janet D. Held, M.S.
Navy Personnel Research, Studies, & Technology
Thomas R. Carretta, Ph.D.
Air Force Research Laboratory, Wright-Patterson AFB, OH
Jeff W. Johnson, Ph.D.
Personnel Decisions Research Institute, Inc.
Rodney A. McCloy, Ph.D.
Human Resources Research Organization

Reviewed by
Tanja F. Blackstone, Ph.D.

Approved and released by
David M. Cashbaugh
Director

Navy Personnel Research, Studies, and Technology (NPRST)
Bureau of Naval Personnel (BUPERS-1)
5720 Integrity Drive
Millington, TN 38055-1300
www.nprst.navy.mil

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From-To)
16-10-2014	Technical Report	October 2012-Sep 2014

4. TITLE AND SUBTITLE

Introductory Guide for Conducting ASVAB Validation/Standards Studies in the U.S. Navy

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROJECT ELEMENT NUMBER

6. AUTHORS

Janet D. Held, Sarah A. Hezlett, Jeff W. Johnson, Rodney A. McCloy, Fritz Drasgow, Eduardo Salas

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATIONS NAME(S) AND ADDRESS(ES)

Navy Personnel Research, Studies, & Technology (NPRST/BUPERS-1)
Bureau of Naval Personnel
5720 Integrity Drive
Millington, TN 38055-1000

8. PERFORMING ORGANIZATION REPORT NUMBER

NPRST-TR-15-1

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Armed Services Vocational Aptitude Battery (ASVAB) is a joint-service battery used by the U.S. military services for enlistment selection and occupational classification decisions. All of the military services validate the ASVAB from time to time to ensure that the composites of ASVAB tests they use to classify their enlisted members to occupations are the most predictive of military performance, and that cutscores are set to manage academically related training attrition. The Navy, however, is the only military service to support an operationally focused ASVAB Validation/Standards program. This Introductory Manual provides the background information and context that both policy makers and researchers should know about the program's objectives and processes. The companion manual, "Technical Guidance for Conducting ASVAB Validation/Standards Studies in the U.S. Navy" is intended for those who actually conduct the work.

15. SUBJECT TERMS

ASVAB, selection and classification, Navy ASVAB validation/standards

16. SECURTIY CLASSIFICATION OF:			17. LIMITATATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
UNCLASSIFIED			UNLIMITED	169	Wendy Douglas
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED			(901)874-2218

Foreword

This report is the first of two that provides guidance on how to conduct predictive validation studies and the setting of aptitude/ability standards for enlisted military occupations using the Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB is the primary enlistment qualification and occupational classification instrument used by all of the U.S. military services. The Navy was the lead on the project because it is the only Service at this time maintaining a continuing ASVAB Validation/ Standards program. This report is considered the Introductory Manual; the second report, the Technical Manual.

This work was sponsored and funded by the Navy's Selection and Classification Office (OPNAV132G) with a contribution of funding from the Defense Manpower Data Center, Personnel Testing Division (DMDC-PTD). The work was executed by Navy Personnel Research, Studies, and Technology (NPRST/BUPERS-1), a department of the Bureau of Naval Personnel, along with a team of experts on the various manual topics. The contract work was conducted under the auspices of the U.S. Army Research Office Scientific Services Program administered by Battelle (Delivery Order 0253, Contract No. W911NF-07-D-0001).

David M. Cashbaugh
Director

Executive Summary

The Armed Services Vocational Aptitude Battery (ASVAB) is the primary tool used by all of the U.S. military services to screen, on a cognitive ability basis, for enlistment eligibility (selection) and for occupational classification. The Navy has about 85 occupational fields that are called *Ratings*. Training differs for each Rating not only in content, but in technical complexity and time allowed for completion. Because Sailors who fail training can be set back in their careers, and because failures add to the Navy's training costs, establishing and maintaining effective ASVAB standards is an essential military operational function.

Factors considered in determining which Navy Ratings require an ASVAB validation /standards study include (a) observed increases in academic failure or setback rates, (b) major curriculum or training platform changes, (c) constrained training time occurring from funding shortfalls, (d) Rating mergers or newly formed Ratings, and (e) a downturn in the military recruiting environment that can depress applicant ASVAB scores. These general criteria for establishing Navy ASVAB validation/standards study requirements could also be used by other military services and replace any existing time-table or periodicity requirement.

Currently there is both a requirement and an opportunity for the military to improve occupational classification by optimizing the ASVAB and its use. The requirement comes from a potential recruiting downturn (associated with an improved U.S. economy) that will necessitate particular attention to the Services' ASVAB standards so as to minimize extra training costs associated with lower recruit population ASVAB scores. The opportunity is that new approved candidate ASVAB tests and adjunct classification tests will indirectly benefit recruiting by improving the person-job fit. Test validation research tells us that cognitive constructs not yet a part of the ASVAB improve the person-job fit but also increase the proportion of annual recruit populations occupation qualified. An improved person-job fit by itself could result in higher enlistee retention rates (via improved job satisfaction), thereby potentially mitigating both recruiting and retention issues (e.g., Sailors leaving the Navy after only one term of enlistment increase both technical training costs projected recruiting requirements).

The above reasons and others combine to make a strong case for conducting ASVAB validation/standards studies on a continual basis that may or may not incorporate new tests. Therefore, it is recommended that each Service maintain a continuing ASVAB Validation/Standards program, preferably integrated as part of an overarching Selection and Classification (S&C) function. The Navy has an S&C function (OPNAV132G) that sponsors ASVAB validation/standards studies and, relatedly, the development of this Introductory Manual for conducting the studies, and the companion Technical Manual.

Chapter 1 establishes the purpose of the Introductory Manual, and briefly, the Technical Manual. The Introductory Manual is intended to provide general guidance on conducting ASVAB validation/standards studies and will be most useful for sponsors, policy makers, and leadership from the recruiting, training, and the enlisted communities. All of these entities benefit from effective ASVAB standards and so are considered stakeholders. The Technical Manual is intended to provide the researcher who conducts ASVAB validation/standards studies with the "how to do it" information.

Chapter 2 provides a brief history of the ASVAB and a description of the aptitudes, abilities, and knowledge constructs that it measures. Also described is a large-scale, joint-service test development and validation effort that was intended to augment the ASVAB with tests that better represent the ability to reason abstractly and solve problems than the ASVAB, which largely measures the ability to apply accumulated knowledge, skills, and abilities. Included also is a list of criteria for evaluating candidate tests developed in the past noting that recent revisions have been made and approved by the Manpower Accession Policy Working Group (MAPWG) that oversees the ASVAB.

Chapter 3 provides a discussion of a framework for understanding predictors and criteria both as overt measures of what we intend to measure, and as underlying constructs. Chapter 4 extends the focus of the ASVAB criterion measure, training performance, to job performance, which the ASVAB does predict depending upon what aspects are measured. The chapter highlights issues involved in predicting job performance such as the sparse literature on the criterion compared to the predictor, the perennial "Criterion Problem", and the difficult decisions about whether to consider single or multiple criteria.

Chapter 5 focuses on training performance and describes the evolution of Navy training from instructor-led, group-paced classroom courses with hands-on laboratory/demonstrations to self-paced, computer-based training. The chapter provides guidelines for best practice performance measurement in training, in particular for simulation-based training. Also provided are lessons learned from the literature on successful and unsuccessful computer-based training, which has guided the Navy's evolution to a blended training solution.

Chapter 6 lays out the process and steps for conducting Navy ASVAB validation/ standards studies. Although the phases of a study generalize for each Rating, the reader should know that each study is tailored to a Rating's particular issues that either directly or indirectly relates to the effectiveness of that Rating's ASVAB standard. The overall goal of any ASVAB validation/standards study is to provide a standard that minimizes academically related failure and setback rates while addressing the Navy's need to fill all Ratings with aptitude/ability-qualified Sailors.

Chapter 7 gives an overview of synthetic validity, an indirect approach to establishing test validity when (a) sample sizes are too small to conduct a formal criterion-related validation study or (b) new jobs are formed and the need to establish an ASVAB standard precedes the availability of the performance criterion data to use in ASVAB validity analysis. The Navy applies a modified version of synthetic validity that involves comparisons of training difficulty, time allowed to train, similar Ratings' ASVAB standards, and referencing cutscores to the ASVAB normative population.

Appendices A, B, and C contain examples of Navy ASVAB validation/standards studies to illustrate that one size does not fit all. That is, the three reports (Navy SEALs, Nuclear Field, and Mineman Ratings) are provided to (a) demonstrate the dynamics and issues experienced by each Navy community, (b) show how the study methods are tailored to address these issues, and (c) address cutscore setting for a range of magnitudes in the ASVAB composites' validity coefficients. Appendix D contains the INTERSERVICE Aptitude/Ability Standards Panel Charter (for conducting joint-service ASVAB validation/standards studies when one Service trains other Service members).

Contents

Figures

Tables

Chapter 1.
Introduction

Purpose

The Armed Services Vocational Aptitude Battery (ASVAB) is the primary tool used by all of the U.S. military services to screen, on a cognitive ability basis, for enlistment eligibility (selection) and for occupational classification. The purpose of this "manual" is to provide guidance to stakeholders and individuals who play policy roles in supporting the process of establishing or revising ASVAB standards for a Service's enlisted occupations with the Technical Manual providing the details.

There are several Department of Defense components that have ASVAB responsibilities. The Office of the Under Secretary of Defense for Personnel and Readiness, Accession Policy Directorate, sets policy for the development and use of the ASVAB for determining military service eligibility. The Defense Manpower Data Center, Personnel Testing Division (DMDC-PTD) is the Executive Agent for ASVAB research, development and maintenance. Headquarters, United States Military Entrance Processing Command (HQ-USMEPCOM) is responsible for enlistment processing, which includes maintaining ASVAB testing sites and equipment. The Manpower Accession Policy Working Group (MAPWG), comprised of technical and policy representatives from the Services, HQ-USMEPCOM, and DMDC-PTD as Chairs of the technical committee and full working group, has the responsibility of overseeing the development, effectiveness, and security of the ASVAB, and the assessment of any new tests that meet the criteria for inclusion in the battery or as adjunct occupational classification tests. Finally, the Defense Advisory Committee on Military Personnel Testing (DACMPT), comprised of nationally recognized experts in the areas of test development and industrial/organizational psychology, provides independent, objective recommendations on ASVAB development and enlistment screening to the Secretary of Defense, through the Under Secretary of Defense for Personnel and Readiness.

Each Service is responsible for developing its own ASVAB occupational classification composites and cutoff scores (hereafter referred to as cutscores), which we refer to as ASVAB standards. Setting effective ASVAB standards means addressing, to the extent possible, the dual objectives of optimizing training performance (quality) while meeting the target fiscal year's recruiting goals (quantity). The two objectives of quality and quantity produce tensions and each has their costs. In many ways, designing an ASVAB validation/standards study is like investing money, or at least minimizing overall costs, and must take into account all of the relevant information and technical issues. Conducting an ASVAB validation study can seem to be a deceptively simple process, partly because researchers have access to relevant data such as ASVAB scores as well as attrition and training data. Standard statistical procedures available in a variety of software packages are straightforward to operate and will generate an abundance of seemingly relevant output. Conducting an effective ASVAB validation study, however, involves much more than number crunching. Such outputs as regression equations derived in a selected sample, a correlation coefficient, or a percentage of a criterion variable's variance accounted for by the predictors are not the end products.

According to the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999), the term "validity" refers to the actual evidence and supporting theory for the interpretations of test scores, whereas the term "validation" refers to the actual process that accumulates the evidence (p. 9). For clarity, we refer to conducting an "ASVAB validation/standards study" fairly often as the latter, but that entails the former. For the purposes of this project the terms "validation" and "setting cutscores" involve distinct, multi-phased processes that result in an ASVAB standard. We mention this distinction because a prior Department of Defense (DoD) large-scale ASVAB validation effort with job performance measures (the Job Performance Measurement [JPM] Project) as the criteria focused on, with some exceptions, ASVAB validity, not the cutscores that created "standards" (Green & Mavor, 1994). We reference the JPM literature in later chapters and recommend here Rostker (2006) for those interested in a broad view of the evolution of the all-volunteer force and the role of testing, and Campbell and Knapp (Eds.) (2001) for a comprehensive view of personnel selection and classification that includes the ASVAB.

In 2006, Human Resources Research Organization, Inc. (HumRRO) developed a framework, or roadmap, to address a DMDC-PTD goal of having a unified approach that could be followed by all of the Services for conducting ASVAB validation research (McCloy, Campbell, Knapp, Strickland, & DiFazio, 2006). The unified framework provides a context for thinking about the ASVAB and other test validation research (e.g., involving candidate ASVAB tests or adjunct occupational classification tests). The framework outlines diverse validation objectives, reviews different criteria that may be used in validation research, and provides an overview of factors that may influence the Services' capacity to interpret and apply the results of validation studies. This Introductory Manual and the accompanying Technical Manual are intended to complement the unified HumRRO framework by providing more specific guidance and with the operational objective of setting effective ASVAB standards.

Of the goals outlined by McCloy et al. (2006), the guidance outlined in the manuals is most congruent with the goal of validating the ASVAB to maximize training performance, not job performance. There are two rationales. First, historically, the ASVAB was developed to map to training constructs, with training being developed form job analysis inputs. Second, inadequate ASVAB standards can result in large up-front Navy costs when Sailors fail training. These costs are from (a) the requirement to reclassify Sailors to other Rating thus adding to the original training costs, (b) transporting the Sailor to another training site with sometimes a long awaiting training status, and (c) shortened Sailor productive status and shortages of them in their jobs (impacting the Fleet's readiness). Further, training failure leads not only to Navy costs, but Sailor personal costs such as decreased motivation that can affect job performance and unit cohesion, and career setbacks. It is *not* the case that we are *not* interested in job performance; however the ASVAB linkage with job performance is already well known. We only have to skim the vast literature about the military's JPM Project (cited in later chapters) to understand the linkages. That is, the ASVAB predicts learning and performance in the training context that, post training, predicts job knowledge learning, that further predicts some important but not all aspects of job performance) (see, e.g., Hunter, 1986 or Kuncel & Hezlett, 2010, for broader contexts).

Approaches to Assigning Individuals to Jobs

There are different approaches that an organization can take in the use of aptitude/ability test scores (along with other applicant information) to make hiring or job classification decisions. One approach is to hire directly to an organization's job. Another approach is to make the hiring decision first and then, in a second stage, the job assignment decision. The same or different personnel selection instruments, or a mix, could be used for both the hiring and job assignment processes. Mental, moral, physical, and education are the primary factors in screening for military enlistment. The ASVAB, as a primary measure of the mental category, is the focus of the two manuals. The Armed Forces Qualification Test (AFQT), a composite of two verbal and two math ASVAB tests, is used to establish service eligibility. The ASVAB is used again in different test combinations tailored for occupational classification. Along with the ASVAB standard (composite with cutscore), many occupations may have additional standards such as eyesight/color blindness (e.g., electricians), hearing (e.g., sonar technicians), security clearances (e.g., intelligence), and language aptitude (e.g., cryptologists who are language interpreters/deciphers).

The military strives to assign all of its enlisted personnel to occupations/jobs for which they are best equipped to perform. In principle, the assignment (classification) system optimizes an array of valued outcomes for all jobs in the organization, not just training performance. However, to optimize job classification, information must be known about the relative value of jobs as well as job complexity, job requirements, predictability of performance, and what constitutes the performance measure. Given sufficient information on these "variables", the classification system will yield gains over pure selection (McCloy et al., 2006). We refer the reader to the Army's extensive work in the area of classification efficiency/effectiveness by including predictors that differentially classify individuals to jobs based upon improved person-job fit, termed differential assignment theory (Johnson & Zeidner, 1991; Lightfoot, Ramsberger, & Greenston, 2000; Rumsey, Walker, & Harris, [Eds.], 1994). The Technical Manual provides an overview of classification effectiveness (Chapter 19).

Organizational Gains from Test Validation

Economists and psychologists, each with different academic backgrounds, provide evidence for cost savings in improving classification effectiveness based upon aptitude/ability test scores and the addition of new measures that improve the person/job fit. For economists, the savings is usually estimated on the military front end – the costs of recruiting, compensation, and training (e.g., Harris, McCloy, Dempsey, DiFazio, & Hogan, 1994). The linear programming algorithm applied by Harris et al. at the time of their research resulted in an estimated 114 million dollar savings over four years by merely adding a composite of spatial ability measures to the existing ASVAB. For psychologists, the savings is usually estimated further out – the costs of attrition, poor job performance, and not being able to retain personnel over time (e.g., Schmidt, Hunter, & Dunn, 1995). Schmidt et al. showed through putting a dollar value on job performance that an estimated 83 million dollar savings could be saved annually by merely adding a perceptual measure to the ASVAB. Further large savings were estimated by adding a psychomotor measure.

No matter the methods, both the economic and psychological/testing disciplines agree that in large-scale testing programs such as the ASVAB, adding tests that provide even a small increment in predictive validity can result in large cost savings. We recognize that the cost savings developed in theoretical frameworks are always mitigated by the operational situation. That is, each Service must factor in their job assignment needs, which are sometimes tied to available training seats for a particular time frame, or monthly goals for a critically undermanned occupation. Even so, an important goal for the Services in their selection and classification programs should be to lower personnel-related costs to the extent possible through optimizing the ASVAB and adjunct classification tests and their use in setting occupational standards, which are tied to the ASVAB's validity in predicting performance outcomes (training, for the Navy).

The validity coefficients for the Navy's operational ASVAB classification composites average at about .55 (1.00 being the maximum value) across Navy Ratings and range from about .25 for the physically and mentally challenging SEAL (Sea, Air, and Land special warfare combat forces) training to about .85 for the highly academic and difficult Nuclear Field (NF) courses. As an example of cost-avoidance, the Navy sets an annual recruiting goal of about 3,000 enlistees for the NF Ratings (there are three). The NF's ASVAB composite cutscore (252) is so high that it qualifies only about the top 7% of the ASVAB normative population (conceptually serving as the military applicant population). A conservative estimate of the cost of training a NF candidate some years ago was $100,000. A conservative estimate of the NF training graduation rate in a good recruiting environment is about 80% (when including graduates and failures for academic reasons and no others in the study sample). Given an estimated .85 validity coefficient, the stringent ASVAB cutscore, and the 80% training graduation rate, lowering the ASVAB cutscore by 12 score points (one third of a standard deviation) to qualify an additional 5% of youth (say roughly from 5% to 10%) would result in an expected 10% decline in the NF graduation rate (see Table 15 of the NF study — Appendix B). The training cost difference between a hypothetical 70% graduation rate compared to 80% is 300 more students failing X $100,000 (cost to train per each student) = $30,000,000 (recurring each year).

We recognize that the $100,000 cost to train a Navy recruit for a technical occupation (truly an underestimate at this point in time for NF) is not a complete waste because most failed students (for academic reasons) are reclassified to other Ratings/training. However, avoiding unnecessary financial costs for training and retraining should be an important military goal, and it is a Navy goal in setting effective ASVAB standards. The point is that, in recalibrating an ASVAB standard for an occupation, or even in developing a cutscore waiver policy, we must consider the difficulty of the training, the number of recruits going through training, the validity of the best fit ASVAB composite, and the appropriateness of the ASVAB cutscore, including the impact of that cutscore on the availability of talent across other Ratings. This can best be done by a systematic, preemptive approach mandated by a centralized selection and classification agency rather than through reflex reactions to currently observed problems that might have been developing over time.

We note that before the Navy's ASVAB Validation/Standards program was established, the Navy Recruiting Command (NRC) and the Enlisted Community Managers (ECMs) could make on-the-fly decisions about how many ASVAB score points could be waivered (exception to policy) and for what Ratings without any supporting empirical evidence (another reason for centralized S&C functions and controls). If NRC or the ECMs were allowed to arbitrarily issue ASVAB cutscore waivers in a recruiting downturn, the consequences would be decidedly negative. The negative impact at the Rating level would depend on such factors as the training difficulty, length, and expense, the level of academically related failure rates at the time, and the magnitude of the ASVAB classification composite's validity coefficient with more negative impact associated with larger validity magnitudes.

During the time of the two manuals' development, the military recruiting environment was extremely positive and therefore with no apparent need for improving military classification effectiveness (or for recalibrating ASVAB standards or issuing ASVAB cutscore waivers). There was, however, recognition that the economy is cyclical and that the positive military recruiting environment would not continue indefinitely (Gilroy, 2011). The opportunity now exists to preemptively improve selection and classification with candidate ASVAB tests and adjunct occupational classification tests recommended by an expert ASVAB review panel (Drasgow, Embretson, Kyllonen, & Schmitt, 2006). These tests will benefit recruiting by improving the person-job fit. That is, test validation research tells us that these tests will increase the proportion of annual recruit populations occupation qualified and also increase performance success rates. An improved person-job fit could also result in higher enlistee retention rates (via improved job satisfaction), thereby potentially mitigating both recruiting and retention issues (e.g., Sailors leaving the Navy after only one term of enlistment thereby increasing both technical training costs in the aggregate and recruiting costs due to the need to recruit replacements above those normally projected).

A good case can be made to conduct ASVAB validation/standards studies on a continual basis, but particularly during recruiting downturns because policy makers will need to know the expected impact on training performance and associated costs due to, on average, lower ASVAB recruit population aptitude/ability scores. Lower ASVAB scores will result in lower average training performance scores in many difficult/ complex training courses merely because of the strong ASVAB relation with final school grades. However, a lower proportion of ASVAB scores in the upper score range can also occur. Adding to the lack of high scores could be the propensity to issue ASVAB cutscore waivers, shifting the whole distribution of scores to the left. Therefore, it is recommended that each Service maintain a continuing ASVAB Validation/Standards program, preferably as part of an S&C function. The Navy has an S&C function (OPNAV132G) that sponsors not only ASVAB validation/standards studies, but the supporting technical processes including the development of the manuals. Appendix D contains the "INTERSERVICE Aptitude/Ability Standards Panel Charter" approved by the MAPWG in 2012 for conduction joint-service aptitude/ability validation studies with the purpose of minimizing training costs incurred by a Service that conducts joint-training when one Service's ASVAB standard is misaligned. This charter and the two manuals could serve to justify S&C functions across Services in this era of fiscal austerity where they might otherwise be at risk of nonsupport.

Manual Structure and Target Audience

The Introductory Manual is intended to provide information and context for conducting an ASVAB validation/standards study for individuals with diverse backgrounds. Readers are not expected to have a sophisticated statistical background; however, the researcher who conducts the studies will benefit from the content. Key audiences for this Introductory Manual include

- policy makers, managers, and researchers responsible for conducting ASVAB validation studies,
- military personnel working with contractors and entry level behavioral scientists,
- Enlisted Community Managers (ECMs) who monitor and manage Navy schools' failure rates and unfilled classroom seats, and
- school instructors responsible for evaluating students.

By outlining the factors that affect validation research, the two manuals are designed to help researchers and practitioners obtain accurate test validation results and make sound decisions about instating or revising military occupations' ASVAB standards. The references provided in both manuals are intended to complement the subject matter and it is the hope that new researchers will benefit from having this starting point for an array of topics threaded together in one place.

After reading both manuals, individuals conducting ASVAB validation/standards studies should be able to

- outline the procedures involved,
- describe key issues that should be taken into consideration,
- understand when and how to seek guidance, technical and otherwise,
- determine the strengths and limitations of a particular ASVAB validation study,
- draw appropriate conclusions from validation research, and
- develop a report with recommendations.

Introductory Manual Chapters

Chapter 1 establishes the purpose of the Introductory Manual, and briefly, the Technical Manual. The Introductory Manual is intended to provide general guidance on conducting ASVAB validation/standards studies and will be most useful for sponsors, policy makers, and leadership from the recruiting, training, and the enlisted communities. All of these entities benefit from effective ASVAB standards and so are considered stakeholders. The Technical Manual is intended to provide the researcher who conducts ASVAB validation/standards studies with the "how to do it" information.

Chapter 2 provides a brief history of the ASVAB and a description of the aptitudes, abilities, and knowledge constructs that it measures. Also described is a large-scale, joint-service test development and validation effort that was intended to augment the ASVAB with tests that better represent the ability to reason abstractly and solve problems than the ASVAB, which largely measures the ability to apply accumulated knowledge, skills, and abilities. Included also is a list of criteria for evaluating candidate tests developed in the past noting that recent revisions have been made and approved by the Manpower Accession Policy Working Group (MAPWG) that oversees the ASVAB.

Chapter 3 provides a discussion of a framework for understanding predictors and criteria both as overt measures of what we intend to measure, and as underlying constructs. Chapter 4 extends the focus of the ASVAB criterion measure, training performance, to job performance, which the ASVAB does predict depending upon what aspects are measured. The chapter highlights issues involved in predicting job performance such as the sparse literature on the criterion compared to the predictor, the perennial "Criterion Problem", and the difficult decisions about whether to consider single or multiple criteria.

Chapter 5 focuses on training performance and describes the evolution of Navy training from instructor-led, group-paced classroom courses with hands-on laboratory/demonstrations to self-paced, computer-based training. The chapter provides guidelines for best practice performance measurement in training, in particular for simulation-based training. Also provided are lessons learned from the literature on successful and unsuccessful computer-based training, which has guided the Navy's evolution to a blended training solution.

Chapter 6 lays out the process and steps for conducting Navy ASVAB validation/ standards studies. Although the phases of a study generalize for each Rating, the reader should know that each study is tailored to a Rating's particular issues that either directly or indirectly relates to the effectiveness of that Rating's ASVAB standard. The overall goal of any ASVAB validation/standards study is to provide a standard that minimizes academically related failure and setback rates while addressing the Navy's need to fill all Ratings with aptitude/ability-qualified Sailors.

Chapter 7 gives an overview of synthetic validity, an indirect approach to establishing test validity when (a) sample sizes are too small to conduct a formal criterion-related validation study or (b) new jobs are formed and the need to establish an ASVAB standard precedes the availability of the performance criterion data to use in ASVAB validity analysis. The Navy applies a modified version of synthetic validity that involves comparisons of training difficulty, time allowed to train, similar Ratings' ASVAB standards, and referencing cutscores to the ASVAB normative population.

Appendices A, B, and C contain examples of Navy ASVAB validation/standards studies to illustrate that one size does not fit all. That is, the three reports (Navy SEALs, Nuclear Field, and Mineman Ratings) are provided to (a) demonstrate the dynamics and issues experienced by each Navy community, (b) show how the study methods are tailored to address these issues, and (c) address cutscore setting for a range of magnitudes in the ASVAB composites' validity coefficients. Appendix D contains the INTERSERVICE Aptitude/Ability Standards Panel Charter (for conducting joint-service ASVAB validation/standards studies when one Service trains other Service members).

Chapter 1. References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999*). Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Campbell, J. P., & Knapp, D. J. (Eds.) (2001). *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Erlbaum.

Drasgow, F., Embretson, S. E., Kyllonen, P. C., & Schmitt, N. (2006). *Technical review of the Armed Services Vocational Aptitude Battery (ASVAB)* (FR-06-25). Alexandria, VA: Human Resources Research Organization.

Gilroy, C. (2011, May) *The All Volunteer Force: Let the good times roll.* Briefing given to the Defense Advisory Committee on Military Personnel Testing, Miami, FL.

Green, B. F., & Mavor, A. S. (Eds.) (1994). *Modeling cost and performance for military enlistment*. Washington, DC: National Academy Press.

Harris, D. A., McCloy, R. A., Dempsey, J. R., DiFazio, A. S., & Hogan, P. F. (1994). *Personnel enlistment testing, job performance, and cost: A cost-effectiveness analysis (*ARI Technical Report 1016*)*. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29,* 340-362.

Johnson, C. D. & Zeidner, J. (1991). *The economic benefits of predicting job performance (Vol 2)*. NY: Praeger.

Kuncel, N. R., & Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science. 19*, 339-345.

Lightfoot, M. A., Ramsberger, P. F., & Greenston, P. M. (1991). *Matching recruits to jobs: Enlisted Personnel Allocation System. (*ARI Special Report 41*)*. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

McCloy, R. A, Campbell, J. P., Knapp, D. J., Strickland, W. J., & DiFazio, A. S. (2006). *A framework for conducting validation research with the Armed Services Vocational Aptitude Battery (ASVAB)* (HumRRO Final Report FR-06-15). Alexandria, VA: Human Resources Research Organization.

Rostker, B. D. (2006). *I Want You! The Evolution of the All-Volunteer Force*. Santa Monica, CA: RAND Corporation.

Rumsey, M. G., Walker, C. B., & Harris, J. H. Eds.) (1994). *Personnel selection and classification*. Hillsdale, NJ: Erlbaum.

Schmidt, F. L., Hunter, J. E., & Dunn, W. L. (1995). *Potential utility increases from adding tests to the Armed Services Vocational Aptitude Battery (ASVAB)* (NPRDC-TN-95-5). San Diego, CA: Navy Personnel Research and Development Center.

Chapter 2.
An Overview of the ASVAB
Fritz Drasgow

Introduction

The military services have an ongoing need to develop and maintain effective enlisted personnel selection and job classification standards. The cognitive aptitude/ability standards, for the most part, are developed using various combinations of the tests of the Armed Services Vocational Aptitude Battery formed into composites (ASVAB; Segall, 2004). The primary goal guiding the development and evaluation of ASVAB standards is to ensure that enlisted personnel are assigned to technical training for which they exhibit a high likelihood of success while at the same time meeting the Services' annual recruiting requirements (not missing goal). ASVAB validation/standards studies are conducted to evaluate the extent to which these objectives are being met. This chapter provides some basic information about the ASVAB that will be helpful to those not familiar with the battery and to those who want to refresh their ASVAB knowledge. We refer the interested reader to the official ASVAB website, http://official-asvab.com/history_res.htm, which provides an array of ASVAB information including the history of the ASVAB's development.

A Brief History of the ASVAB and ASVAB Candidate Cognitive Tests

Until 1976, the individual Services researched and maintained their own enlistment program selection and classification tests. For example, the Navy maintained a battery of tests prior to the ASVAB called the Basic Test Battery (BTB). The consolidation of the military testing enterprise enabled efficiencies in testing and the application of test scores for all Services so that individuals changing their minds on which Service to join would not have to retest on a different Service-specific test battery. With a common ASVAB, however, it has become more difficult to consider and operationalize Service-specific classification tests that show value, sometimes in a limited context. The current position of the Manpower Accession Policy Working Group (MAPWG), which has a vested interest in the ASVAB's development, is that an enlistment testing model should be adopted to allow for the flexibility of adding tests. Such a model could be based upon a common core set of ASVAB tests administered to all military applicants, with Service-special tests administered either seamlessly after the computer adaptive version of the ASVAB (CAT-ASVAB) or later in the occupational classification process.

The addition of new military service classification tests to the CAT-ASVAB platform requires a stringent review of their psychometric properties. Not only would computer-administered tests be the most expedient method for test delivery, but Department of Defense (DoD) (headed by the Defense Manpower Data Center, Personnel Testing Division [DMDC-PTD], executive agency for ASVAB development, research, and maintenance) is seeking to eliminate the paper-and-pencil (P&P) ASVAB. The CAT-ASVAB is administered at the nation's 65 Military Entrance Processing Stations (MEPS).

The P&P ASVAB forms have historically always been administered to applicants in the testing sites outside of the MEPS, the more remote Military Entrance Testing Sites (METS). However, P&P forms at the METS are now mostly replaced by internet-delivered proctored CAT-ASVAB (iCAT). High school students still take the P&P version of the ASVAB under the Career Exploration Program (CEP), but iCAT is also being considered for this program. High school students can make up from 10-14% of military accessions, so the CEP is considered an important military youth market segment. We note that DoD is in the process of completely eliminating P&P ASVAB delivery.

Historically, all of the Services have been involved heavily in the research and development of candidate ASVAB tests, including those that were more representative of fluid intelligence (as compared to the ASVAB, which is more representative of crystallized intelligence; Cattell, 1943). As CAT-ASVAB came on line, it became plausible to consider more fluid intelligence-based tests with their graphically rich formats that are well suited for computer administration. The DoD and the Services supported a joint-service developed and validated battery of tests of fluid intelligence called the Enhanced Computer-Administered Test (ECAT) Battery (Alderton, Wolfe, & Larson, 1997).

DoD was supportive of adding at least one of the ECAT tests to the ASVAB in the late 1990s if it met psychometric and practical criteria (discussed later in this chapter). The test that best met the criteria at the time was Assembling Objects (AO). The AO test, and Army developed test (Held & Carretta, 2013) had the advantage over other ECAT tests because it could be administered both in P&P and computer formats. Most recently, an ECAT working memory test, Mental Counters (MCt) (Alderton et al., 199 7; Larson & Saccuzzo, 1989), has met multiple criteria, but only for computer delivery, which does not seem to be a problem as P&P ASVAB is being phased out.

Military applicants who are administered the P&P ASVAB can qualify for an occupation that uses AO in their ASVAB classification composite by qualifying on another composite (an alternative) formed with only ASVAB tests. The alternative composite model is currently used only by the Navy when the primary most valid composites contains AO or Coding Speed (CS), a formerly official ASVAB test that was retained by the Navy after it was eliminated from the ASVAB (Held & Carretta, 2013). The alternative standards model is a practical solution for occupational qualification that can benefit from added ASVAB constructs; however, there are concerns on a psychometric basis that those just qualifying at the cutscore on one composite and not at all on the other, would do so based upon test measurement error.

The ASVAB Tests

The ASVAB consists of nine tests that measure various aptitudes, abilities, skills, and knowledge. Each ASVAB test is standardized to a mean score of 50 and a standard deviation of 10, based on a representative sample of U.S. youth. The most recent ASVAB norms were developed for the ASVAB 1997 Profile of American Youth (PAY97) (Segall, 2004).

Brief descriptions of the ASVAB tests are provided in Table 2-1.

Table 2-1
ASVAB Tests Content

Test Name	Test Content
General Science (GS)	Knowledge of biological and physical sciences
Arithmetic Reasoning (AR)	Solving arithmetic word problems
Word Knowledge (WK)[a]	Identifying synonyms or the meaning of words in context
Paragraph Comprehension (PC)[a]	Obtaining information from written passages
Mathematics Knowledge (MK)	Performing operations in algebra, geometry, fractions, decimals, and exponents
Electronics Information (EI)	Knowledge of electrical principles and electronics
Auto and Shop Information (AS)	Knowledge of automobile, tool, shop terminology and practices
Mechanical Comprehension (MC)	Understanding mechanical and physical principles
Assembling Objects (AO)	Determining how an object will look when its parts or connection points are put together

[a]WK and PC are combined to form the Verbal (VE) composite that is a component of the AFQT and several Navy ASVAB classification composites.

In Table 2-1, WK and PC are combined to form the Verbal (VE) test combination, with WK weighted approximately 2/3 and PC 1/3. The VE test is used in the Armed Services Qualification Test (AFQT) for enlistment eligibility. The ASVAB individual tests, including VE, are scored on a standard score scale that was derived to have a mean of 50 and standard deviation (SD) of 10 developed for the PAY97 ASVAB normative population. Each Service, however, has the latitude to score their composites differently, as described in Segall (2004). Briefly, Navy composite scores are calculated as the sum of subtest standard scores; for Army and Marine Corps, composite scores are further standardized to have a mean of 100 and SD of 20; for Air Force, on a percentile metric tied to a representative applicant population.

Two major score scale changes were operationalized with implementation of the PAY97 ASVAB norms. First, P&P ASVAB tests as well as CAT-ASVAB are scored based on Item Response Theory (IRT) (see, for example, Embretson & Reise, 2000; Lord, 1980) rather than the former number of items correct used in P&P forms. Second, with implementation of PAY97, there are no upper and lower score bounds (formerly 20 and 80). (We note that the bulk ASVAB scores for military applicants are within the 20 to 80 score range.)

The CAT-ASVAB

A major advantage of the computerized adaptive version of the ASVAB, the CAT-ASVAB, is that it takes the typical applicant less time to complete (about 1.5 hours) than its counterpart P&P version (about 3 hours). The shorter CAT-ASVAB testing time allows the Services to (a) contain applicant processing time to one day in many cases rather than two, thus lowering Recruiter oversight and overnight lodging costs, and (b) consider the addition of classification tests that demonstrate substantial classification payoffs.

CAT-ASVAB, as with all computerized adaptive tests, reduces testing time because the test items administered to an examinee are quickly tailored to an appropriate difficulty level. That is, the adaptive algorithm takes into account the examinee's response (correct or incorrect) to previous items (initial items are of average difficulty) and then efficiently establishes the level of difficulty for the next item. In contrast, in the P&P ASVAB version, the same set of items of each subtest are administered in lock step to a group of examinees (by a Test Administrator) that have individual aptitude/ability differences.

Table 2-2 displays the number of items for each ASVAB test in the P&P and CAT versions and the testing time limit for P&P ASVAB.

Table 2-2
ASVAB Test Number of Items and Test Times

Test Name	P&P Number of Items	P&P Testing Time	CAT Number of Items
General Science (GS)	25	11	16
Arithmetic Reasoning (AR)	30	36	16
Word Knowledge (WK)	35	11	16
Paragraph Comprehension (PC)	15	13	11
Mathematics Knowledge (MK)	25	24	16
Electronics Information (EI)	20	9	16
Auto and Shop Information (AS)	25	11	22[a]
Mechanical Comprehension (MC)	25	19	16
Assembling Objects (AO)	25	9	16
Total	**225**	**143**	**145**

Notes. (1) In CAT-ASVAB, Auto Information and Shop Information are administered as separate 11-item subtests. (2) CAT-ASVAB time limits can be found at http://official-asvab.com/docs/asvab_fact_sheet.pdf. These published time limits were set so that 99% of examinees can finish with the average time spent much lower.

Table 2-2 on the previous page shows a fixed number of items for both the P&P ASVAB and the CAT-ASVAB. The CAT-ASVAB has a smaller number of items because of the efficient item delivery algorithm. Time limits are only displayed for the P&P ASVAB subtests; however, there are very generous time limits for the CAT-ASVAB, which are provided in the "ASVAB Fact Sheet" available on the official ASVAB web site (http://www.officialasvab.com). More information about the history of the development of the P&P and CAT versions of the ASVAB can be found on this website and in Sands, Waters, and McBride (Eds.) (1997).

As noted earlier, the AFQT is used to determine eligibility for military service. The AFQT is the sum of 2VE + AR + MK subtest standard scores transformed from IRT theta scores, transformed again into a uniform score distribution ranging from 1 to 99 (the percentile %ile metric) applied to the PAY97 (Segall, 2004). Each Service has its own minimum AFQT score requirement (35 is the current cutscore for the Navy). The AFQT score of 50 (50% of the ASVAB normative population scoring at or above 50) is a particularly meaningful score as it allows for certain military enlistment incentives. For the interested reader, percentile-to-standard score equivalents for several scales are provided in Appendix A of Ghiselli, Campbell, and Zedeck (1981).

What the ASVAB Predicts

Much research has demonstrated that scores on the ASVAB predict success in training (e.g., see the three ASVAB validation/standards studies in this manual's appendices). Ree and Earles (1991), for example, found an average multiple correlation greater than .60 when predicting final school grades across 82 job training courses for 78,041 Air Force enlistees. As noted in Chapter 1, the Navy, from individual ASVAB validation/standards studies has estimated an average validity coefficient of .55 (corrected for the range restriction effects due to the ASVAB standard) when predicting final school grades (the range spanning from .25 to .85). None of these efforts, however, tracked military personnel into jobs to estimate the ASVAB's prediction of job performance. As we will see in the Technical Manual, a project of this type requires greater sophistication in statistical design when job performance prediction is of interest, including accounting for those who failed training and therefore were not in the jobs to be measured on job performance. Also noted in Chapter 1 is that, for the Navy, training performance is the criterion upon which to establish ASVAB standards.

The Army has long supported a research program to develop alternative measures (cognitive and non-cognitive) that predict important job performance dimensions, recognizing that the ASVAB was developed primarily to predict training performance. Oppler, McCloy, Peterson, Russell, and Campbell (2001), for example, found the ASVAB to predict core technical proficiency and general soldiering proficiency with average multiple correlations (from regression analysis) above .60 across a range of MOS, including Infantryman, Cannon Crewmember, and Medical Specialist. In contrast, Oppler et al. found the ASVAB to predict effort and leadership with an average multiple correlation of .37 but to predict maintaining personal discipline and physical fitness and military bearing with mean multiple correlations of only .17 and .16, respectively. These findings highlight that ASVAB composites are expected to better predict task performance than contextual performance.

Guidelines for Evaluating New Predictors

DMDC-PTD, within the MAPWG committee, has set guidelines that each Service follows in proposing a new test be administered on the CAT-ASVAB platform (either as a candidate new ASVAB test or a Service-special test). This section outlines a set of studies and analyses that were considered in developing the DMDC-PTD "Checklist". No single study is viewed as adequate for making such a complex and significant decision about adding a test to the ASVAB; rather, several studies are needed addressing a wide range of psychometric and other issues. We turn now to the information Dr. Bruce Bloxom, the former MAPWG technical committee chair during the CAT-ASVAB development days, requested when the tests of the Enhanced Computer-Administered Test (ECAT) Battery were being considered for inclusion in the ASVAB (see Alderton et al., 1997; Wolfe, Alderton, Larson, Bloxom, & Wise, 1997; and the special issue of *Military Psychology*, Vol. 9[1], 1997, for more information about the ECAT). Dr. Bloxom's list of documentation required for adding tests to the ASVAB included

1. construct definition;
2. rationale;
3. item taxonomy;
4. information regarding the likelihood of subgroup bias;
5. measurement precision;
6. internal consistency;
7. score-conditional precision by subgroup;
8. model-based precision;
9. subgroup differences in item functioning;
10. validity (internal and external);
11. incremental validity over ASVAB;
12. information about whether equating was subgroup dependent;
13. feasibility for addition to CAT-ASVAB;
14. minimum and maximum times;
15. evidence that instructions were appropriate for low-AFQT applicants;
16. low susceptibility to cheating, compromise, practice, coaching;
17. evidence of no floor or ceiling effects;
18. how applicants have perceived the test (affectively);
19. limitations of the test for CAT-ASVAB administration;
20. uses of standard equipment; and
21. Appendix: Instructions, items, and feedback

In the following sections, we review some of the above criteria for tests to be considered for inclusion to the ASVAB that are especially relevant for ASVAB validation efforts. Many of the other listed criteria are addressed by the specific Service that developed a new test and submitted as part of the MAPWG's as new test implementation requirements.

Construct Definition, Rationale, and Item Taxonomy

The first three criteria (construct definition, rationale, and item taxonomy) should be addressed when a new measure is being developed. The developer should carefully define the characteristic assessed by the test and provide a rationale for its use. The types of items included on the test should flow logically from the construct definition. Information about the construct definition, rationale, and item taxonomy should be included in the documentation describing the development of the new measure.

Measurement Precision

Measurement precision (Criterion 5) refers to a test's accuracy in estimating an examinee's ability level and is reflected in the reliability coefficient and the index of measurement error (standard error of measurement). Two theoretical frameworks have been used for assessing measurement error for the ASVAB, the Classical Test Theory (CTT) (early use for P&P tests) and Item Response Theory (IRT) (highly suitable for CATs but also applicable for P&P tests). The CTT method, explained in depth in the Technical Manual, assumes the same precision of measurement over the total ability range, which may or may not be the case, whereas the IRT method does not (IRT being more item-centric).

Reliability estimates from correlating scores across parallel P&P versions of the ASVAB are documented in the published book on the development and psychometrics of the CAT-ASVAB (Moreno & Segall, 1997, p. 172). Reliability information can also be found on the official ASVAB website for both P&P forms (http://www.officialasvab.com/docs/asvab_techbulletin_4.pdf) and CAT-ASVAB for IRT methods (http://official-asvab.com/reliability_res.htm.

The potential ASVAB test, Information and Communications Technology Literacy Test (ICTL) (referred to now as the "Cyber test") is operational for the Air Force and currently being validated by several of the Services. Trippe and Russell (2011) show that the test has greater precision of measurement in the higher ability range, which is desired when the military is assessing high ability relevant for complex occupations with difficult training courses. In general, the measurement precision/reliability of all predictors used for military selection and classification should be, at a minimum, comparable to those of the ASVAB tests, and even higher if there is an ability measurement gap in the ASVAB for a construct that is important to consider for qualification to critical occupations.

Internal Consistency

Internal consistency (Criterion 6) means the extent to which the test is measuring the same underlying construct. There are several methods for estimating internal consistency – one being coefficient alpha (Cronbach, 1951). Internal consistencies estimated from scores on several P&P versions of the ASVAB have ranged from .80 to .90 (Drasgow, 2003). The classical test theory methods of establishing internal consistency are not appropriate for a computerized adaptive test (CAT) because a CAT does not provide item level data that are the same for each examinee. Alternatively, the standard error of measurement (measurement precision, Criterion 5) can be used for adaptive tests, not only for a total group reliability assessment, but also to assess if there are differences in reliability for important subgroups (e.g., gender and race/ethnic groups – Criterion 9). Another approach to assessing internal consistency for a CAT involves administering an over-length version of the test (a larger number of items). The items could be split to make two forms whereby odd-numbered items make up one form and even-numbered items, the other form. Then, the IRT method would be used to create two estimates of each examinee's ability level on the construct or trait being assessed by the test (a split-half reliability method applied to a CAT).

The reliability coefficient derived from the split-half method for a CAT (with sufficient items and sample size) should be compared to the reliability derived from the IRT method that estimates a "marginal" reliability (across the whole ability range) from the test information function and error of measurement - usually uneven across that range. We refer the reader who is interested in historical developments in psychometric theory to Guilford (1936) for more on the factors that affect test reliability and also to other widely available resources, including Sands et al. (1997).

The test-retest method of estimating test reliability is a measure of the stability of examinees' test scores over time and is therefore technically not a measure of a test's internal consistency. Test-retest takes into account the state of mind and motivation differences of examinees between testing times, but also real learning that has taken place if the period between tests is long. It is difficult to ascertain examinees' motivation levels during either P&P ASVAB or CAT-ASVAB over time, especially because some examinees may not be taking the ASVAB seriously on the first occasion - just seeing what the test is all about without serious intentions to enlist in the military. We note that examinees who initially take the P&P ASVAB must be administered the CAT-ASVAB upon retest. Examinees who take the CAT-ASVAB initially will test on CAT-ASVAB again with a different item pool (loosely referred to a different "form"). Current retest policy is a 30-day wait from the initial test before a retest, another 30-day wait for a second retest, and 6-months thereafter. Other types of reliability are described in the Technical Manual.

Validity

For Criterion 10, internal validity refers to the extent to which a relation between two variables is free from the effects of uncontrolled influences, such as an examinee's experience taking an initial ASVAB on a subsequent ASVAB retest (which we hope is not substantial). External validity refers to the generalizability of the results of a study to future similar situations, which we would hope occurs when we estimate the predictive validity of a new test, say Cyber Test, on one sample of cyber course students and repeat the process on another. A predictive validity, or criterion-related validity study (Criteria 11) refers to the observed relation between the test and the immediate criterion, say training performance. However, the first phase of estimating the relation is a concurrent validity study "for research purposes only" and should include several military occupations strategically selected on the basis of their importance to the Services and their apparent differences in job demands.

Generally, a newly developed test is administered initially in a concurrent validity design to military trainees in the schoolhouse – the same point at which the training performance measures are administered (and later, in a predictive validity design if warranted by the initial findings). This concurrent validation study can address a number of the criteria on Dr. Bloxom's list, including internal validity (construct validity). For example, the item-level data can be factor analyzed to assess the latent structure of the new test. If a preliminary definition of the expected construct measured by the test is sufficient, then the factor analysis results should show a strong first factor with little test score variance accounted for by additional factors. Evidence of multiple strong factors would cast doubt on the construct definition of the test and possibly suggest that separate constructs be measured by separate tests.

A criterion-related validity study, or predictive validity study, should assess not only the validity of the new test, but also the incremental validity that the new test provides over and above the ASVAB. This assessment would be straightforward if there were no prior selection based on ASVAB scores (or any other variables): (a) Enter the ASVAB tests in step 1 of a hierarchical multiple regression equation, (b) add the new test in step 2, and (c) examine the increase in R^2 when the new test is added. However, enlistment eligibility and occupational classification are both based on ASVAB scores, so there is restriction in range on score variability that most likely will result in biased outcomes (addressed in the Technical Manual).

Timing Study

It is important to conduct a study to determine the amount of time to allow examinees to complete the new test (Criterion 14). One rule of thumb is that the time is appropriate if 90% of the examinees complete the test. Groups of representative examinees could be administered the test with varying time limits to determine how much time is needed for 90% of them to finish. We note that operationally, the CAT-ASVAB has more lenient test time limits.

Susceptibility to Cheating, Compromise, Practice, or Coaching

A study is needed to examine the possibility that practice, coaching, or some other non-trait-related manipulation can substantially affect the new test's scores (Criterion 16). Tests of fluid intelligence, for example, sometimes can be coached to produce large score gains (Flynn, 1987). Establishing the coach-ability of item types is especially important for non-cognitive measures for which it may be possible to "fake good", as demonstrated by Army researchers in a carefully developed and validated study of the Military Applicant Profile (MAP). When the MAP was used in an operational setting, scores greatly increased and its validity declined to near zero (White, Young, & Rumsey, 2001).

Norms

Although not listed as a criterion, test norms, such as the ASVAB PAY97 norms (Segall, 2004), are needed so that applicants' scores can be meaningfully interpreted. For example, the military develops policy for enlisting recruits with a specific AFQT score requirement. Each applicant can be placed on a score scale that was developed for the full range of youth between the ages of 18 and 23. The Services cannot expect to establish norms immediately for a new test developed for inclusion in the ASVAB, but there are analytical methods for projecting the new test's norms given that (a) there is not severe restriction in range in ASVAB test scores for the sample used for the new test's administration (due to a stringent ASVAB cutscore) and (b) we have a sufficient sample size. Establishing norms (at least test means, standard deviations, variances, and intercorrelations with the ASVAB tests) for new tests (assuming they relate to important military performance dimensions) will enable the military to assess expected improvements in differential assignment capability resulting from the test's use in classification, that is, improving the person-job fit.

Score Reports

Criterion 21 is "Appendix: Instructions, items, and feedback". The importance of providing good feedback to both examinees and test users is obvious to those involved in the test development and administration enterprise. The importance, however, is often overlooked. It is more difficult than generally realized to develop accurate and comprehensible score reports and the associated interpretive materials. Therefore, sufficient time and resources should be planned for pre-testing with the actual materials, recognizing that revisions may be required.

Additional Criteria on Bloxom's List

The data from a test validation study can be used to address several additional criteria on Dr. Bloxom's list. Likelihood of adverse impact (as opposed to differential item functioning [DIF]) can be examined by comparing mean scores for important subgroups (Criterion 4). The distribution of scores should be inspected to determine whether any floor or ceiling effects are present (Criterion 17). Revision of the test would be needed if such effects were found on either Criterion 4 or 17.

To evaluate the suitability of instructions for examinees of all abilities (Criterion 15) and assess how applicants perceive the test (Criterion 18), a post-test questionnaire should be administered to a sub-sample of validation study participants. Also, IRT could be applied to the item level test data to address several criteria on the list. First, depending on the nature of the construct assessed, the two-parameter logistic (2PL) model (e.g., for personality) or the three-parameter logistic (3PL) model (e.g., for cognitive ability that takes into account guessing) could be fitted to the data. This would enable an assessment of model-based precision (Criterion 8) as well as score-conditional precision. Second, given an adequate sample size, IRT DIF studies could be conducted for important subgroups. This would ascertain whether items functioned differently across groups (Criterion 9) and address the concern underlying Criterion 4 (likelihood of subgroup bias). If only inconsequential DIF is found, then score-conditional precision by subgroup (Criterion 7) and equating by subgroup (Criterion 12) would be automatically satisfied. On the other hand, if a meaningful amount of DIF were found (see Stark, Chernyshenko, & Drasgow, 2004, for an effect size measure for overall DIF), Criteria 7 and 12 would need to be explicitly addressed.

The remaining criteria - feasibility for addition to the CAT-ASVAB (Criterion 13), limitations of the test for CAT-ASVAB administration (Criterion 19), and uses of standard equipment (Criterion 20) - are important for tests being considered for inclusion in ASVAB or as part of the enlistment process. Criterion 20, for instance, would make it too costly to add tests that require special peripheral devices (e.g., headphones for dichotic listening, control sticks and foot pedals for psychomotor coordination). Such tests would need to demonstrate significant improvements in predictive validity or classification efficiency to justify the cost of implementation.

Finally, for any new test, suitability for operational implementation should be assessed by a slow rollout plan. A small number of MEPS might implement the new test operationally with individuals tracked and their performance monitored. CAT-ASVAB was initially implemented at five MEPS sites, and its success was an important factor in the decision to implement it at all MEPS. Although not listed as a criterion, there is a logical requirement for an evaluation of the new test under operational conditions.

DMDC Updated Checklist

DMDC-PTD through the MAPWG has updated Dr Bloxom's "Checklist for Including Tests on the ASVAB Platform". Some of the checklist items, such as establishing job families to demonstrate the utility of a test for more than one occupation are being addressed by the Services. Other items, such as Human Subjects Protection protocols, are specifically addressed by the Services' Institutional Review Boards (IRB). The Services' IRB also interfaces with HQ-USMEPCOM's IRB for final approval of a new test's administration on the CAT-ASVAB platform and the transfer of data for validation efforts. Finally, not all new tests reach the MEPS evaluation stage. Further, the MAPWG must consider not only the MEPS operational schedules and enlistment processing impact from introduction of a test, but also the target population to be tested (i.e., all applicants, or only applicants who reach a threshold score on an ASVAB classification composite). At this point in time, all tests nominated and researched by the Services for CAT-ASVAB administration have been approved and accommodated.

Chapter 2. References

Alderton, D. L., Wolfe, J. H., & Larson, G. E. (1997). The ECAT battery. *Military Psychology, 9,* 5-37.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40,* 153-193.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Drasgow, F. (2003). Intelligence and the workplace. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology, Volume 12, Industrial and organizational psychology* (pp. 107–130). NY: Wiley.

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171-191.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: W. H. Freeman and Company.

Guilford, J. P. (1936). *Psychometric methods*. NY: McGraw-Hill.

Held, J. D., & Carretta, T. R. (2013). *Evaluation of Tests of Processing Speed, Spatial Ability, and Working Memory for use in Military Occupational Classification*. (NPRST-TR-14-1). Millington, TN: Navy Personnel Research, Studies and Technology

Larson, G. E., & Saccuzzo, D. P. (1989). Cognitive correlates of general intelligence: Toward a process theory of g. *Intelligence, 13,* 5-31.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Moreno, K. E. & Segall, D. O. (1997). *Reliability and construct validity of CAT-ASVAB*. In W. A. Sands, B. K. Waters, & J. R. McBride, (Eds.), Computerized adaptive testing: From inquiry to operation. (pp. 219-226). Washington, DC: American Psychological Association.

Oppler, S. H., McCloy, R. A., Peterson, N. G., Russell, T. L., & Campbell, J. P. (2001). The prediction of multiple components of entry-level performance. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 349-388). Mahwah, NJ: Erlbaum.

Ree, M. J. & Earles, J. A. (1991). Predicting training success: Not much more than *g*. *Personnel Psychology, 44,* 321-332.

Sands, B. K., Waters, B. K., & McBride, J. R. (Eds.) (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Segall, D. O. (2004). *Development and evaluation of the 1997 ASVAB score scale* (Technical Report No. 2004-002). Seaside, CA: Defense Manpower Data Center.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item functioning and differential test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology*, *89*, 497-508.

Trippe, M. D., & Russell, T. L. (2011). *Information and communications technology literacy test norming study: Phase III final report* (AFCAPS-FR-2011-0011). Randolph AFB, TX: Air Force Personnel Center Strategic, Research and Assessment Branch.

White, L. A., Young, M. C., & Rumsey, M. G. (2001). ABLE implementation issues and related research. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 525-558). Hillsdale, NJ: Erlbaum.

Wolfe, J. H., Alderton, D. L., Larson, G. E., Bloxom, B. M., & Wise, L. L. (1997). Expanding the content of CAT-ASVAB: New tests and their validity. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 239-249). Washington, DC: American Psychological Association.

Chapter 3.
Mapping Predictors and Criteria
Sarah A. Hezlett

Introduction

Criterion-related validity studies examine the extent to which one or more predictors (e.g., ASVAB tests) relate to a criterion (e.g., a measure of military training performance). To study the relations, individuals' scores on the predictor test are mapped to their standing on the criterion measure, generally through the correlation coefficient. However, there is both an operational concern and a theory-based concern when examining the relations through correlation analysis. The U.S. military has an operational concern in that the correlations between observed ASVAB scores and a measures of training performance are directly used to aid in setting an effective ASVAB standard, which in turn is expected to improve training performance and graduation rates if they are insufficient to begin with. The personnel research laboratory, on the other hand, has a theory-based concern in that analyses should accurately shed light on the "theoretical" relation – that is, the relation between the underlying constructs of both the predictor and criterion, neither of which is directly observable (Guion, 1998). This chapter provides basic information from both operational and theoretical perspectives, as well as more about the ASVAB, the nature of different criteria, and how they relate.

The Predictor/Criterion Framework

Criteria are constructed abstractions that represent outcomes of interest to an organization. Underlying constructs cannot be observed directly, but are behaviors or results that, in principle, can be experienced or realized by the organization. Well-developed criterion measures should be good representations of the criterion constructs.

Scores on predictor tests also represent underlying constructs. Therefore, the system of relations in an ASVAB validation/standards study includes: (a) inferred relations between predictor constructs and predictor measures, (b) inferred relations between the criterion constructs and criterion measures, (c) unobservable relations between predictor and criterion constructs, (d) observable relations between the predictor and criterion measures, and (e) the inferred relations between the predictor measures and criterion constructs. These types of relations are fully discussed by Binning and Barrett (1989) and Schmitt and Landy (1993) and we consider them to be helpful for researchers in designing good validation/standards studies and in drawing more informed conclusions from them.

Figure 3-1 illustrates the relations as a system or framework and is referred to at various points in the chapter.

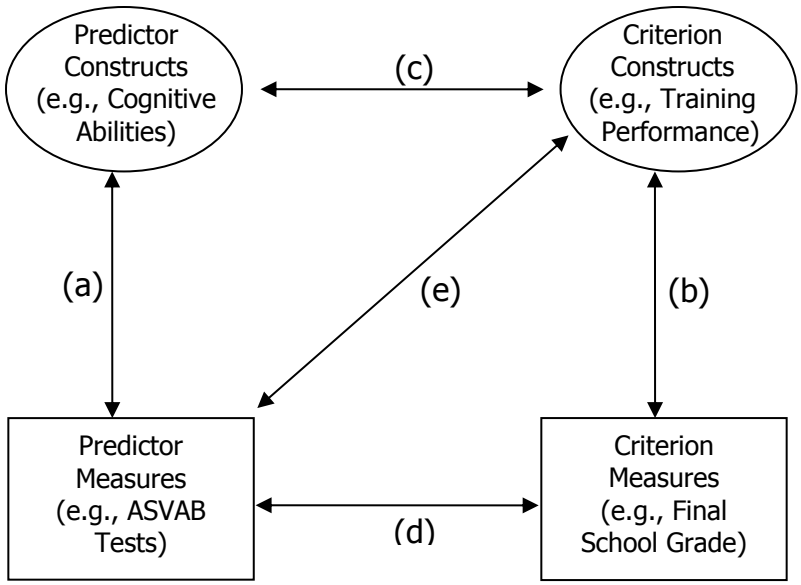

Figure 3-1. System of relations in an ASVAB validation study.

The framework in Figure 3-1 may seem abstract or academic, but it provides valuable insight for planning validation efforts. The operational relation "d" is what the ASVAB validation/standards researcher observes when validating ASVAB test scores against training performance measures (i.e., Final School Grade). The qualification of recruits to occupations is based on relation "d" and the ASVAB cutscore is set to maximize the utility of the ASVAB standard (e.g., acceptable training graduation rates). As diligent researchers, however, we are really interested in the *inferred* relations between the ASVAB tests scores and the underlying criterion constructs (relation "e" in Figure 3-1). The accuracy of relation "e", however, rests on the validity of inferences "a" and "c" (Guion, 1998). As researchers validating a newly developed candidate ASVAB test we would want to know relation "a" for that predictor, relation "b" for the criterion, and ideally, relation "c" between the two constructs before estimating relation "d".

Because psychological measures are not perfectly reliable, the correlations observed between predictor and criterion measures (relation "d" in Figure 3-1) are lower than the true relations between the predictor and criterion constructs (relation "c" in Figure 3-1). If it is found that the criterion is poorly measured, steps can be taken to improve the measurement process or other corrective actions. Thus, knowing the reliability of both the predictor and criterion is important not only in understanding correlation attenuation, but also because steps can be taken to improve both measures and therefore the criterion-related validity of the predictor. The mechanics of estimating reliability and the attenuating effect of unreliability on validity magnitude are discussed in several chapters of the Technical Manual. The sections that follow in this chapter address the relations in Figure 3-1 starting with the predictor, followed by the criterion, and then finally, models of predictor/criterion relations.

The Predictor

Understanding the relations between a set of predictor constructs and their measures (e.g., relation "a" in Figure 3-1) requires consideration of (a) the nature of the constructs, (b) evidence regarding the validity of the measures, and (c) information about the reliability of the measures. These issues are interrelated and the knowledge of the nature of the constructs and validity evidence for specific measures are reciprocally related. That is, construct definitions are used to guide measurement development and validation and validity evidence provide insight into constructs. However, a measure's reliability constrains its validity, potentially curtailing the degree of validity in assessing the construct.

The Construct Domain of Cognitive Abilities

The current purpose of the ASVAB is to provide an assessment of whether military recruits will, at high rates, successfully complete the training for their respective military occupational specialties (referred to as MOS for Army and Marines, Rating for Navy, and Air Force Specialty Code [AFSC] for Air Force). To this end, the ASVAB tests measure multiple aptitudes that generally represent cognitive ability constructs. The best way to characterize the relations between cognitive ability constructs has been a noteworthy and long-running controversy (Kuncel, Hezlett, & Ones, 2004). Following the lead of Spearman (1904, 1927), many researchers have argued that all intellectual abilities are related, with a general component (*g*) accounting for the positive correlations typically observed between tests of cognitive abilities.

Other researchers (e.g., Guilford, 1959; Sternberg, 1985; Thurstone, 1938) have proposed models of intelligence based on the notion that there are multiple specific abilities (*s*). Spearman's two-component model is not inconsistent with the specific ability theory but proposes that scores on each cognitive ability measure are a function of both *g* and *s* (*s* being a specific component unique to a test or a limited set of tests). No matter the theoretical perspective, considerable empirical evidence supports the concept of *g*.

Much of the evidence for *g* is from exploratory and confirmatory factor analyses, which seek to explain the correlations between measures with smaller numbers of unobserved "latent" constructs. A hierarchical model, with *g* being the highest order factor, is generally accepted as the best representation of the latent structure of cognitive abilities (Carroll, 1993; Drasgow, 2003; Kuncel et al., 2004). Carroll's massive review and reanalysis of the factor analytic literature on cognitive abilities found support for several second-order factors. These second order factors included the constructs of fluid intelligence (*Gf*; reasoning abilities) and crystallized intelligence (*Gc*; abilities reflecting the application of acquired knowledge) posited by Cattell (1943), along with memory, visual perception, auditory perception, retrieval, cognitive speed, and processing speed (Carroll, 1993; Drasgow, 2003).

What the ASVAB Measures

Factor analytic results suggest that the ASVAB tests measure some, but not all, of the factors Carroll (1993) identified (Drasgow, 2003). Based on their analyses of an earlier paper-and-pencil (P&P) version of the ASVAB, Ree and Carretta (1995) concluded that the ASVAB's structure was best represented by a hierarchical model with the general factor (*g)* as the highest order factor and three second-order factors. The three second-order factors reflect speed, verbal/math, and technical knowledge. Alternative models, including one with just *g,* and several models with different second-order factors, either fitted the data slightly less well or were more difficult to interpret. Based on similar data, Drasgow also concluded that a model with just a *g* factor was a poor fit to the data and that a better fitting model was one with four correlated factors. These four factors reflected quantitative, verbal, technical, and speed.

The ASVAB measures specific acquired technical knowledge in addition to the academically based AFQT and aptitudes (i.e., spatial ability for Assembling Objects [AO]). The Auto and Shop Information (AS), Electronics Information (EI), Mechanical Comprehension (MC), and, to a lesser extent, General Science (GS) tests capture specific technical knowledge but also, at least for GS, academic achievement, and for MC, some degree of mechanical aptitude. The military views the technical tests as adding to the AFQT's prediction of important training and job performance criteria for many of its occupations. In ASVAB validation studies, the criterion, even for training performance measures, is generally multidimensional; therefore, a narrowly defined ASVAB composite (e.g., AFQT) will generally have lower validity than one that combines several training-relevant ASVAB tests (such as the MC test for mechanics occupations).

The ASVAB Assembling Objects (AO) test is currently the only test that does not require verbal ability (somewhat of a non-verbal reasoning test). AO is a spatial ability test that provides incremental validity to the ASVAB in predicting training grades for some Navy mechanics Ratings and is used operationally in two Navy ASVAB classification composites. In addition to providing incremental validity, the test has an added benefit to the Navy by offsetting lower scores for women and some minority groups when the technical knowledge-based tests are used in a composite (Anderson et al., 2011; Held & Carretta, 2013).

The former speeded ASVAB tests, Coding Speed (CS) and the Numerical Operations (NO), are no longer a part of the ASVAB; however, the Navy has retained CS as a special classification test administered seamlessly to Navy applicants on the CAT-ASVAB platform. The CS test, as with AO, provides incremental validity to the ASVAB and also lowers score barriers for women and some minority groups for some Navy Ratings. We note that DMDC-PTD is evaluating a reworked CS test that reduces the impact on scores due to computer hardware changes, and is also a purer measure of processing speed. The Navy is retaining the CS test as is because a recent hardware effects study showed that a required change in response input — mouse replacing the CAT-ASVAB specialized keyboard, showed no difference in CS scores. Further, there is some evidence that the CS test measures an underlying motivational construct (or perseverance on task) that is relevant for a range of occupations (Segal, 2012). A purer measure of processing speed, whoever, would be a useful addition to the ASVAB.

The ASVAB is not a completely comprehensive measure of the cognitive ability construct space. Although its coverage of the construct space is sufficiently broad to assess *g* and several second-order factors, the ASVAB does not reliably assess all of the second-order cognitive ability domains identified by Carroll (1993) and Drasgow (2003). Memory, visual perception, auditory perception, and reasoning (i.e., fluid intelligence, *Gf*) do not appear to be thoroughly measured, if at all, by the ASVAB. As noted earlier, however, the Services and DMDC-PTD are working toward the inclusion of more measures, the consolidation of redundant content, and a process for ASVAB content changes for future versions as training and jobs change. Currently, the most promising edition to the ASVAB (mentioned in the previous chapter) is a Cyber test, originally called the Information and Communications Technology Literacy Test (ICTL) (Trippe & Russell, 2011) and a working memory test called Mental Counters developed as part of the DoD sponsored Enhanced Computer Administered Test (ECAT) Battery (Alderton, Wolfe, & Larson, 1997; Larson and Saccuzzo, 1989).

The Criterion

Understanding the relations between a set of criterion constructs and their measures (e.g., relation "b" in Figure 3-1) requires consideration of (a) the nature of the constructs, (b) evidence regarding the correlations among the measures (reserving the term "validity" for the predictor), and (c) information about the reliability of the measures. As with the predictor, these considerations are interrelated for the criterion. This section begins with a description of what constructs the criterion would typically measure in an ASVAB validation study that involves the relevant training construct domain. Then, models of job and training performance are discussed to provide a fuller treatment of the criterion space, followed by a discussion of the predictive validity of the ASVAB and other cognitive predictor measures.

The Construct Domain of Training Success

Several taxonomies have been proposed to describe the concept of training success. Kraiger, Ford, and Salas (1993) developed a theoretically grounded taxonomy of learning outcomes based on constructs drawn from diverse fields of research. Their model differentiates and defines three kinds of outcomes: (a) cognitive, (b) skill-based, and (c) affective. Cognitive outcomes include constructs related to knowledge, such as declarative knowledge (e.g., knowledge of facts, principles, and what to do), manipulation of that knowledge, mental models (e.g., our perceptions of a real-world situations that influences our behaviors), and meta-cognitive skills (e.g., high-level mental strategies organizing cognitive skills to solve a problem). Skill-based outcomes encompass the later stages of skill acquisition - compilation and automaticity. Affective outcomes include attitudes (e.g., appreciation of diversity) and motivation (e.g., self-efficacy and goal-setting). Kraiger (2002) refined this taxonomy, providing each category with more detailed descriptions and examples of narrower outcomes.

In contrast, Campbell and Kuncel (2001) outlined a taxonomy of capabilities that potentially can be trained. The four major categories of capabilities are (a) knowledge, (b) observable skills, (c) problem-solving skills, and (d) attitudes and beliefs. Observable skills—which include cognitive, psychomotor, physical, interpersonal, expressive, and

self-management skills—are used in the application of knowledge to solve *structured* problems or to achieve a specific goal. In contrast, problem-solving skills are used in the application of strategies to solve *ill-structured* problems. Whereas Campbell and Kuncel classify the use of cognitive strategies (e.g., heuristics) as problem-solving skills, Kraiger et al. (1993) refer to the term "cognitive" not as a skill but as a learning outcome (cognitive outcome).

Despite the differences in the two taxonomies of training success just described, both illustrate the breadth of potential training outcomes. The construct space of training performance is complex and multidimensional, and the training may be developed to meet diverse objectives (Campbell & Kuncel, 2001). For example, suppose the desired learning outcomes include (a) knowledge of when to use particular medical equipment, (b) skill in using the equipment, and (c) skill in working with other team members. The measures developed to evaluate training should assess all three constructs. Failure to do so may result in a "deficient criterion measure".

Finally, learning outcomes represent only one way of conceptualizing training success and the military may consider training *completion* as a critical outcome. However, the relation between learning outcomes and training completion is complex. For some recruits, difficulty achieving target learning outcomes can lead to challenges in completing the training, especially if recruits are set back. Those who are set back may feel stigmatized and lose confidence in their ability to learn, further affecting their ability to master the material. In other instances, other factors (e.g., personal problems, family demands) may impede both learning and training outcomes.

The Army's Project A – Multiple Job Performance Criteria

We note that the Army has long supported a research program to develop alternative measures (cognitive and non-cognitive as well as psychometric ability) that predict important job performance dimensions, recognizing that the ASVAB was developed primarily to predict training performance. Oppler, McCloy, Peterson, Russell, and Campbell (2001), for example, found the ASVAB to predict core technical proficiency and general soldiering proficiency with average multiple correlations (from regression analysis, test score range restriction, and shrinkage, p. 356) above .60 across a range of MOS, including Infantryman, Cannon Crewmember, and Medical Specialist. In contrast, Oppler et al. found the ASVAB to predict effort and leadership with an average multiple correlation of .37 but to predict maintaining personal discipline and physical fitness and military bearing with mean multiple correlations of only .17 and .16, respectively. These findings highlight that ASVAB composites are expected to better predict task performance than contextual performance.

The U.S. Army Research Institute's "Project A" (Campbell, 1990; Campbell & Knapp, 2001), the Army's portion of the Department of Defense's Job Performance Measurement (JPM) project (Wigdor & Green, 1991), provides insight into job performance as the criterion in military settings. As part of what may be the largest validation study ever conducted, Project A researchers took great care to understand and define what constitutes "job performance" for their military occupation specialties (MOS) and what types of measures predict important elements of job performance.

For first-tour Service members, Campbell, Hanson, and Oppler (2001) identified and described five basic dimensions of performance:

1. *Core technical proficiency*. The ability to perform the tasks that define the MOS. For example, disarming explosives would be a key activity for the Explosive Ordinance Disposal MOS.
2. *General soldiering proficiency*. The ability to perform tasks that soldiers in every MOS should be able to perform. Recognizing friendly and threat vehicles is an example provided by Campbell et al (2001).
3. *Effort and leadership*. "The individual exerts effort over the full range of job tasks, perseveres under adverse or dangerous conditions, and demonstrates leadership and support toward peers. … While appropriate knowledge and skills are necessary for successful performance, this construct is meant only to reflect the individual's willingness to do the job required and to be cooperative and supportive with other soldiers" (p. 310).
4. *Maintaining personal discipline*. "The degree to which the individual adheres to Army regulations and traditions, exercises personal self-control, demonstrates integrity in day-to-day behavior, and does not create disciplinary problems" (p. 310).
5. *Physical fitness and military bearing*. "The degree to which the individual maintains an appropriate military appearance and bearing and stays in good physical condition" (p. 310).

Project A inspired several researchers to refine the job performance construct space. Campbell outlined a taxonomy with eight latent performance components that could be used to describe performance in any job (Campbell et al., 1993), Pulakos led a research initiative on adaptive performance (Pulakos, Arad, Donovan, & Plamondon, 2000), and Borman and his colleagues' differentiated the concepts of task and contextual performance (Borman & Motowidlo, 1993; Motowidlo, Borman, & Schmit, 1997). The Army's Project A work also was instrumental in the development of non-cognitive measures (e.g., personality). The Navy is evaluating the Navy Computerized Adaptive Personality Scales (NCAPS) (Houston, Borman, Farmer, & Bearden, 2006) and for the Army, their more recent personality instrument, the Tailored Adaptive Personality Assessment System (TAPAS) (Drasgow et al., 2012). Non-cognitive measures were recommended for evaluation by the ASVAB Review Panel (Drasgow et al., 2006).

Task versus Context in Measuring Job Performance

Task performance refers to two kinds of activities that tend to be emphasized in job descriptions. These are activities that (a) transform raw materials into the goods or services produced by an organization, and (b) support the production of goods and services, including maintaining the supply of raw materials, distributing finished products, and providing planning, supervising, or staff functions (Borman & Motowidlo, 1993; Motowidlo, 2003; Motowidlo et al., 1997). Contextual performance consists of activities that support the broader environment in which the technical core must function, including behaviors such as volunteering for tasks not formally part of the job,

demonstrating effort, helping and cooperating with others, following organizational rules and procedures, and supporting organizational objectives (Borman & Motowidlo, 1993). Models involving task and contextual performance have attempted to explain the antecedents of job performance, providing further insight into how cognitive ability constructs map to the performance domain. Cognitive ability is thought to play a role in both contextual and task performance (Motowidlo et al., 1997). Cognitive ability is posited to influence three direct determinants of task performance (task knowledge, skills, and habits), but only one of the direct determinants of contextual performance - contextual knowledge. Consequently, task performance is expected to be more strongly predicted by cognitive ability and contextual performance is expected to be more strongly predicted by non-cognitive constructs, such as personality traits.

Research to date has placed more emphasis on examining the paths through which non-cognitive predictors influence performance. Findings have been mixed, with results varying by the specific personality traits, intervening variables, and dimensions studied (for a review, see Johnson, 2003). Research has generally found, however, that cognitive abilities more strongly predict task performance than contextual performance (McHenry, Hough, Toquam, Hanson & Ashworth, 1990). Consequently, researchers can anticipate that scores on ASVAB tests will more strongly predict task performance dimensions than contextual performance dimensions.

Predicting Performance Criteria

Cognitive Tests Predict both Training and Job Performance

Research has consistently found that cognitive abilities are good predictors of not only training performance, but also academic outcomes and job performance. For example, standardized tests used in making decisions to admit individuals to undergraduate and graduate programs have been found to be good predictors of a variety of measures of academic performance including faculty ratings and degree completion (Kuncel & Hezlett, 2007). Much like the ASVAB, college admission tests assess cognitive abilities (Kuncel et al., 2004). Research applying meta-analyses to many research samples has demonstrated that general mental (or cognitive) ability (*g)* is a good predictor of both training criteria and job performance (Ree, Carretta, & Steindl, 2001). Schmidt and Hunter (1998) also used meta-analysis to quantitatively summarize prior research on diverse personnel selection tools and concluded that the average estimated validity of general mental ability was .56 for predicting training performance, and .51 for predicting job performance. Corrected for range restriction on the predictor and unreliability in the criteria, these validity coefficients are good estimates of the tests' average operational validities, or relation "e" in Figure 3-1.

In other research, cognitive ability has been identified as one of the major indirect determinants of job performance (Campbell, McCloy, Oppler, & Sager, 1993; Johnson, 2003; Motowidlo et al.,1997). Models explaining the antecedents of training and job performance provide insight into why cognitive ability is a strong predictor of performance. Models of training effectiveness posit that cognitive abilities influence learning outcomes, such as knowledge and skill acquisition. On the job, knowledge and skill acquisition from learning during training has a direct impact on the transfer of

training and, ultimately, job performance (Colquitt, LePine, & Noe, 2000; Noe & Colquitt, 2002). Consistent with these models, research on job performance has found that job knowledge mediates the relation between cognitive ability and job performance (Borman, White, & Dorsey, 1995; Borman, White, Pulakos, & Oppler, 1991; Schmidt, Hunter, & Outerbridge, 1986).

Results from the training literature are less clear cut. General cognitive ability (*g*) has been found to directly affect the acquisition of new job knowledge and skills, and, to a lesser extent, directly or indirectly influence trainees' motivation and attitudes (Colquitt et al., 2000; Ree et al., 2001). There is also some evidence that *g* indirectly influences the acquisition of new job knowledge through its direct influence on prior job knowledge (Ree et al.). One meta-analysis, however, suggested that only skill acquisition, not job knowledge, has a direct relation with transfer of training (Colquitt et al.). A reasonable conclusion is that *g* primarily contributes to individuals' job performance by influencing their acquisition of job knowledge and/or mastery of skills. Another important issue addressed by some job performance models is determining the categories of job performance needed to describe performance in all jobs. Several models posit that job performance, like training performance, is multidimensional.

Models of Training and Job Performance

First, it seems obvious that learning from training can influence job performance. In the research setting, taxonomies of training outcomes and effectiveness (Campbell & Kuncel, 2001; Kirkpatrick, 1977, 1998; Kraiger et al., 1993), models of training (Noe & Colquitt, 2002), and models of job performance (Campbell et al., 1993; Johnson, 2003) independently posit that the learning that occurs during training influences job performance. Second, empirical research on training and job performance supports these models. For example, acquisition of learning outcomes during training has been found to predict subsequent job performance (Colquitt et al., 2000). Vineberg and Joyner (1982) in their review of predictors of job performance from military studies found miniaturized training and self-paced training improved prediction of job performance over other measures. These findings and others support the assumption that training and job performance are related. It is also important, however, that knowledge, skills, and motivation all contribute to or determine the valid portions of the dimensions of job performance (Johnson, Duehr, Hezlett, Muros, & Ferstl, 2008; McCloy, Campbell, & Cudeck, 1994).

Third, empirical research indicates that learning during training is a necessary but not sufficient determinant of training transfer. That is, additional individual and work environment factors influence the extent to which trainees apply on-the-job what they have learned (Kraiger, 2003; Noe & Colquitt, 2002). An observed weak relation between training performance and job performance may be due to what researchers have observed as the "transfer problem," which can be due to characteristics or conditions that occur within three broad categories: (a) trainee characteristics, (b) training design, and (c) the work environment (Grossman & Salas, 2011). More specifically, Grossman and Salas applied the training transfer model developed by Baldwin and Ford (1988) and posited implications for practical consideration such as cognitive ability, self-efficacy, motivation, perceived utility of training, training design (behavioral modeling,

error management, realistic training environments) and the work environment (transfer climate, support, opportunity to perform, follow-up). Grossman and Salas argued that these factors have exhibited the strongest, most consistent relations with training transfer.

As we can see, the problem of training transfer is complicated, and there may not be an easy solution. Kraiger (2003) defined key areas that should be evaluated in efforts to solve the problem: (a) training needs assessment, (b) training design, and (c) training evaluation. For the military, the matter is complicated by two related factors. First, the military (at least the Navy) has adopted computer-based training (CBT) to varying extents because it has been perceived to be a more efficient learning platform than instructor-led classroom lectures (although now CBT is being integrated into a more effective blended training solution). Second, just as there are optimal person-job matches, there are optimal person-training matches. Some research has shown aptitude-treatment interactions with the treatment being either the occupation or the training (e.g., Statman, Gribben, Naughton, & McCloy, 1998).

Finally, the strength of the relation between training performance and job performance may be diminished if training emphasizes only knowledge, skills, or attitudes related to a subset of key performance dimensions. For example, training that is designed with the single objective of helping Sailors acquire technical skills may enable them to perform the technical aspects of their jobs. It may not, however, help them learn how to perform other important aspects of their roles, such as working together as a team or demonstrating personal discipline. As we have noted earlier, task and contextual performance are distinct; training designed to enhance one may have minimal impact on the other. But also, the Army's Project A has clearly established that there are "Can Do" factors that relate to technical aspects of the job as well as "Will Do" factors that relate to one's character and motivation to perform well on the job (Campbell, Hanson, & Oppler, 2001).

Improving training and training transfer is outside the scope of setting ASVAB standards. Nevertheless, having a clear understanding of training performance as the criterion variable will aid in interpreting not only the ASVAB validity coefficient, but also the linkage of training performance to job performance. Often a Navy ASVAB validation/standard review is triggered by complaints from the Fleet that Sailors reporting for duty do not have the requisite knowledge to perform the job. The question is whether the deficiency is due to the training itself or to the ability level of the Sailor to learn the material. Having a clear understanding about the training dynamics and aspects of the job being trained allows us to make credible recommendations in our ASVAB validation/standards studies. For example, a clearly needed training improvement may moderate the need to raise the ASVAB standard to such an extent that it affects the overall personnel classification system (e.g., the fill of the Rating).

Is "*g*" Enough for Prediction?

There has been some controversy about whether *g* is enough in predicting training and job performance. For example, it is probably not disputed that *g* is the best predictor (e.g., McHenry et al., 1990; Ree & Earles, 1991), but is *g* enough? Some research on cognitive measures shows that specific cognitive abilities do not add much

incremental validity beyond *g* in predicting performance (Ree et al., 2001). Carretta and Ree (2000) concluded that measures of specific cognitive ability typically increment the validity of *g* by no more than about .02 or .03. However, in large scale testing programs like the ASVAB, this level of incremental validity can be meaningful and result in large savings in training and other personnel related costs.

At times, specific abilities do contribute substantially beyond *g* in accounting for variance in training or job performance, and that conclusion is logical given there can be many training and job dimensions. For example, Olea and Ree (1994) examined the predictive validity of the Air Force Officer Qualifying Test (AFOQT) for several pilot and navigator training criteria. The AFOQT measures general cognitive, verbal, math, spatial, and perceptual speed, as well as aviation job knowledge (Carretta & Ree, 1996). Olea and Ree found that specific job-related knowledge (i.e., aviation knowledge) augmented the *g* in the AFOQT in predicting pilot training performance criteria by about .08. However, for navigator training, that job knowledge component of the AFOQT was not specific enough to provide incremental validity.

Thorndike (1985, 1986) concluded that specific tests can meaningfully improve the prediction of training and job performance when validation samples are large (i.e., $n > 200$) and when individuals already have been screened on *g*. Also, specific ability tests can better predict performance on tasks that are consistent and can be learned to the point of automaticity (i.e., do not require careful attention) as opposed to tasks that are inconsistent and require people to continually think about what they are doing and thus better predicted by *g* (Campbell & Kuncel, 2001).

Given the literature is generally positive about specific abilities, researchers and practitioners conducting ASVAB validation/standards studies are encouraged to align the content of ASVAB tests with the content of training when applying rational (rather than empirical) methods in developing ASVAB composites. Although additional research is needed, some of which is currently being addressed, it seems likely that tests of specific technical knowledge, like the military's Cyber test, will facilitate the acquisition of job-specific knowledge.

Concluding Remarks

This chapter provided an overview of predictors and criteria typically involved in personnel selection and classification systems and a framework for relating them as both observed and latent variables. We noted that training performance serves as the ASVAB validation criterion, not job performance, but that they are related. Awareness of the training context improves the design of validation studies by helping to (a) identify important training performance constructs, (b) make judgments about how well available training performance measures represent these constructs, and (c) identify and evaluate possible explanations for high student failure rates other than the ASVAB.

The following chapter expands the criterion space and addresses issues in measuring job performance, should we be looking past training performance to evaluate measures other than the ASVAB (e.g., non-cognitive measures such as personality).

Chapter 3. References

Alderton, D. L., Wolfe, J. H., & Larson, G. E. (1997). The ECAT battery. *Military Psychology, 9,* 5-37.

Anderson, L., Hoffman, R. R. III., Tate, B., Jenkins, J., Parish, C., Stachowski, A., & Dressel, J. D. (2011). *Assessment of Assembling Objects (AO) for improving predictive performance of the Armed Forces Qualification Test* (ARI-TR-1282). Arlington, VA: United States Army Research Institute.

Baldwin, T. T., & Ford, K. J. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology, 41*, 63-105.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478-494.

Borman, W. C., White, L. A., & Dorsey, D. W. (1995). Effects of ratee task performance and interpersonal factors on supervisor and peer performance ratings. *Journal of Applied Psychology*, *80*, 168-177.

Borman, W. C., & Motowidlo, S. M. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 71-98). San Francisco: Jossey-Bass.

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models of supervisory job performance ratings. *Journal of Applied Psychology, 76,* 863-872.

Campbell, J. P. (1990). Project A: The U.S. Army selection and classification project [Special issue]. *Personnel Psychology*, 43(2).

Campbell, J. P., Hanson, M. A., & Oppler, S. H. (2001). Modeling performance in a population of jobs. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 307-333). Mahwah, NJ: Erlbaum.

Campbell, J. P., & Knapp, D. J. (Eds.) (2001). *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Erlbaum.

Campbell, J. P. & Kuncel, N. R., (2001). Individual and team training. In N. Anderson, D. S. Ones, H. K. Sinangril, & C. Viswesvaran (Eds.), *Handbook of industrial, work, and organizational psychology* (pp. 287-312). London: Sage.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp.35-70). San Francisco: Jossey-Bass.

Carretta, T. R., & Ree, M. J. (1996). Factor structure of the Air Force Officer Qualifying Test: Analysis and comparison. *Military Psychology, 8*, 29-42.

Carretta, T. R. & Ree, M. J. (2000). General and specific cognitive and psychomotor abilities in personnel selection: The prediction of training and job performance. *International Journal of Selection and Assessment, 8,* 224-233.

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

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40,* 153-193.

Colquitt, J. A., LePine, J. A., & Noe, R. A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology, 85,* 678-707.

Drasgow, F. (2003). Intelligence and the workplace. In W. C. Borman, D. R. Ilgen, R. J. Klimoski (Eds.), *Handbook of psychology, Volume 12, Industrial and organizational psychology* (pp. 107–130). NY: Wiley.

Drasgow, F., Embretson, S. E., Kyllonen, P. C., & Schmitt, N. (2006). *Technical review of the Armed Services Vocational Aptitude Battery (ASVAB)* (FR-06-25). Alexandria, VA: Human Resources Research Organization.

Grossman, R., & Salas, E. (2011). The transfer of training: What really matters. *International Journal of Training and Development, 15,* 103-120.

Guilford, J. P. (1959). Three faces of intellect. *American Psychologist, 14,* 469-479.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Erlbaum.

Held, J. D., & Carretta, T. R. (2013). *Evaluation of tests of processing speed, spatial ability, and working memory for use in military occupational classification* (NPRST-TR-14-1). Millington, TN: Navy Personnel Research, Studies, and Technology.

Houston, J. S., Borman, W. C., Farmer, W. F., & Bearden, R. M. (Eds.) (2006). *Development of the Navy Computer Adaptive Personality Scales (NCAPS).* NPRST-TR-06-2). Millington, TN: Navy Personnel Research, Studies, and Technology.

Johnson, J. W. (2003). Toward a better understanding of the relationship between personality and individual job performance. In M. R. Barrick & A. M. Ryan (Eds.), *Personality and work: Reconsidering the role of personality in organizations* (pp. 83-120). San Francisco: Jossey-Bass.

Johnson, J. W., Duehr, E. E., Hezlett, S. A., Muros, J. P., & Ferstl, K. L. (2008). *Modeling the direct and indirect determinants of different types of individual job performance* (ARI Technical Report #1236). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Kirkpatrick, D. L. (1977). Evaluating training programs: Evidence vs. proof. *Training and Development Journal, 31*, 9-12.

Kirkpatrick, D. L. (1998). *Evaluating training programs: The four levels (2nd Edition).* San Francisco: Berrett-Koehler Publishers.

Kraiger, K. (2002). Decision-based evaluation. In K. Kraiger (Ed.) *Creating, implementing, and managing effective training and development* (pp. 331-375). San Francisco: Jossey-Bass.

Kraiger, K. (2003). Perspectives on training and development. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology, Volume 12, industrial and organizational psychology* (pp. 171-192). NY: Wiley.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology, 78*, 311-328.

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science, 315*, 1080-1081.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86,* 148-161.

Larson, G. E., & Saccuzzo, D. P. (1989). Cognitive correlates of general intelligence: Toward a process theory of g. *Intelligence, 13,* 5-31.

McCloy, R. A., Campbell, J. P., & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology*, *78*, 493-505.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43,* 335-354.

Motowidlo, S. J. (2003). Job performance. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology, Volume 12, Industrial and organizational psychology* (pp. 39–53). NY: Wiley.

Motowidlo, S.J., Borman, W.C., & Schmit, M. (1997). A theory of individual differences in task and contextual performance. *Human Performance*, *10*(2), 71-83.

Noe, R. A. & Colquitt, J. A. (2002). Planning for training impact: Principles of training effectiveness. In K. Kraiger (Ed.) *Creating, implementing, and managing effective training and development*. (pp. 53-79). San Francisco: Jossey-Bass.

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than *g*. *Journal of Applied Psychology*, *79*, 845-851.

Oppler, S. H., McCloy, R. A., Peterson, N. G., Russell, T. L., & Campbell, J. P. (2001). The prediction of multiple components of entry-level performance. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 349-388). Mahwah, NJ: Erlbaum.

Pulakos, E. D., Arad, S., Donovan, M., A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology, 85,* 612-624.

Ree, M. J., & Carretta, T. R. (1995). *Factor analysis of the ASVAB: Confirming a Vernon-like structure*. Interim Technical Paper, AL/HR-TP-1995-0007, U.S. Department of Defense, June 1995. Brooks Air Force Base, TX.

Ree, M. J., Carretta, T. R., & Steindl, J. R. (2001). Cognitive ability. In N. Anderson, D. S. Ones, H. K. Sinangril, & C. Viswesvaran (Eds.) *Handbook of industrial, work, and organizational psychology, Volume 1, personnel psychology*. (pp. 219-232). London: Sage.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71,* 432-439.

Schmitt, N., & Landy, F. J. (1993). The concept of validity. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 275-309). San Francisco: Jossey- Bass.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science, 58,* 1438-1457.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology, 15,* 201–293.

Spearman, C. (1927). *The abilities of man*. NY: Macmillan.

Statman, M. A L., Gribben, M. A., Naughton, J. A., & McCloy, R. A. (1998). *The development of a new research paradigm for studying aptitude-treatment interactions* (HumRRO FR-WATSD-97-25). Alexandria, VA: Human Resources Research Organization.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*.New York: Cambridge University Press.

Thorndike, R. L. (1985). The central role of general ability in prediction. *Multivariate Behavioral Research, 20,* 241-254.

Thorndike, R. L. (1986). The role of general ability in prediction. *Journal of Vocational Behavior, 29,* 332-339.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs, 1.*

Trippe, M. D., & Russell, T. L. (2011). *Information and communications technology literacy test norming study: Phase III final report* (AFCAPS-FR-2011-0011). Randolph AFB, TX: Air Force Personnel Center Strategic, Research and Assessment Branch.

Vineberg, R., & Joyner, J. N. (1982). *Prediction of job performance: Review of military studies* (NPRDC-82-37). San Diego, CA: Navy Personnel Research and Development Center.

Wigdor, A. K., & Green, B. F. (1991). *Performance assessment in the workplace (Vol. 1 & 2)*. Washington, DC: National Academy Press.

Chapter 4.
Issues in Predicting Job Performance
Rodney A. McCloy

Introduction

Personnel selection, classification, and training systems all intend to maximize job performance in some capacity. Selection systems seek to identify those job applicants who will perform best when hired. Classification systems seek to assign all of the hired individuals to jobs so as to maximize the organization's performance (but also to optimize the person-job fit). Training systems seek to prepare individuals to perform their jobs to the highest level possible. Each of these personnel systems requires the measurement of performance. This chapter takes the focus away from the ASVAB and training performance as the criterion and expands on the topic of job performance with particular emphasis on the nature of job performance as a construct and the difficulties organizations and researchers have had in measuring it.

The Nature of Job Performance: Three Ironies

In experimental psychology, the researcher seeks to establish direct control over and manipulate one or more *independent* variables and then evaluates the results of the experimental manipulation by examining changes in the outcome, or *dependent* variable. Outside the laboratory, where we lack experimental control, we tend to use the term *predictor* instead of "independent variable" for the measure included on the "*X*" side of the prediction equation. Similarly we tend to use the term *criterion* instead of "dependent variable" for the primary research outcome of interest, the measure included on the "*Y*" side of the prediction equation.

Without a criterion, there is no prediction equation (and thus no ASVAB validity coefficient in the case of ASVAB validation/standards studies). The criterion is a variable of primary interest in nay test validation/ standards study. However, whereas we have a lot of information about predictors in the literature, we have little information about the criterion. Ideally, we want to know the standing of the "subjects'" (applicants, trainees, employees) on the criterion. If we could observe their standing at the applicant level, we could use that information to select/hire/promote the "best" among them. However, the costs and/or danger of direct observation of criterion standing for a pool of applicants typically are too great. That is, practical constraints prohibit us from hiring/admitting everyone who applies for a job, observing their job performance, and then retaining satisfactory performers and firing those who perform in an unsatisfactory manner. Therefore, we go to great lengths to identify variables that *relate* strongly to the actual criterion so that we can use *hired* individuals' standings on these ("surrogate") variables to make our organizational decisions.

In the field of industrial-organizational (I-O) psychology, the criterion of greatest importance arguably is job performance. It is probably surprising, therefore, that I-O psychologists live with three great ironies regarding job performance and our understanding and measurement of it.

Irony #1: A Sparse Literature

The first irony is the surprising dearth of literature regarding job performance. Given the pre-eminence of performance in I-O psychology, one might expect an expansive sea of papers and books devoted to the topic. At least the relatively modest size of the literature has not resulted from a failure on the part of researchers to recognize the importance of studying the criterion when examining prediction systems. From nearly the beginning of efforts to quantify important I-O variables, I-O psychologists developed, validated, and proposed theories for all types of predictors while at the same time hearing numerous cries in the wilderness that researchers should devote a corresponding amount of time and energy developing, studying, and theorizing about criteria. Consider the following statements:

- "Much time and care should be expended on the collection of adequate variables. . . . Possibly as much time should be spent in devising the criterion as in constructing and perfecting the tests. This important part of a research seldom receives half the time or attention it requires or deserves. If the criterion is slighted the time spent on the tests is, by so much, largely wasted" (Toops, 1944, p. 290).
- "It should be axiomatic in any prediction study that the criterion is worthy of at least as much attention as the predictors" (Patterson, 1946, p. 277).
- "While everybody is concerned about the criterion, not very many people have spent very much time on it . . . instead of spending 95 per cent of our time going through a certain set of operations to get ourselves items and 5 per cent with the criterion, or with what those items are measuring, perhaps even the reverse percentage for a while might pay off in results" (Flanagan, 1948, p. 35).
- "Research can be no better than the criteria used. One must, therefore, approach the prediction process in a logical fashion, developing criteria first, analyzing them, and then constructing or selecting variables to predict the criteria" (Nagle, 1953, p. 273).

Recent efforts have been expended to model job performance (Campbell, 1990a; Campbell, McCloy, Oppler, & Sager, 1993; McCloy, Campbell, & Cudeck, 1994) and to introduce and explore criterion constructs such as contextual performance (Borman & Motowidlo, 1997; Motowidlo, Borman, & Schmit, 1997; Motowidlo & Van Scotter, 1994). These efforts have served as catalysts to increase published studies on the criterion; however, the literature remains unexpectedly manageable.

Irony #2: The Criterion Problem

The second irony is that one of our most vexing issues is the “criterion problem”—the fact that despite the criterion being the most important variable in the study, desirable criterion measures can prove quite difficult to obtain and generally require development. The advent of World War II forcefully drove this point home. Jenkins (1946) stated that psychologists were well equipped to predict a criterion, given that the criterion existed. However, rather than choosing criteria that adequately represented the

behaviors of interest, psychologists often chose variables that were the most convenient to obtain. Nagle (1953) stated that taking the criterion as a given "overcame the problem of criterion development by ignoring it" (p. 271). In a refreshingly honest appraisal of the state of affairs, Jenkins wrote that, before the war, "...psychologists in general tended to accept the tacit assumption that criteria were either given of God or just to be found lying about" (p. 93). The task of selecting men who would prove successful in combat, however, exposed the significant flaws in this assumption. There simply were *not* many criteria lying about. Criteria needed to be developed, and their adequacy (both extant criteria and manufactured) needed to be ascertained.

Even today we continue to bear the weight of the criterion problem. Given that the criterion is the variable of greatest importance in our predictive validation studies, the notion of obtaining any old measure that happens to be available and slapping a label of “job performance” on it should give one pause. Yet, as will be argued, it remains a relatively common practice. Aside from being sloppy science, failing to spend the time to develop appropriate performance criteria can have major real-world impact that might lead us to engage in counterproductive behaviors. In particular, the results of a validation study depend heavily on the particular criterion variable employed. Failure to pay due diligence in assessing the criterion could lead to putting in place selection measures that fail to provide the desired level of predictive power when applied in a setting that uses a different performance criterion.

The notion that validation results depend on the particular criterion variable chosen is not a new one. Jenkins (1946) provided an example involving the prediction of military aircraft gunners’ performance in World War II. Scores on an intelligence test correlated strongly with the performance criterion. Psychologists conducting the study, however, believed the criterion was too "intellectualized" (as might happen if the criterion were a job knowledge test). When the intellectual nature of the criterion was reduced (as might happen if the new criterion were a work sample test), the intelligence test no longer predicted criterion performance. Obviously the predictive validity of the intelligence test depended on the criterion measure used to validate it.

Weitz (1961) also discussed the importance of criteria to validity results but from an experimentalist's perspective, arguing that the criterion is another parameter to investigate when determining the impact of an independent variable. He illustrated that the outcomes obtained from an experiment investigating the effects of different types of verbal association on the mediation of transfer of learning (Cramer & Cofer, 1960) depended substantially on the criterion chosen. Although Weitz never mentioned a validity coefficient, his conclusions ring true for every psychologist conducting test validation research.

Wallace (1965) expressed grave concerns that still may apply regarding a major problem faced by all psychologists trying to predict the criterion of performance. He lamented that the empirical nature of much predictive research has failed to lead us to any new insights or developments regarding the variables incorporated in validation studies.

Wallace (1965) examined the progress of researchers who have attacked the prediction problem from a more theoretical viewpoint and concluded that they had often failed to find significant results and have no way to test the hypotheses they propound. That is, the criterion-related validity coefficient is not in and of itself meaningful. Echoing Weitz (1961), Wallace argued that

> "We have yet to pay enough attention to the processes by which we use our criteria to establish predictive relationships. In the rare cases, where we have examined different types of criteria for the same job and against the same predictors we have failed to arrive at interpretations of the different validity functions which emerge" (p. 414).

Wallace implored psychologists to use criteria that would be relevant not only to true job performance, but also to testing hypotheses about predictor-criterion relations. That is to say, he suggested determining the construct validity of our measures (predictors and criteria) so as to gain greater understanding of why a particular set of interrelations is observed.

James (1973) later gave an explicit statement that construct validation could/should be applied to criteria. Prior to his article, it seemed that construct validity was reserved for predictors - trying to understand the traits that each predictor assessed. Wallace (1965) was among the first to advocate the use of construct validation procedures with regard to criteria, but the James article showed how the procedures could be applied. James tested three types of models and recommended an integrated model and more elaborate construct validation procedures. The following is one of James' clearest statements on the topic:

> "While it could be argued that ascertaining what has been measured by a criterion is more a problem of content validity or operational definition, it is argued here that, because of the many possible sources of variance contributing to the measurement of a criterion, it is necessary to investigate the construct validity of criteria. The need to identify criterion constructs becomes crucial whenever contaminated measures, such as ratings, are employed, especially multiple ratings from different sources, or whenever operationally defined objective criteria (typically global) are not available . . . . Finally, an understanding of what constructs have been measured by criteria should greatly promote effective test construction in the sense that a conceptual framework attempting to overlap criterion and test constructs would replace much of the 'raw empiricism' still prevalent today" (p. 79).

Despite the illuminating thoughts that James (1973) provided us, the bulk of the literature that followed involving job performance measures still focused on predictors and the predictive relation rather than the criterion. Researchers still all too often seem willing to use whatever measures are available as their criterion and then turn their attention to the predictive validity of whatever predictors they have included in the study. Clearly, this practice can prove hazardous to our understanding of which predictors might prove useful in which situations.

Irony #3: A Confusing Literature

The third irony is that the literature regarding job performance has long been confused and, therefore, confusing. One might list several culprits for the muddle that constitutes the performance literature. Certainly one of the primary contributors is the sloppiness with which job performance has been operationally defined. Campbell et al. (1993) provided the following list of various operational definitions for job performance:

- "Time to complete a training course
- Grades or achievement test scores earned in training
- Number of errors made in a simulator
- Number of tinker toy figures assembled in a forty-five-minute experimental session
- Number of one-minute marketing interviews completed outside a shopping center in one day
- Number of defective pieces produced
- The total or average cost of the pieces produced
- Number of proposals written
- Total value of contracts won
- Total value of sales
- Number of grievances or complaints incurred
- Length of tenure in the organization
- Total days absent
- Salary level
- Promotion rate within an organization
- Percentage over budget
- Supervisor, peer, subordinate, or self ratings of 'overall' performance
- Scores on a paper-and-pencil job knowledge test
- Scores on a professional certification test
- Number of citations in the citation index over a three-year period
- Number of refereed journal articles published in a six-year period" (p. 36).

Labeling all such myriad criteria as "job performance" (almost certainly a result of the vaunted criterion problem) can serve only to cloud our discussion and understanding of what performance is.

The work of Campbell and his colleagues (Campbell, 1990a; Campbell et al., 1993; McCloy et al., 1994) has helped the field focus more clearly on a definition of what performance is and what it is not, as well as how we should conceptualize the performance construct and latent space. Their theory grew partly out of the substantial empirical research conducted during the Army's Project A described briefly in Chapter 3 (Campbell, 1990b; Campbell & Knapp, 2001). In brief, they define *performance* as behavior. Campbell et al. state that performance is

> "…something that people do and can be observed. By definition, it includes only those actions or behaviors that are relevant to the organization's goals and that can be scaled (measured) in terms of each individual's proficiency (that is, level of contribution). Performance is what the organization hires one to do, and do well. Performance is *not* the consequence or results of action, it is the action itself" (p. 40).

Effectiveness involves the evaluation of the results of an individual's performance. As a result, effectiveness is due to more than just the individual's actions/behavior. This definition highlights "the point that if the research questions deal with predictor validities, or training effects, or any other strategy focused on the individual, then the dependent variable should not be something that the individual cannot influence" (Campbell et al., 1993, p. 41). Finally, *productivity* is usually defined as the ratio of effectiveness (output) to the cost of obtaining said effectiveness (input).

In the 1980s and early 1990s, military personnel researchers undertook a major effort to define and measure job performance and to use those performance measures to inform a variety of military personnel issues (Wigdor & Green, 1991). Much of this work was conducted by Campbell and his colleagues and served as the catalyst for what was subsequently published as Campbell's theories of job performance measurement. As a result of the miscalibration of the military entrance test (the ASVAB), large numbers of young adults who normally would *not* have qualified for enlistment were accepted for service and yielded subpar training performance (Sellman & Valentine, 1981; Sims & Truss, 1980). One consequence of the ASVAB miscalibration is that it focused attention on the fact that the ASVAB had not been validated against job performance, but rather against a limited spectrum of criteria such as training grades, levels of indiscipline, and early attrition from service.

As a result of the ASVAB miscalibration, a 10-year research program was launched to refocus on job performance as the criterion upon which to validate the ASVAB. A substantial amount of effort and resources were applied to develop hands-on job performance measures and to relate those measures to enlistment standards. Ultimately, a mathematical model was developed that related recruit quality (as measured by enlistment test scores and educational achievement) to the surrogates of on-the-job performance, which could then be used to project the costs to recruit, train, and equip new recruits (Green & Mavor, 1994; McCloy et al., 1992). The effort enabled military personnel planners to build recruiting budgets that would allow the Services to attract sufficient numbers of high-ability youth who were high school graduates that could then satisfy a specified level of overall job performance.

A recent account of the details regarding the enlistment test miscalibration, job performance measurement research, and the development and validation of the recruit quality cost-effectiveness model can be found in Sellman, Born, Strickland, and Ross (2010).

One important feature of the performance theory developed by Campbell and his colleagues concerns their characterization of performance as inherently multidimensional. The notion of "overall job performance"—perhaps the most frequently adopted criterion measure—does not constitute a construct at all, being instead an unwieldy amalgamation of often moderately or poorly related dimensions. These dimensions should be examined separately if we are to maximize our understanding of an individual's performance and how best to predict or modify it. The theory specifies three determinants of individual differences in job performance that apply to all jobs, which are (a) declarative knowledge, (b) procedural knowledge and skill, and (c) motivation. The theory also specifies eight major performance dimensions (components of a job) that do not necessarily apply to all jobs depending upon the nature of the job. These eight performance dimensions are (a) job-specific task proficiency, (b) non-job-specific task proficiency, (c) written and oral communication, (d) demonstrating effort, (e) maintaining personal discipline, (f) facilitating peer and team performance, (g) supervisor/leadership, and (h) management/administration (Campbell, 1990a; Campbell et al., 1993).

Borman and Motowidlo (1993), as mentioned earlier and in the previous chapter, proposed expanding the job performance criterion domain to include behaviors that broadly support the organization at a level other than its technical core; that is, the dimensions of contextual performance that support the social and psychological well-being of the organization. The five contextual performance dimensions are (a) volunteering to carry out task activities that are not formally a part of the job, (b) persisting with extra enthusiasm when necessary, (c) helping and cooperating with others, (d) following organizational rules and procedures, and (e) endorsing, supporting, and defending organizational objectives.

Despite the call for clarity regarding the definition and measurement of job performance, the criterion problem still exists, as does the confusion regarding what performance is and is not.[1] The confusion has been particularly noteworthy in meta-analytic studies regarding the validity of various measures for predicting job performance (Oswald & McCloy, 2003).[2]

[1] Confusion regarding what constitutes performance also plagues other fields of study. Currently, there is much furor over how best to evaluate teachers, with many proponents arguing that student performance should serve as one of the indicators. Student performance is an indicator of teacher effectiveness, perhaps, but not of teacher performance, for much of a student's performance is not under the teacher's control.

[2] The military is considering personality measures as an adjunct to the ASVAB for classifying enlisted personnel to specific jobs, and so it would be appropriate to evaluate which of the many potential criteria are most important.

Multiple Criteria or a Single Criterion

One critical lesson from job performance modeling regards the dimensionality of performance—namely, that performance is multidimensional. This conceptualization of performance flies in the face of a notion that has had far too much influence—namely, Thorndike's (1949) notion of the *ultimate criterion*, which refers to "the concept of a single criterion measure that could reflect overall job success" (Catano, Wiesner, Hackett, & Methot, 2010, p. 190). The ultimate criterion proves so vexing for job performance researchers because it implies that performance is unitary (rather than multidimensional) and thus possibly could be represented with a single criterion measure (rather than several). Despite Thorndike's declaration that the ultimate criterion "is multiple and complex in almost every case" (p.121), the strong desire remains for a single performance criterion.

Thus, a specific performance criterion issue (unitary vs. multidimensional) has introduced a conflict between (a) those who trumpet the virtues of collecting multiple measures of performance, each connected to one of the multiple performance dimensions; and (b) those who advocate the need for a single performance composite. As cited previously, Wallace (1965) lamented that little had been done to that point with regard to understanding the various predictive relations that arose when a given job had multiple criterion measures. His views greatly resemble those of other researchers who argued for the use of multiple criteria in prediction research (e.g., Guion, 1961; Dunnette, 1963a, 1963b). With the multiple criterion approach, several criterion measures are obtained and used, each in turn, to validate a set of predictors. Dunnette (1963b) argued that amassing such validity evidence permits a much clearer explication of the meaning of the predictor scores than can be obtained when using a single criterion—the type of "plumbing the depths" that Wallace was seeking. For example, when multiple measurements are made, construct validation could be employed to attain a fuller understanding of the factors underlying the criterion variables (Inn, Hulin, & Tucker, 1972; James, 1973).

Schmidt and Kaplan (1971) pointed out that the multiple criterion perspective stands in contrast to the procedure advocated by those more concerned with practical (perhaps economic) issues. Researchers having this pragmatic bent have argued that, although multiple criteria may be obtained, they should be combined to form a single composite criterion. This composite may then be used to order individuals along a single dimension (if that were to be the case) upon which decisions about those individuals may be based. Some researchers have even asserted that the use of a single composite criterion was "indispensable" (Nagle, 1953; Toops, 1944). Perhaps the major difference between the multiple and composite criterion perspectives is that researchers espousing the use of a composite criterion do not concern themselves with the further understanding of behaviorally defined constructs, whereas researchers espousing the use and maintenance of multiple criteria believe such understanding to be a major goal of any validation study (Schmidt & Kaplan).

Ironically, neither the multiple nor the composite criterion approach as traditionally applied aids clarification of what underlies a particular predictive relation. If numerous criterion measures are obtained and then combined into a composite, and if that composite is then used to validate predictor measures, the researcher is left with a single criterion consisting of several criteria that are to some unknown degree deficient (i.e., fail to measure at least some portion of the construct of interest) and contaminated (i.e., measure at least some portions of other constructs not of interest). The advantage of obtaining multiple measurements is reduced and information lost unless we examine their intercorrelations. With intercorrelations, we can attempt to assess whether the multiple measures are measuring the same or different constructs.

It is difficult to assess the extent to which the observed validity coefficient arises from the prediction of relevant variance (i.e., variance that is due to the construct of interest; cf. Thorndike, 1949). In addition, the interpretation of the composite criterion is questionable if the criteria constituting it do not share a common underlying factor. Using a formative model (Podsakoff, MacKenzie, Podsakoff, & Lee, 2003) to combine criteria that do not contain a common factor may be correct statistically, but it could easily result in a psychologically nonsensical measure (Guion, 1961; Inn et al., Tucker, 1971). The interpretation difficulty arises because the resulting composite does not assess a construct but is instead a conglomeration of different constructs that might not relate highly with one another. As Campbell commented, "'Overall' performance is not a construct and cannot be *substantively* defined except in terms of mentioning all of its components in the same paragraph. The general factor found in covariance matrices is artifactual and results from method variance and common *determinants* (e.g. IQ, conscientiousness, etc.). Obviously, if there is a decision to be made about somebody (e.g. promotion), information must be combined, but the combinational rules are decision-specific" (John Campbell, personal communication, September 9, 2010).

On the other hand, if one obtains multiple criteria and keeps them separate for the validity analyses, then the number of indeterminably deficient and contaminated criteria increases *k*-fold; "the criterion problem" then becomes "the criteria problem" (Hakel, 1986). Although the multiple criteria approach has been advocated for its contribution to understanding the meaning of predictor test scores (Dunnette, 1963b), the procedure can provide such information only if one knows which factors determine variation in the criterion measures.

In applied research, the best solution might be to combine the two approaches. One would begin by collecting multiple measures of job performance, thus explicitly recognizing the multidimensional nature of the construct. Relations between the predictor set and these various criteria would be evaluated closely to determine which predictors were best for predicting which performance dimensions. At the end of the day, however, the applied research setting typically requires some type of decision be made: hire/do not hire, pass/fail, promote/do not promote. As a result, if multiple measures were collected, they would typically be combined in some way to yield a single composite criterion score that could support personnel decision-making.

Although it is difficult to interpret a single composite criterion that is inherently multidimensional, this is frequently done when conducting overall performance appraisals or any other administrative function that requires that a single decision be made about a single individual. We can think about co-workers in our office or our unit and readily identify who are the most highly regarded or top-performing employees. Perhaps we would identify these individuals as top performers because of their technical skills, sales skills, or interpersonal skills. However, to make such judgments about the general rank order of co-worker performance, we have to implicitly weight the various components that form the performance composite. If multiple variables are combined to yield a composite measure, the weighting must be made explicit. The question then becomes how to weight the components.

During the early time when many of the previously cited articles initially appeared in the scientific literature, researchers had access to relatively few resources that would allow them to simultaneously weight multiple predictors and criteria. Today, of course, the near ubiquity of sophisticated statistical analysis software presents researchers with numerous options for this task including simple to use drop down menus. There are both positive and negative aspects to widespread availability of easy to use software and the material in this chapter addresses some of the substance that must be considered when generating predictor/criterion models. The specific topic of weighting variables in these models is addressed in the Technical Manual.

The following summary points are offered on how to measure job performance, especially with regard to forming a composite meant to represent "overall" job performance:

- Define performance carefully and fully prior to measuring it. Create measures designed to assess important and relevant job behaviors.
- Recognize the multidimensional nature of performance by creating measures of each of the identified dimensions so that the full performance domain may be captured and understood (per Dunnette, 1963b) and the most likely successful predictors selected. Do not settle for convenient, available measures that might be either deficient or contaminated.
- If forming a single composite criterion meant to represent "overall" job performance, think carefully about how to weight the components that constitute the composite. Several methods are available that will allow the composite to reflect stakeholders' policy valuations.
- Keep in mind the distinction between nominal weights and effective weights (see the Technical Manual). Applying equal weights to a set of components will almost certainly result in unequal contributions of the components to the composite. In most instances involving rationally determined weights, alternate empirical weights need to be calculated and applied to the components to obtain effective weights that equal the desired nominal weights.

Concluding Remarks

Although the Navy has focused on training performance in ASVAB validation/ standards studies as the criterion variable, job performance remains an important outcome to measure and predict. However, job performance proves to be a complex, multidimensional target and we know that personnel psychologists have wrestled for decades with how best to conceptualize, define, and measure it (Wallace, 1965). Understanding the criteria of job performance for the military will be especially important in the future as new predictors are considered for inclusion in the ASVAB (or as adjunct classification instruments). The next chapter extends the discussion of performance measurement, but mainly to performance in Navy training with documentation of best practices.

Chapter 4. References

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 71-98). San Francisco, CA: Jossey-Bass.

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance*, *10*, 99-109.

Campbell, J. P. (1990a). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial & organizational psychology* (2nd Ed., Vol. 1) (pp. 687-732). Palo Alto, CA: Consulting Psychologists Press, Inc.

Campbell, J. P. (Ed.) (1990b). Project A: The U.S. Army selection and classification project [Special Issue]. *Personnel Psychology*, *43*, 231-378.

Campbell, J. P., & Knapp, D. J. (Eds.) (2001). *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Erlbaum.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35-70). San Francisco, CA: Jossey-Bass.

Catano, V. M., Wiesner, W. H., Hackett, R. D., & Methot, L. L. (2010). *Recruitment and selection in Canada* (4th ed.). Scarborough, ON: Nelson Education Ltd.

Cramer, P., & Cofer, C. N. (1960). The role of forward and reverse association in transfer of training. *American Psychologist*, *15*, 463. (Abstract).

Dunnette, M. D. (1963a). A modified model for test validation and selection research. *Journal of Applied Psychology*, *47*, 317-323.

Dunnette, M. D. (1963b). A note on *the* criterion. *Journal of Applied Psychology*, *47*, 251-254.

Flanagan, J. C. (1948). Discussion [Panel I]. *Proceedings of the 1948 invitational conference on testing problems: Validity, norms, and the verbal factor* (pp. 35-42). Princeton, NJ: Educational Testing Service.

Green, B. F., & Mavor, A. S. (Eds.) (1994). *Modeling cost and performance for military enlistment*. Washington, DC: National Academy Press.

Guion, R. M. (1961). Criterion measurement and personnel judgments. *Personnel Psychology*, *14*, 141-149.

Hakel, M. D. (1986). Personnel selection and placement. *Annual Review of Psychology*, *37*, 351-380.

Inn, A., Hulin, C. L., & Tucker, L. (1972). Three sources of criterion variance: Static dimensionality, dynamic dimensionality, and individual dimensionality. *Organizational Behavior and Human Performance*, *8*, 58-83.

James, L. R. (1973). Criterion models and construct validity for criteria. *Psychological Bulletin*, *80*, 75-83.

Jenkins, J. G. (1946). Validity for what? *Journal of Consulting Psychology*, *10*, 93-98.

Maier, M. H., & Truss, A. R. (1983). *Original scaling of ASVAB forms 5/6/7: What Went Wrong?* Research Contribution No. 457, Center for Naval Analyses, Alexandria, Va.

McCloy, R. A., Campbell, J. P., & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology*, *79*, 493-505.

McCloy, R. A., Harris, D. A., Barnes, J. D., Hogan, P. F., Smith D. A., Clifton, D., & Sola, M. (1992). *Accession quality, job performance, and cost: A cost-performance tradeoff model* (HumRRO Final Report FR-PRD-92-11). Alexandria, VA: Human Resources Research Organization.

Motowidlo, S. J., Borman, W. C., & Schmit, M. (1997). A theory of individual differences in task and contextual performance. *Human Performance*, *10*, 71-83.

Motowidlo, S. J., & Van Scotter, J. R. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology*, *79*(4), 475-480.

Nagle, B. F. (1953). Criterion development. *Personnel Psychology*, *6*, 271-289.

Oswald, F. L., & McCloy, R. A. (2003). Meta-analysis and the art of the average. In K.R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 311-338). Mahwah, NJ: Lawrence Erlbaum Associates.

Patterson, C. H. (1946). On the problem of the criterion in prediction studies. *Journal of Consulting Psychology*, *10*, 277-280.

Podsakoff, P. M., MacKenzie, S. B., Podsakoff, N. P., & Lee, J. Y. (2003). The mismeasure of man(agement) and its implications for leadership research. *The Leadership Quarterly*, *14*, 615-656.

Schmidt, F.L. & Kaplan, L.B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. *Personnel Psychology, 24,* 419-434.

Sellman, W. S., Born, D. H., Strickland, W. J., & Ross, J. J. (2010) Selection and classification in the U.S. military. In J.L. Farr & N.T. Tippins (Eds.), *Handbook of employee selection* (pp. 679-704). NY: Routledge - Taylor and Francis Group.

Sellman, W. S., & Valentine, L. D. (1981). *Aptitude testing, enlistment standards, and recruit quality*. Paper presented at the 89th Annual Convention of the American Psychological Association, Los Angeles, CA.

Sims, W. H., & Truss, A. R. (1980). *A reexamination of the normalization of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 6, 7, 6E, and 7E (CNS 1152)*. Alexandria, VA: Center for Naval Analysis, Marine Corps Operations Analysis Group.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement technique*. NY: Wiley.

Toops, H. A. (1944). The criterion. *Educational and Psychological Measurement*, *4*, 271-297.

Wallace, S. R. (1965). Criteria for what? *American Psychologist*, *20*, 411-417.

Weitz, J. (1961). Criteria for criteria. *American Psychologist*, *16*, 228-231.

Wigdor, A. K., & Green, B. F. (1991). *Performance assessment in the workplace (Vol. 1 & 2)*. Washington, DC: National Academy Press.

Chapter 5.
Navy Training and Best Practice Performance Measurement

Eduardo Salas and Sarah A. Hezlett

Introduction

The Navy currently sets ASVAB standards using training grades and training outcomes as the performance criteria. The Navy maintains a large organizational infrastructure to support training, including the development of tests to assess learning. This chapter provides an overview of the Navy's training, training challenges, and guidance for best practice performance measurement. The more that we know about any criterion variable and its measurement process, the more confidence we will have that the criterion problems discussed in the previous chapter are not obscuring real underlying predictor/criterion relations.

The Navy's Evolving Training

There are many inputs to the development of the Navy's training programs. For example, the Navy Education and Training Command (NETC) located in Pensacola, Florida, develops training policy. NETC also conducts the fundamental job duty task analysis (JDTA) workshops with the input of subject matter experts (SMEs) from the Fleet to establish the important job tasks that require recruit training. The JDTA information is used in part to develop the training course curriculum. The Navy Manpower Analysis Center (NAVMAC) located in Millington, Tennessee, also has a role in curriculum development by way of the occupational standards that they develop for each Navy Rating and each job category within the Rating. The training schools in turn provide SMEs, some of whom are instructors, for assisting in the development of the curriculum along with expert contractors. Contractors are integral to the development and maintenance of many of the Navy's training systems and curricula.

The overarching goal of the Navy's training programs is to establish a level of Sailor proficiency that is necessary to perform a job to some set standard (occupational standards set by NAVMAC). In fulfilling this goal, every Navy training course is developed with specific objectives. These objectives are categorized as course, enabling, and terminal objectives. Course objectives broadly define what knowledge and skills are required to perform the job and are developed through JDTA working groups. Higher level and lower level tasks (broad and narrowly defined) are used to guide curriculum development. Terminal objectives are developed to describe the standard that must be met to demonstrate that sufficient learning has taken place to perform a specific task. Enabling objectives are the individual steps, or building blocks, of declarative or procedural knowledge required to demonstrate a terminal objective.

Around the early 2000 time-frame, the Navy started to move largely from an instructor-led, group-paced method of delivering training with hands-on application that took place either in laboratories or field exercises, to self-paced computer-based training (CBT). The CBT training transition was part of the Navy's "Revolution in Training" that was projected to realize large cost savings because fewer instructors were required to actually teach students. Many of the instructors were Navy senior enlisted with important skill sets used in critical Navy occupations from which they were taken (to conduct training). Cost savings were also projected from CBT because many students would complete their courses early and move more quickly into a productive status. However, over time the savings realized from CBT were more than offset by the Navy's loss in capability of filling Fleet and Shore personnel requirements with fully proficient Sailors. Further, even if Sailors were proficient at the end of training, the Navy lost the capability to project when Sailors would report for their duty assignments because CBT at the time was largely self-paced.

Because CBT was self-paced (within a reasonable time limit), students who were very smart could more easily progress through the CBT modules, but this did not necessarily result in deep learning. Others struggled to digest the early CBT's one-dimensional, uninteresting formats that did not afford opportunities to apply what was learned. A combination of factors most likely resulted in the shortfalls of the Navy's initial CBT that included (a) insufficient funding to support effective CBT development and updates, (b) training policy gaps in specifying the roles and responsibilities for input to CBT system development, (c) a shortage of schoolhouse instructors to oversee students' progress (as many instructors were replaced by CBT without a plan for student oversight), and (d) an inadequacy in the CBT for engaging the Sailor in deep learning with the tendency for student to largely pass CBT courses through test-taking strategies.

The Navy has learned many lessons from their initial CBT instructional materials and is now fully engaged in core initiatives to improve training. One of these improvement initiatives is for Navy schools to return to instructor-led, group-paced instruction, but within a blended solution that incorporates (to varying degrees) best practice CBT. The CBT platforms are now used merely as learning resource adjuncts in many schools (see Singh, 2003 for a concept of effective blended solutions).

Pros and Cons of CBT Environments

Bedwell and Salas (2010) reviewed CBT and considered the pros and cons. In their opinion, the most positive aspects of CBT are (a) the standardized presentation of material that cannot occur when different instructors lead classes, (b) accessibility in remote geographical locations, including ships, (c) cost savings because a large number of instructors is no longer required to teach students (instructor to student ratio), and (d) dynamic, engaging presentation of technical material that cannot occur in textbooks.

The most negative aspects of poorly conceived or developed CBT are (a) the course content can quickly become outdated, and the updating process can be so expensive that it negates the savings that were forecasted from the elimination of instructors, (b) updating to replace outdated or incorrect information can be time-consuming, so students may be exposed to incorrect information for an inordinate amount of time, (c) an uninteresting or inauthentic learning environment can result in students not relating

to the content and therefore becoming inattentive (Rosenberg, 2001), (d) learning can become superficial and retained only long enough to pass a test, and (e) students may not have learned enough to be prepared to perform their jobs.

In developing its blended solutions, the Navy has strived to reintroduce realistic and engaging learning environments (with some schools reintroducing hands-on learning) that are augmented by effective CBT. Below are some practical considerations for assessing CBT when visiting Navy schoolhouses (distilled from Bedwell & Salas, 2010):

- CBT is becoming a more sophisticated learning tool, socially intelligent, and agent-guided.
- CBT is not the answer to all training needs, but it has a place in training.
- Blended learning with CBT as a component is a popular training solution but may not always lead to increased learning.
- CBT will not be successful unless the purpose is clearly defined prior to design and development.
- Although the purpose drives the selection of the most appropriate CBT, the selection should consider the metrics that will be used for CBT effectiveness.
- As with any training, CBT success depends heavily on the elements of good training.
- CBT developers should not try to reinvent the wheel, but rely on what works. Years of CBT research, including multimedia design, have produced useful principles.
- CBT environments still require instructors.
- CBT that is strictly preprogrammed cannot adapt to the learner's needs – that is, provide intelligent tutoring capability.
- CBT will require maintenance and content updates, so CBT is not free of costs.

Performance Measurement in Classroom/CBT Environments

Having presented some Navy history and information about CBT, we now turn to the ASVAB validation/standards researcher's interest in CBT performance measurement. We refer back to Figure 3-1 for linkage to our predictor criterion relations. First, we note that measures used to assess knowledge acquisition during self-paced CBT may have low reliability if they involve change over time. The low reliability occurs mainly because of the self-paced nature of CBT. That is, individuals differ in how long it takes to learn the CBT material, and if performance is developed as a composite measure, it will include not only the learning outcome, but also the time to achieve the outcome. These composite scores—known by various names, including change, growth, difference, or slope scores—have an inherently lower level of reliability than the raw performance data (Carter, Krause, & Haberson, 1986).

Because ASVAB validation/standards studies have an operational focus, it is not necessary to correct the CBT performance criterion for unreliability. That is, the operational decision to academically fail a student is based on the observed performance scores, not the perfectly reliable "true" score. However, if we can obtain sufficient reliability evidence and it appears to be low, then we will want to inform the training command about the deficiency. For research purposes, say in estimating the validity of a new cognitive or non-cognitive test, we will want to correct the ASVAB validity for unreliability in the criterion if only for estimating relation "e" in Figure 3-1. Correcting for criterion unreliability is addressed in the Technical Manual.

We should be aware that models of skill acquisition suggest that gaining knowledge is a critical but insufficient step towards becoming a skilled performer (Kraiger, Ford, & Salas, 1993). The Navy Fleet would like to receive already skilled Sailors directly from training, but training is just a first step in the continuum of a Sailor's learning. Furthermore, just because individuals acquire specific levels of knowledge in training does not mean they will be able to "enact" that knowledge, either in the training itself, or in transferring that knowledge to the job (as briefly discussed in Chapter 3). Taxonomies of learning outcomes recognize knowledge and skills as distinct constructs (Campbell & Kuncel, 2001; Kraiger et al., 1993). A large literature exists on the problem of transferring what is learned in training to application on the job and this literature comprises various perspectives (system design, organizational environment, individual differences). We refer the reader to just a few of the widely available sources (Gist, Bavetta, & Stevens, 1990; Holton & Baldwin, 2003; Kraiger, Salas, & Cannon-Bowers, 1995; Royer, 1979; Tracy, Tannenbaum, & Kavanagh, 1995).

Training Success

In receiving the performance measures, the ASVAB validation/standards team will want to understand if they were developed in such a way as to have a role in improving training effectiveness (Goldstein & Ford, 2002). We should ask questions about the performance measures and their capacity for the following:

- evaluating the effectiveness of the training program so that it can be improved,
- providing a platform for practice so that the students can instantiate learning,
- providing diagnostic feedback to the students so that they take corrective actions, and
- providing diagnostic feedback to the instructors so that they can identify students who require remediation.

The Navy training community considers the above goals in their development of meaningful training performance measures; however, the ASVAB validation/standards researchers should follow through with these and many other questions about the integrity of the measures as they serve the basis for establishing the ASVAB's predictive validity. This chapter and Chapter 3 will provide a solid basis for the criterion variable's evaluation in the training context.

Simulation Environments

Simulation environments are realistic teaching tools, and as such, they act as a bridge between knowledge and procedural learning, demonstration of learning, and application of learning on the job (training transfer). Simulation-based training (SBT) has always had a role in military training (e.g., flight simulators and constructed war games). The realism dynamic depends upon the job context, job criticality, and cost effectiveness of the technology. For example, the Navy's Air Traffic Control School in Pensacola has a complete mock-up of an air traffic control tower where different flight scenarios and problems are played out for students to resolve. Obviously, Air Traffic Controller is a critical job, and the investment in SBT is well justified. In contrast, the Apprentice Technical Training (ATT) at Great Lakes has a CBT system that simulates different electronic failure scenarios for students to diagnose and repair. Both simulation environments appear to be appropriately developed for the schools' contexts.

SBT can be viewed as a general method for providing systematic and structured learning experiences and, as such, is a viable means of performance measurement (Salas, Rosen, Held, & Weismuller, 2009). In fact, one dimension of SBT effectiveness is the quality of the practices of developing SBT performance measurement. The importance of performance measurement, however, may depend upon the nature of the job tasks that SBT is intended to train. For example, the organization may consider SBT as a platform for merely providing practice opportunities under the assumption that all individuals will eventually become proficient in the task on their own. In this case, performance measurement may not be a priority. Or SBT may be intended to provide diagnosis, feedback, and guidance to ensure the trainee takes the correct path to proficiency or actually attains proficiency, in which case, providing diagnosis and feedback requires performance measurement.

The following points should be considered when trying to establish the integrity of SBT and performance measures:

- It might be difficult to construct exercises or scenarios that capture the relevant job-related behaviors.
- Performance measurement may not be automated in any fashion so that the instructor may be more involved in logging results than paying attention to the performance behavior.
- Unlike knowledge assessments taken by standardized tests, instructors are the mechanism for making laboratory or field practical assessments, and they all might not perceive an individual's performance the same way.

In general, the instrument used to evaluate laboratory or field exercise performance is a checklist. Checklists are useful for grading purposes and generating feedback, but they are often ambiguously related to psychological constructs. Checklists also are designed to guide students to the completion of simulations, not to differentiate their performance. Further, rating officials (usually instructors) can introduce bias in the checklist procedure. Problems of bias in observer ratings are well documented and pose a threat to the measurement of training performance (Goldstein & Ford, 2002).

Instructors, as performance observers, are sensors that detect and capture performance (Bakeman & Gottman, 1997). Any sensor, whether human or machine, must be initially calibrated, and intermittently recalibrated, so that its readings are always accurate. Part of the calibration process involves training observers to reliably rate performance (see Baker & Dismukes, 2003; Goldsmith & Johnson, 2002; Holt, Hansberger, & Boehm-Davis, 2002; Mulqueen, Baker, & Dismukes, 2002). As mentioned earlier, the Technical Manual provides a full discussion of the psychometrics of unreliability and how it influences the validity coefficient.

Measures of performance in "simulated" settings such as laboratory demonstrations and field exercises can act as surrogate measures of some dimensions of job performance. Surrogate measures are indirect measures of performance that are related to and predictive of on-the-job performance. Surrogate measures sacrifice "face validity" that perhaps sponsors of research would want to see, but they add value in important ways (Kennedy, Lane, & Kuntz, 1987), mainly by providing higher reliability of the performance measures and lower developmental costs. We note that optimal (maximal) sustained performance is usually observed in the training environment or where the individual perceives high stakes outcomes (e.g., promotion) and so the reliability of the surrogate measure may be higher than the actual job performance measure, where we would generally observe actual performance.

Surrogate job performance measures have five important characteristics: (a) stability over time, (b) high correlation with the performance construct of interest, (c) sensitivity to the same factors as those affecting performance, (d) a higher degree of reliability than the direct measure of performance, and (e) reduced training time requirements (Kennedy et al., 1987; Lane, 1986).

Not all available performance related data have the requisite characteristics to be surrogates. For example, promotion to higher rank exists in Navy databases, but it is not a suitable criterion for ASVAB validation because promotion is constrained to few individuals, may be contaminated by a political process, and is essentially a dichotomous variable (although having a dichotomous variable as the criterion in not a fatal flaw, as we will see in Chapter 12 of the Technical Manual). The next section provides a discussion about the challenges involved regarding obtaining meaningful criterion measures.

General Navy Training Performance Measurement Challenges

A challenge for the ASVAB validation/standards team during their Navy school visits will be how to consider intermediate measures of performance such as remediation and retests, setbacks that cause a student to restart in a later class, student action boards (SAB), and after action reviews (AARs). These data only exist at the schoolhouse and only for a most recent two-year period. The final school grade (FSG) from a course of training is maintained for most schools in the Navy's corporate training database (explained in the next chapter), but FSG does not tell the whole story. The FSG variable is the performance measure used to develop the ASVAB composites' validity coefficients, but only some schools factor in retests into that measure.

One must consider that a student who has experienced three setbacks but eventually passed the course had more opportunities to learn the material than the student who just breezed through the course. All other things being equal, the student who had more exposure to the training material should be more prepared to perform their jobs. On the other hand, the Navy conducts training in a limited time-frame so as to realize a reasonable time of productive status for a Sailor's first term of enlistment. Students who are setback or retest multiple times impinge on that training time efficiency model. It is up to the ASVAB researchers to understand the curriculum difficulty as it relates to time allowed to train, observed student challenges, and how the school's performance measurement process takes challenges into account. We note that the one component of the Navy's algorithm for Rating classification considers the ASVAB's validity in predicting first pass pipeline (training) success (explained in Chapter 18 of the Technical Manual), meaning, without a setback incident.

Another challenge is how to deal with performance measures that are graded only as pass or fail (i.e., dichotomous outcome variables). Schools often allow students to repeat a performance-based task several times, considering each attempt as practice (until proficiency). In fact, feedback is instrumental in training so that students can obtain guidance in both the behaviors and strategies required to perform the task correctly. There is little to no variance in the criterion if most students eventually perform the task correctly, making it impossible to establish a meaningful validity coefficient (Sackett, Lievens, Berry, & Landers, 2007; Sackett & Yang, 2000; Shadish, Cook, & Campbell, 2002). What is optimal for the student and the Navy (training until the task is mastered) is suboptimal for the empirical analysis of the relation between the ASVAB and performance measure.

As mentioned before, in dealing with performance measurement challenges the ASVAB validation/ standards team should recognize that performance demonstrated at a school, particularly in simulations, might reflect students' maximum level of performance, not their typical level of performance that might be demonstrated later in the job. This distinction is important because it has been shown that individuals' maximum performance correlates only modestly with their typical performance (DuBois, Sackett, Zedeck, & Fogli, 1993; Sackett, Zedeck, & Fogli, 1988). Generally, students in training are highly motivated to succeed because failing produces a stigma and also setbacks in their careers. The Navy's validation research involving personality measures is not yet mature enough to address typical and maximum performance differences in training or on the job.

Performance measurement problems that may reduce the magnitude of the ASVAB validity coefficient are always a concern, and include the following:

- the performance measure inadequately differentiates individuals because it is too easy or too hard;
- the performance measure is a subject mastery test, and a threshold score is all that is required to establish when progression to another subject is to occur; and
- the performance measure is too short to measure the performance outcomes reliably.

Guidelines for Best Practice Performance Measurement in SBT

Salas et al. (2009) provided guidelines for developing best practice training performance measurement in SBT. Although developed for SBT, many of the guidelines also apply to classroom training that is knowledge-based and so the guidelines as stated below should assist the ASVAB validation/standards team in assessing the training performance criterion.

1. Develop measurable learning outcomes.
2. Know the behavioral, attitudinal, and cognitive competencies necessary for performance.
3. Derive a set of specific metrics for each performance objective.
4. Develop behavioral markers of performance for each learning outcome.
5. Develop metrics that are diagnostic of performance.
6. Use multiple data sources and types to capture performance.
7. Capture performance at multiple levels.
8. Create a plan for integrating multiple sources and types of measurement.
9. Automate as much of the performance measurement collection and analysis as possible.
10. Develop and implement training programs for observers/instructors.
11. Provide structured tools/protocols for observations.
12. Use a checklist for observations that link discrete behaviors to scripted events.
13. Focus performance measurement on discrete, observable behaviors.
14. Create and maintain a systematic, organized representation of performance.
15. Do not over-burden observers; maintain a good ratio of observers to trainees.

Salas et al. (2004) also reviewed the theoretical foundations underpinning performance measurement systems in SBT and the various methods historically and currently used for measuring performance. The review also provided example applications of qualitative and quantitative methods of measuring performance. Finally, Salas et al. provided four categories to group 21 best practices as most effective for the specified purpose of the measurement system, described here as to (a) obtain multiple levels of measurement, (b) know about the process as well as the performance outcome, (c) describe, evaluate, and diagnose the performance, and (d) inform about the requirement for remediation (Table 3, p. 361).

With regard to the second listed guideline, “Know the behavioral, attitudinal, and cognitive competencies necessary for performance”, we can assume that attitudes, and therefore personality or temperament (especially motivation), may be important predictors of Sailors’ training performance for some occupations (possibly moderated by learning styles and types of training). Both the Navy and Army have developed computer adaptive personality/temperament instruments (with somewhat different

algorithms and objectives). The Navy has the Navy Computerized Adaptive Personality Scales (NCAPS) (Houston, Borman, Farmer, & Bearden, 2006) and the Army has the Tailored Adaptive Personality Assessment System (TAPAS) (Drasgow et al., 2012). The Army is using TAPAS operationally in some of their applicant screening programs, and for limited MOSs classification. The Army considers a spectrum of criterion measures that are close in time to enlistment to further out in a Soldiers career. The Navy uses NCAPS for only a select few Ratings and not at all for applicant enlistment screening. NCAPS is an ongoing research project and it is yet to be determined if personality measures would be applied globally in Rating classification.

Concluding Remarks

The Navy has undergone several transformations in training and the latest blended platforms should greatly improve Sailors' readiness to perform their first duty jobs. Part of this transformation involves establishing for each Navy Rating the optimal balance of traditional instructor-led classroom training, effective computer-based training, and hands-on laboratory or field exercises. The ASVAB validation/standards team will encounter various mixes of training techniques and platforms when conducting school visits; this chapter was intended to provide assessment guidelines. The most important observation to be made during school visits is not how sophisticated or flashy the training systems are, but whether they follow good training principles and were developed to produce good performance measures. The next chapter describes the various steps for conducting an ASVAB validation/standards study, including the protocol for conducting a Navy school visit and performance data collection.

Chapter 5. References

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge, UK: Cambridge University Press.

Baker, D. P., & Dismukes, R.K. (2003). *A gold standards approach to training instructors to evaluate crew performance* (No. NASA/TM--2003--212809). Moffett Field, CA: National Aeronautics and Space Administration.

Bedwell, W. L., & Salas, E. (2010). Computer-based training: Capitalizing on lessons learned. *International Journal of Training and Development, 14*, 239-249.

Campbell, J. P., & Kuncel, N. R. (2001). Individual and team training. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of work and organizational psychology*. London: Blackwell.

Carter, R. C., Krause, M., & Haberson, M. M. (1986). Beware the reliability of slope scores for individuals. *Human factors, 28*, 673-683.

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support Army selection and classification*. (ARI Technical Report 1311). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

DuBois, C. L. Z., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and white-black differences. *Journal of Applied Psychology, 78*, 205-211.

Gist, M. E., Bavetta, A. G., & Stevens, C. K. (1990). Transfer training method: Its influence on skill generalization, skill repetition, and performance level. *Personnel Psychology*, *43*, 501-523.

Goldstein, I., & Ford, J. K. (2002). *Training in organizations* (4th ed.). Pacific Grove, CA: Wadsworth Thomson Learning.

Goldsmith, T. E., & Johnson, P. J. (2002). Assessing and improving evaluation of aircrew performance. *The International Journal of Aviation Psychology, 12*, 223-240.

Holt, R. W., Hansberger, J. T., & Boehm-Davis, D. A. (2002). Improving rater calibration in aviation: A case study. *International Journal of Aviation Psychology, 12*, 305-330.

Holton, E. F. & Baldwin, T. T. (2003). Making transfer happen: An action perspective on learning transfer system. In E. F. Holton & T. T. Baldwin (Eds.), *Improving learning transfer in organizations* (pp. 3-15). San Francisco: Jossey-Bass.

Houston, J. S., Borman, W. C., Farmer, W. F., & Bearden, R. M. (Eds.) (2006). *Development of the Navy Computer Adaptive Personality Scales (NCAPS)*. NPRST-TR-06-2). Millington, TN: Navy Personnel Research, Studies, and Technology.

Kennedy, R. S., Lane, N. E., & Kuntz, L. A. (1987). Surrogate measures: A proposed alternative in human factors assessment of operational measures of performance. *Proceedings of the 1st Annual Workshop on Space Operations, Automation, & Robotics* (pp. 551-558). Houston, TX: NASA Lyndon B. Johnson Space Center.

Kraiger, K., Salas, E., Cannon-Bowers, J. A. (1995). Measuring knowledge organization as a method for assessing learning during training. *Human Factors, 37,* 804-816.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology, 78*, 311-328.

Lane, N. E. (1986). *Issues in performance measurement for military aviation with applications to air combat maneuvering* (No. NTSC TR-86-008). Orlando, FL: Naval Training Systems Center & U.S. Army Research Office.

Mulqueen, C., Baker, D. P., & Dismukes, R. K. (2002). Pilot instructor rating training: The utility of the multifacet item response theory model. *International Journal of Aviation Psychology, 12*, 287-303.

Rosenberg, M. J. (2001). *E-Learning: Strategies for delivering knowledge in the digital age*. New York, NY: McGraw-Hill.

Royer, J. M. (1979) Theories of the transfer of learning. *Educational Psychologist*, *14*, 53-69.

Tracy, J. B., Tannenbaum, S. I., & Kavanagh, M. J. (1995). Applying trained skills on the job: The importance of the work environment. *Journal of Applied Psychology, 80*, 239-252.

Sackett, P. R., Lievens, F., Berry, C. M. & Landers, R. N. (2007). A cautionary note on range restriction and predictor intercorrelations. *Journal of Applied Psychology, 92,* 538-544

Sackett, P. R. & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*, 112-118.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73,* 482-486.

Salas, E., Rosen, M. A., Weaver, S. J., Held, J. D., & Weissmuller, J. J. (2009). Guidelines for performance measurement in simulation-based training. *Ergonomics in Design, 19,* 12-14.

Shadis, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Singh, H. (2003). Building effective blended learning programs. *Educational Technology, 43,* 51-54.

Chapter 6.
The Navy's ASVAB Validation/Standards Process
Janet D. Held

Introduction

A primary goal of the Navy is to ensure that enlisted personnel are assigned to technical training for which they exhibit a high likelihood of success. Policy makers recognize, however, that there will always be some level of recruit failure rates in training because (a) training is time-constrained so that Sailors will have a certain degree of productive status during their first term of enlistment, (b) the ASVAB is not a perfect predictor of training success, and (c) recruit ASVAB scores as a population can vary in different recruiting environments. The Navy's ASVAB Validation/Standards program has been institutionalized so that there can be a continuous monitoring process of all Ratings' ASVAB standards. In general, the Navy's approach in the past was to revise ASVAB standards as needed; however, a broader but more structured schedule starts in fiscal year 2015. Many factors will determine the prioritization of Ratings that require an ASVAB validation/standards study. This chapter describes these factors and the high level steps that are taken in conducting a study; the Technical Manual provides the details.

The Strategy for Conducting ASVAB Validation/Standards Studies

Any one of a multitude of factors described in this chapter can trigger the requirement for an ASVAB validation/standards study for a particular Navy Rating. This does not mean, however, that a particular study is conducted in a vacuum. The Navy recognizes that raising an ASVAB standard for one Rating can have a negative impact on the availability of recruit talent for all of the other Ratings, and so every effort is made to assess a "System of Ratings" impact.

Besides conducting individual Rating studies, the Navy conducts some ASVAB validation/standard studies for occupational groups of Ratings (usually fewer than 10 Ratings). It some cases it makes sense to consider the same ASVAB standards for Ratings that have an overlap in major job duties, require similar training, and are managed within the same community (e.g., Surface, Air, and Submarine). One reason is that it is administratively easier for Sailors to cross Ratings if required to do so (e.g., in a Navy drawdown). Another reason is that it allows the Navy to have some flexibility in determining where and when to make Rating assignments. For example, some applicants sign contracts for a "Program," or "Occupational Specialty", which involves two or three Ratings within the same community and with the same ASVAB standard (and similar occupational standards). Actual Rating assignments are made at the Navy's Recruit Training Center (RTC) upon arrival, or during initial core training when the Navy has greater clarity on the immediate community needs and school seat availability.

The ASVAB Validation/Standards Framework

Setting or revising ASVAB standards for Navy Ratings is theoretically a continuing dynamic process that is linked with other Navy functions and components, including occupational standards development and training development processes. Figure 6-1 shows how the ASVAB validation/standards studies align with these processes.

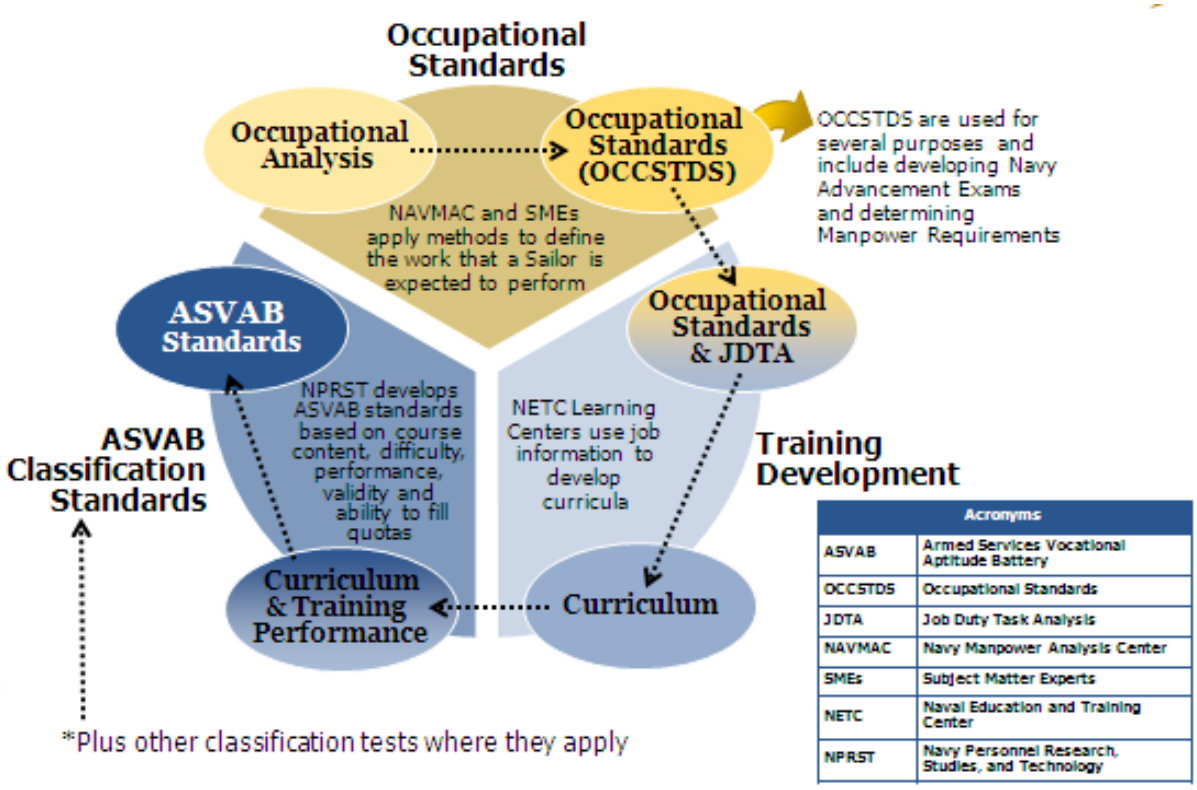

Figure 6-1. Relations between an ASVAB validation/standards study and other Navy components and processes.

Figure 6-1 shows that conducting an ASVAB validation/standards study is a third link preceded by the processes of occupational standards development and training/curriculum development. The occupational standards for each Rating are developed by the Navy Manpower Analysis Center (NAVMAC) and the training, curriculum, and performance measures are developed by the Naval Education and Training Command (NETC) and the Ratings' Learning Centers.

Each year, the Navy Personnel Research, Studies, and Technology (NPRST) ASVAB validation/standards team, the program sponsor (N132G - the Navy Selection & Classification Office), and the Navy Ratings' Enlisted Community Mangers (ECMs) meet to discuss the various issues experienced by each Rating, such as high schoolhouse failure rates. Besides this fact-finding or information-gathering method, an empirical-based method is applied to create a decision aid tool that provides input to the rank

ordering of Ratings that are candidate for an ASVAB review. A part of the decision aide tool involves a matrix that contains the following factors across Ratings, most of which can be scored numerically to provide a relative ranking of Rating complexity: (a) length of time to complete training, (b) number of skills and abilities, (c) number of times each of these skills and abilities is required to perform each subset of duties at an entry-level requirement, (d) the percentages of the most recent recruit population meeting the Ratings' ASVAB standards (qualification rate) (Johns, 2011). Other inputs to the priority ranking are the last time an ASVAB study was conducted and the estimated ASVAB validity coefficient for the operational composite.

Initiating a Navy ASVAB Study

Studies typically have been conducted in the past at the request of a Rating's ECM and after a subjective evaluation of criticality, with the process now being more structured and objective. An ECM manages a "community" of Ratings that are interrelated by occupational or mission similarities. For example, the following Ratings belong to the Construction Battalion community: Construction Electricians, Construction Mechanics, Steelworkers, Builders, Engineering Aids, Utilitiesman, and Equipment Operators. Each ECM is responsible for ensuring the health of their community of Ratings and therefore continually monitors Sailors' training success, job performance, retention, and promotion potential. The ECMs consider an ASVAB validation/standards study to be an important tool in managing the competing goals of quality and quantity, most commonly referred to as Rating "Fit and Fill".

An ECM may request an ASVAB validation study whenever a school's failure rate becomes unacceptably high or a significant number of school seats go unfilled. Other important reasons to conduct a validation study include major curriculum changes due to the merger of two or more Ratings, as well as changes in training time or mode of curriculum delivery (e.g., computer-based training replaces instructor-led training). These changes may affect the skills and abilities needed to perform well in training and, consequently, may require an ASVAB standards adjustment. The initiating factors are listed below - each single factor can trigger an ASVAB validation/standards study.

- High academically related failure rates or setback rates.
- A merger of two or more ratings, or a new rating.
- A significant change in the curriculum.
- A change in training time or training delivery methods.
- The inability to fill school seats.
- An observation by the schoolhouse that the ASVAB validity is too low.
- The requirement to adjust the ASVAB standards in a difficult or exceptional recruiting market.
- To provide common ASVAB standards for occupational groups where warranted.

An ECM initiates an ASVAB validation study by directly contacting the Director of Navy Selection and Classification (N132G - Washington, DC) or the ASVAB validation study team at NPRST, a branch of the Bureau of Naval Personnel (BUPERS) co-located with the ECMs at the Navy Personnel Command (NPC) in Millington, TN. The ECMs are familiar with the ASVAB Validation/Standards program because NPRST provides a tutorial type of briefing for them at the beginning of each fiscal year. However, even though the studies are decided upon early in a fiscal year, the ECMs can contact the team to discuss evolving ASVAB-related issues any time during the year. When a study requirement becomes apparent, the ASVAB team provides the ECM with information that will help in understanding the study fundamentals during the fact-finding meeting.

The Navy's ASVAB Classification Composites

Table 6-1 lists the Navy's classification composites (Table 2-1, the ASVAB tests).

Table 6-1
Navy Operational Selection and Classification Composites

Composite Name	Composite Tests[a]
Armed Forces Qualification Test[b]	2VE[c] + AR + MK
General Technical	VE + AR
Administrative	VE + MK
Hospitalman	VE + MK + GS
Electronics	AR + MK + EI + GS
Basic Electricity & Electronics	AR + 2MK + GS
Nuclear Field	VE + AR + MK + MC
Engineering	VE + AR + MK + AS
EOD/SEAL	GS + MC + EI
Mechanical	AR + MC + AS
Mechanical_2	MK + AS + AO
Operations	VE + AR + MK + AO
Air Traffic Control	VE + MK + MC + CS[d]
Business/Clerical	VE + MK + CS[d]

[a]ASVAB composites are integer weighted sums of the subtest standard scores. [b]AFQT, percentile scored, is used to qualify applicants for military enlistment. [c]VE is formed from 2/3 WK plus 1/3 PC (transformed to subtest standard scores). [d]CS (Coding Speed), a former ASVAB subtest, is now a Navy special classification test.

Two of the Navy's classification composites in Table 6-1 include the former ASVAB Coding Speed (CS) test (now a Navy special classification test); two other composites include the Assembling Objects (AO) test. The two tests are not administered to *all* Navy applicants (see Chapter 2) and so an alternative "ASVAB only" composite must be in place for a Rating when either of these tests is part of the operational ASVAB composite.

The Navy classification composites listed in Table 6-1 are the result of small and large scale ASVAB validation/ standards studies that were conducted over many years. The last major large-scale ASVAB validation project was conducted by Navy Personnel Research and Development Center (NPRDC, the former NPRST) in collaboration with Personnel Decisions Research Institute (PDRI) (Hedge, Carter, Borman, Monzon, & Foley, 1992). The task at the time was to address the apparent ASVAB composite overlap/redundancies and the potential for other ASVAB candidate composites to become operational based upon higher predictive validity levels. More than 70 Navy Ratings were a part of the study and several recommendations were implemented after conducting further single Rating ASVAB validation/standards studies to see if the large-scale study results held up. Over time, some ASVAB composites were added and some removed, but only when the Navy Rating specific ASVAB validation/standards study supported the composite change.

All ASVAB validation/studies should always start with a research plan. Such a plan would ideally state the purpose for conducting the study and the steps to be followed (Drasgow, Whetzel, & Oppler, 2007, p. 351). The phases described in the next section could be part of the ASVAB validation/standards study plan recognizing that not all of the statistical procedures described in the Technical Manual would become part of the plan. That is, as illustrated by the reports included in the Appendices (Nuclear Field, SEALs, and Mineman Ratings), each study has the objective of addressing issues that are specific to a Rating. Further, each Rating may have only some forms of performance measures, which somewhat dictates what statistical procedures can be applied. Therefore, the validation plan would be at a higher "Phase" level.

Summary of Navy ASVAB Validation Phases

In general, a Navy ASVAB validation/standards study can be divided into the following phases, in chronological order:

1. Visit the school site
2. Collect ASVAB data and determine the sample
3. Assess performance measurement
4. Collect in-house performance data
5. Analyze non-graduation and setback data
6. Form curriculum-based and empirical-based ASVAB composites
7. Validate ASVAB composites
8. Analyze cutscores
9. Simulate Rating assignments
10. Develop and submit the report with recommendations.

The remainder of the chapter provides a description and explicit instructions for each ASVAB validation/standards study phase.

Phase 1: Visit the school site. Each ASVAB validation/standards study requires a school site visit. The researcher should ask the ECM for a school point of contact (POC). Some schools require a formal Navy visit request, which should be requested via the NPRST security officer who will coordinate with the Navy Personnel Command security office and submit request into an electronic system accessible by the schoolhouse base. Upon arrival at the school, the researcher should brief the Commanding Officer (CO) and school officials about the study objectives and timelines (in-brief) and again upon departure (out-brief) to let the CO and leadership know about the particulars of the completed visit. During the visit, the researcher should (a) meet with school instructors to understand the teaching challenges and perspectives about student capabilities, (b) arrange for a tour of the classrooms, including the laboratories, (c) obtain a curriculum outline, the learning, enabling, and terminal objectives, and if possible, the curriculum itself (if warranted), and (d) obtain the school's testing plan, which will indicate how performance on each training module (including laboratories) is scored and weighted in computing the final school grade (FSG). Finally, the researchers should arrange to follow up with the school officials and establish if any of them would want to review the study before its final submission.

Phase 2: Collect ASVAB data and determine the sample. Both the size and representativeness of the student sample should be considered (Drasgow et al., 2007). Larger sample sizes are preferred (i.e., more than 200). When sample sizes are small, it may be appropriate to combine small datasets for the same Rating over a number of years where there have not been major changes in the curriculum, training methods, or course length. A change in training curriculum might alter the determinants of performance in a given school, therefore possibly changing the ASVAB composite that best predicts training performance. The schoolhouse retains student performance records for two years; however, FSG for most schools are reported in the Navy's corporate training database (described in Phase 4).

Phase 3: Assess performance measurement. School performance measurement is taken for students at various points in the course and again at the end as a summary measurement (i.e., FSG). The individual training module test scores may provide valuable information if most students clearly perform poorly in one or two and so could result in a training recommendation included in the ASVAB validation/ standards report. There have been studies where a recommendation was made to add a module rather than raising the ASVAB standard.

Performance measurements are always taken for classroom-based knowledge, but only sometimes for laboratory demonstrations (practicals or field exercises); although more often the grade is pass/fail for the hands-on training. The laboratory exercises, as explained in the last chapter, may serve as credible surrogates of important job tasks (recognizing most individuals improve with practice), and thus as measures of important aspects of job performance. In some studies, validity ties between two or more ASVAB composites have been broken by using the laboratory grades at the criterion, or incorporating them with FSG (a more academic measure).

Phase 4: Collect in-house performance data. NPRST's Research Information Systems Management Office (RISMO) receives monthly training data extracts from the Navy Integrated Training Resources Administration System (NITRAS, also referred to as CeTARS) database (authorized by an Interface Control Document – ICD with stipulated conditions). NITRAS is the corporate training database that is maintained by Naval Education Training Professional Development Technology Center (NETPDTC) in Pensacola, FL (the agency that developed the ICD). The NITRAS files contain student performance variables, including Personal Identifying Information (PII) that requires special training and data storage handling. The files are secured and stored by NPRST's Research information Systems Management Office (RISMO) Department under strict information security protocols. The data come in as flat files ("Noe" on the front end of the file names).The files are stored in folders called "Nitras" (one for each fiscal year) in a secure NPRST/RISMO server environment. The "Nitras" folders also contain a copy of the data dictionary of "Person Event", or "PEV", codes. The PEV codes have brief descriptions of what the student event entailed, such as setback for a specific reason (code). The PEV codes generally fall into the categories of academic, non-academic, graduate, attrite, and setback. The term "attrite" is a Navy discharge code associated with any number of reasons (e.g., detection of drug use).

Attrite cases are eliminated from an ASVAB validation/standards study (as are cases that clearly dropped for non-academically related reasons); however, a breakout is reported if there are significant occurrences. These eliminated cases may be of interest for other studies that take a broader look at recruiting and youth attributes, or personality prediction studies. However consent must be given by NETPDTC as per the ICD for sharing data with other than the ASVAB Validation/Standards program studies. The NITRAS data extract file also contains official ASVAB scores that HQ-USMEPCOM transmits to the Navy. The ASVAB validation/standards researcher processes the flat file into an SPSS data file for each Rating that is involved in a study. A unique Course Description Processing (CDP) code is associated with each Rating's training course/location and the researcher selects that CDP to develop the study data file. An SPSS syntax file concatenates all PEVs (e.g., PEV1, PEV2, etc.) chronologically for each student record on one record line along with the dates of occurrence and other file variables. The ASVAB validation/standards researcher is responsible for editing the data for out-of-scope PEVs, other anomalies, and to resolve missing values.

Phase 5: Analyze non-graduation and setback data. Most students who have academic or non-academic difficulties do not attrite, but are setback one or more times before either graduating or being reclassified to another Rating. Besides PEV codes (that also describe setback reason), a student disposition code (DIS) appears in the NITRAS data that logs what happened to the student upon leaving the training (e.g., "Reclassified"). The school's academic setback rate and number of setbacks are useful variables to analyze in cutscore analysis along with pass/fail outcome. A high graduation rate along with a high academic setback rate may indicate that graduation rates are "managed", usually to lower training costs and address Fleet demands for training Sailors. There are costs associated with a trainee being remediated as well as failing (e.g., stigma, morale, and career setback). One major cost not addressed is the Navy "Readiness" cost, which is a conglomerate of many factors.

Phase 6: Form curriculum-based and empirical-based ASVAB composites. Both rational (mapped to the curriculum) and empirical approaches (developed from the data) should be used to formulate a set of "experimental" composites. A rational approach would involve studying the curriculum, testing plan, and other course documents obtained from the school to identify the apparent course constructs. All Navy courses are multidimensional, as is the ASVAB, and there may or may not be an obvious mapping of their constructs. An empirical approach would involve regression methods; however, as we will learn in the Technical Manual, the procedure has limitations if solely used on the ASVAB range restricted study sample (due to the ASVAB cutscore). We learn how to estimate the ASVAB correlations with FSG for a relevant applicant population in the technical manual. Past Navy studies have used both the sample and the estimated population correlation matrix for developing "experimental" ASVAB composites.

The researcher should limit the number of tests in the experimental composite to three or four to minimize the redundancy with other composites (recognizing a possible sacrifice of some predictive validity). The number of ASVAB subtests in Navy operational composites range from one to four and all are integer weighted (see Table 6-1). Redundancy in a set of ASVAB classification composites can be measured by averaging their intercorrelations in a full range population - the ASVAB 1997 Profile of American Youth (PAY97) (Segall, 2004) displayed in Appendix A of the Technical Manual. The lower the average intercorrelation, the lower the composite redundancy and the greater the capability of the set of composites to differentially assign enlistees to jobs. Ideally, each composite should tap into a core set of aptitudes, abilities, and knowledge that are important for training success in one school but not necessarily as important for training success in other schools. The ASVAB composite formulation phase should produce a small number of viable candidates for validation to reduce change relations, but should always include (a) the operational composite(s) for the Rating and (b) operational composites for similar Ratings (limiting the proliferation of composites).

Phase 7: Validate ASVAB composites. Validity coefficients are calculated for the operational ASVAB composite(s) and the candidate replacement(s) (i.e., rationally and empirically derived). Calculation of the ASVAB composite validity coefficients involves use of the multivariate range restriction correction formulas (fully described in the Technical Manual and available as an SPSS file in the manual's Appendix A). Validity coefficient calculations in the sample at hand are downwardly biased due to the reduction of ASVAB score variance that occurs from use of the operational ASVAB cutscore. The correction for range restriction provides an estimate of the ASVAB validities of interest – the values that apply to full range youth populations from which; theoretically, future recruits will be selected for military service. The ASVAB normative population, PAY97, serves as the unrestricted youth population for the Navy; however, the other Services may use their applicant populations. The Navy takes the position that a single ASVAB normative population use in the correction (validities estimated for that population) enables researchers to track validity trends over time and possibly across the Services for similar occupations when join-service studies are involved.

Phase 8: Analyze cutscores. ASVAB cutscores are set within a dynamic and changing recruiting environment, and so the ASVAB validation/standards team must consider a host of factors:

- academic attrition and setback rates,
- observed waiver rate (indicating stress in recruiting),
- criticality of the Rating,
- yearly school input requirement,
- cognitive complexity of the training and job, and
- time allowed for training and training cost.

If alternative composites are recommended, which would need to occur if the most valid ASVAB composite contains either CS or AO (see Chapter 2 for a description of the tests), a cutscore that achieves the same relative aptitude/ability level can be set for each composite. A common way of estimating the same aptitude ability level is to establish the standard score point (z-score) for a ASVAB composite using the mean and standard deviation of that composite derived for the PAY97 population in a linear equation (Segall, 2004). This process is described in the Technical Manual chapter on setting ASVAB cutscores (Chapter 17).

Phase 9: Simulate Rating assignments. Simulating recruit assignments to jobs (Ratings and programs within Ratings) is optional but really required if the Rating under study has a substantial yearly fill requirement (say more than 300) and the ASVAB composite is changed and the cutscore is substantially raised, particularly in a difficult recruiting environment. The data required for simulating recruit assignments to Ratings are obtained by RISMO annually specifically for the ASVAB Validation/ Standards program from the Navy Recruiting Command (NRC). This data acquisition process requires NPRST/RISMO to submit a Data Transfer Compliance Review Checklist and other materials that are part of the Navy's Manpower, Personnel, Training and Education (MPTE) Enterprise Information Management (EIM) protocols managed by the BUPERS Chief Information Officer (BUPERS-07). Included in the package are the study's IRB protocol and IRB approval letter issued by the NPRST Institutional Review Board (IRB).

Approval for the data transfer from NRC to NPRST/RISMO signals NPRST to contact NRC to schedule a meeting to discuss NPRST's requested list of variables. These files, as with the NITRAS files, are secured and stored by NPRST/RISMO. Special classification tests taken at the MEPS will not have their scores transferred to the Navy data systems, with the exception of the CS test. For example in order for NPRST to receive Navy applicants' Cyber test scores, or the working memory Mental Counters test scores HQ-USMEPCOM requires a Data Sharing Agreement (DSA) with NPRST with NPRST's submission of the study IRB protocol and approval documents. Direct electronic transfer of these special classification test scores occurs only upon OPNAV132G notifying HQ-USMEPCOM that the tests are operational (prior to that, they are considered in a research phase). The results of the simulation of recruits' Rating assignments will help inform policy makers about the difficulty in filling Ratings

and the potentially impact of ASVAB score point-waiver policy on training performance. The Technical Manual provides more information about the Navy's two recruit classification simulation applications, both of which were used to study how the Coding Speed and Assembling Objects test improved gender and race/ethnic Rating qualification rates.

Phase 10: Develop and submit the report with recommendations. The ASVAB validation/standards study is written in letter report format with an attached cover letter summarizing the study and recommendations. The letter report is signed by the NPRST Director, serialized, and submitted to the Director, Navy Selection and Classification (N132G) for review and policy action. If the recommendations are approved, N132G issues an ASVAB change directive to all activities that use and maintain ASVAB standards (e.g., recruiting manual, military personnel manual, classification software, etc.). The NPRST ASVAB validation/standards team maintains binder books that contain historical ASVAB validation studies/letter reports/technical notes. Three letter reports, written at a level that is comprehensible to the customers (NRC, NETC, ECMs) are provided as examples in Appendices A, B, and C (Held, 2011; Held, 2012; Held, Alderton, & Britton, 2010).

Chapter 6. References

Drasgow, F., Whetzel, D. L., & Oppler, S. H. (2007). Strategies for test validation and refinement. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 349-384). Mahwah, NJ: Earlbaum.

Hedge, J. W., Carter, G. W., Borman, W. C., Monzon, R. I., & Foley, P. P. (1992). *Evaluation of Navy operational and alternative Armed Services Vocational Aptitude Battery (ASVAB) composites* (Institute Report #216). Minneapolis: Personnel Decisions Research Institute.

Held, J. D. (2011). *Armed Services Vocational Aptitude Battery (ASVAB) Standards: Basic Underwater Demolition/SEALs*. (Letter Report 3900 BUPERS-1/10092 of 11 Aug 12) NPRST, Millington, TN.

Held, J. D. (2012). *Armed Services Vocational Aptitude Battery (ASVAB) Standards: Mineman (MN) Rating*. (Letter Report 3900 BUPERS-13/83 of 17 Aug 12) NPRST, Millington, TN.

Held, J. D., Alderton, D. L., & Britton, D. (LCDR). (2010). *Armed Services Vocational Aptitude Battery (AS VAB) and Navy Advanced Placement Test (NAPT) Standards: Nuclear Field (NF) Ratings*. (Letter Report 3900 BUPERS-1/00158 of 24 Aug 10) NPRST, Millington, TN.

Johns, C. (2011). *Job similarity profiles and algorithms for Armed Services Vocational Aptitude Battery (ASVAB) validation priority determination* (Contract Number: TCN 10-177-0989 ARO/Battelle). Report and application prepared for Navy Personnel Research, Studies, and Technology (NPRST).

Segall, D. O. (2004). *Development and evaluation of the 1997 ASVAB score scale* (Technical Report No. 2004-002). Seaside, CA: Defense Manpower Data Center.

Chapter 7.
Applications of Synthetic Validity
Jeff W. Johnson

Introduction

The military is in a good position to conduct criterion-related validity studies using the ASVAB as the predictor and training performance as the criterion. This is so because a large number of youth are enlisted for the military services each year and subsequently trained in occupations. Training grades are generally available from the military schoolhouses and ASVAB scores are always available from Department of Defense (DoD) and military service databases. However, at least for the Navy, traditional criterion-related validity studies cannot be performed well when (a) there are newly formed occupations (Ratings) or Rating mergers due to emerging requirements and criterion data are simply not yet available or (b) sample sizes are too small for statistically sound analyses. The Navy takes a fallback position in these cases in setting *initial or interim* ASVAB standards until there are better study conditions. Specifically, the Navy's fallback position is for the ASVAB validation/standards team to consider a comprehensive set of factors in an ASVAB standards decision, which are (a) the training curriculum (through formal training task analysis workshops), (b) the complexity and difficulty of the training and similarity to other Ratings' training, (c) the time allowed for training (which, if too short, may influence the level of cutscore as much as the training difficulty), (d) the similarity of the ASVAB classification composites and cutscores in similar types of occupations, and (e) the validity magnitude of the ASVAB composites under consideration. .

Obviously when a sample size is small, one could wait several years for an adequate sample size to materialize; however, the Navy schoolhouses only retain performance data for two years, and even at that, the sample size may fall short or the training could change dramatically over the time-frame. The Navy's approach to establishing the ASVAB's predictive validity in the absence of training performance data is only loosely akin to what industrial/organizational psychologists call validity generalization. For the interested reader, the Society of Industrial and Organizational Psychologist (SIOP) "Principles for the Validation and Use of Personnel Selection Procedures" provides a brief discussion of validity generalization, meta-analysis, validity transportation, and this chapter's topic, synthetic validity (pages 27-30 at http://www.siop.org/_principles/principles.pdf. We provide this chapter on synthetic validity as it is an empirically-based approach that is legally defensible and endorsed by SIOP (see Steel, Huffcutt, & Kammeyer-Mueller, 2006 for a broader discussion of synthetic validity) in contrast to some validity generalization methods. We do not go into the details in the Technical Manual as the Navy does not apply the formal synthetic validity methods in ASVAB validation/standards studies but only provide this chapter as a starting point for those (typically in industry) who might want to conduct a thorough review of the various validity generalization/synthetic validity methods.

The Synthetic Validation Approach to Test Validation

Synthetic validation is the process of inferring validity in a specific situation based on (a) the identification of basic components of the job (i.e., cluster of related work behaviors), (b) a determination of test validity for predicting performance on those components that are similar for other jobs, and (c) a combination or synthesis of the component validities into a whole (Cascio, 1987). Synthetic validity can be used to assemble a job-relevant test battery or to calculate validity coefficients for jobs in which there are too few incumbents to conduct a traditional criterion-related validity study, or when necessary criterion data are otherwise unavailable (e.g., Guion, 1965; Hoffman & McPhail, 1998; Hollenbeck & Whitener, 1988). Procedures for conducting a synthetic validation study are described by Johnson and Carter (2010).

Synthetic validation is based on two assumptions (Johnson, 2007). First, when a job component (clusters of related work behaviors) is common across multiple jobs, the human attributes predictive of performance on that component are similar across jobs. The second assumption is that the validity of a test for predicting performance of a job component is similar across jobs and situations. In other words, the assumption is that any differences found in the relation between a test and a component of job performance across jobs is merely due to sampling error, unreliability, or other random factors (e.g., see Hunter & Schmidt, 1990, for correcting error and bias within the meta-analytic framework).

Synthetic validation assumptions are similar to those made for validity generalization (Jeanneret, 1992), a concept that has received considerable research support (AERA, APA, & NCME, 1999; Schmidt, Hunter, & Pearlman, 1982; SIOP, 2003). Hoffman and McPhail (1998) discussed validity generalization as one of four options available for organizations in establishing test validity in less than optimal data analytic conditions for any particular job: (a) synthetic validation, discussed in this chapter; (b) validity generalization, which essentially summarizes validities of similar instruments across similar jobs; (c) test transportability, which essentially assumes validity based upon an in-depth evaluation of the similarities of like jobs; and (d) test implementation without validity estimation (never an option for the military).

Synthetic validity is not a "type of validity" but rather a *process* by which evidence for the interpretation of a test, or test battery score, is *inferred* for a particular job. Lawshe (1985) discussed the misrepresentation of the validity coefficient in general, in that some psychologists refer to the validity of the test, when it is the validity of the inferences we can make from a test's scores (see Cizek, 2012 for a recent discussion). In synthetic validity, the inference is *based* on the validities (either empirical or judgmental) of the test(s) for predicting performance on the components that constitute the job (Johnson, 2007).[3]

[3] The Navy's approach to synthetic validity and validity transportation is to consider both the training community's work-oriented job task analysis approach (geared towards curriculum development) and the manpower community's work- and worker-oriented job analysis (geared towards establishing Sailors' required skills and abilities).

The validity of a test for predicting a particular job component may be determined using any of a variety of research strategies (e.g., content sampling, experimental design, criterion-related correlations, expert judgment, or a combination of strategies).The two primary synthetic validation methods are Job Component Validity (JCV) and what Steel, Huffcutt, and Kammeyer-Mueller (2006) called the Job-Requirement Matrix (JRM), each imposing a quantitative rigor suitable for relating worker requirements with worker-related job components.

Job Component Validity (JCV)

JCV is a term coined by McCormick (1959) to describe a specific type of synthetic validation technique that indirectly links selection test (or battery) scores and "worker-related" job components, as opposed to "work-related" components. The Position Analysis Questionnaire (PAQ) developed by McCormick and colleagues (e.g., McCormick, DeNisi, & Shaw, 1979; McCormick, Jeanneret, & Mecham, 1972) has been widely used in worker-oriented (behavior) linkages to work-related (requirements). The PAQ is a structured job-analysis instrument used to quantify by rating scales characteristics of lower-level job elements (the smallest unit of work that has a clear beginning, middle, and end) such as time performed, the element's applicability, and importance to the job. Formulas incorporating element weights produce PAQ dimension scores that are then correlated with test (or battery) scores. We note that the PAQ is often used as an adjunct to other job analysis methods because the instrument does not meet every job analysis purpose.

Factor analysis of the PAQ lower level job elements produces over 40 higher level job dimensions, referred to as job components because of the worker orientation in JCV. We note that all synthetic validation techniques involve job dimensions, factors, or components (all referring to analytically derived higher order categories containing lower level elements), but that the specific term Job Component Validity (JCV) is usually reserved for McCormick's approach. The approach has been to score the dimensions/components and to use them in regression equations to predict level of worker attributes, such as verbal, numerical, and spatial aptitudes as measured by aptitude/ability tests. The results have shown a sufficient linkage and thus the potential to develop validity coefficients for aptitude/ability tests in predicting job analysis data that is quantitatively scored (without a requirement to collect actual job performance measures from job incumbents). The JCV approach can also incorporate personality/temperament measures in predicting job analysis data, not yet addressed by the Navy but well addressed by the Army (e.g., see Campbell et al., 2007).

Job Requirement Matrix (JRM)

The Job Requirement Matrix (JRM) approach has its roots in the J-coefficient approach introduced by Primoff (1957; 1959). This approach uses job analysis to identify the job components that are common across multiple jobs and the predictor measures that predict performance on those job components. The J-coefficient used in JRM is a mathematical index of the relation between the test battery and job performance. There are many J-coefficient formulas, most of which involve a vector of the relations between the predictors and the job components, and a vector of the relations between the job

components and job performance (Hamilton & Dickinson, 1987).The elements of the J-coefficient formula can be estimated either empirically through the correlations between predictors and job components or rationally through subject matter expert (SME) judgments. Because empirical correlations between predictors and job components require large sample sizes and performance ratings, the use of SME judgments is more typical (Scherbaum, 2005). For example, the relation between predictors and job components could be estimated by asking SMEs to rate the relevance of each predictor to each job component. Alternatively, test experts or Industrial/Organizational psychologists could estimate the validity coefficients between predictors and job components (e.g., Schmidt, Hunter, Croll, & McKenzie, 1983).

There are two primary differences between the JCV and JRM synthetic validation techniques. First, the JCV approach links tests and performance constructs indirectly by demonstrating that, across jobs, job incumbents' test scores or test validity coefficients are related to the importance of the attribute measured by the test, as determined by a standardized job analysis survey (Mossholder & Arvey, 1984). Other types of synthetic validation approaches, including JRM, depend on more direct linkages between tests and performance constructs. Second, the JCV approach is based on the assumption that high-ability individuals tend to gravitate toward jobs with high-ability requirements and low-ability individuals tend to gravitate toward jobs with low-ability requirements (McCormick et al., 1979). This idea of, basically, self-selection, is commonly referred to as the "gravitational hypothesis" (e.g., Wilk, Desmerais, & Sackett, 1995) and somewhat downplays the role of any cutscore that was set in the hiring process. Cutscores applied to the ASVAB in the military context result in restriction in range of test scores for those who end up in military occupations, so it is not clear how much influence the individual has in choosing an occupation or if the gravitational hypothesis even marginally holds.

Recent Applications of JCV

There have been several interesting applications of the JCV approach. For example, Hoffman and McPhail (1998) compared synthetic validity coefficients calculated using the JCV model to validity generalization (meta-analysis) data for an assortment of clerical jobs from Pearlman, Schmidt, and Hunter (1980). The JCV model uses PAQ (McCormick et al., 1972) "dimension" scores that can be used to predict validity coefficients that apply to the General Aptitude Test Battery (GATB) cognitive ability constructs across studies (Jeanneret, 1992) as well as mean GATB scores found for incumbents across jobs. (The GATB is a multiple test battery used for employment and occupational assessment.) In a study of clerical jobs within a utility company, Hoffman and McPhail applied five PAQ dimensions (general mental ability, verbal ability, quantitative ability, perceptual ability, and spatial/mechanical ability) for linkage to the counterpart GATB constructs (General = G; Verbal = V; Numerical = N; Clerical = Q; and Spatial = S). The Hoffman and McPhail results showed a substantial correspondence between mean JCV estimates and mean meta-analysis estimates for each construct, supporting the viability of JCV studies in situations in which sample sizes are too small for a criterion-related (local) validation study.

In another study, Hoffman, Holden, and Gale (2000) used JCV to estimate validities for small-sample jobs, and created test batteries based on validity data gathered on large-sample jobs in the same organization. Jobs were grouped into job families on the basis of a cluster analysis of PAQ dimension scores and rational judgment by SMEs. One or more jobs within each job family had validity data for certain tests, and the validity of these tests was transported to the other jobs within the family. JCV predictions were made for each GATB construct for jobs within each job family and compared to the validation results for existing tests measuring the same constructs. The PAQ dimension scores were then used to determine the appropriate predictor constructs to include in test batteries for each job family. For the job family specific test batteries, the validity results showed that mean validity estimates obtained at the level of individual predictor constructs via JCV reflected useful levels of validity within each job family.

Other more recent JCV work has focused on extending the JCV method to job analysis tools other than the PAQ. For example, Brown and Harvey (1996) used JCV to predict personality dimension scores with job analysis data obtained using the Common Metric Questionnaire (CMQ). D'Egidio (2001) applied the JCV method to the Occupational Information Network (O*NET), linking O*NET descriptors within domains to mean scores on published cognitive ability tests. Jeanneret and Strong (2003) predicted mean GATB scores using the Generalized Work Activity (GWA) ratings from O*NET. Johnson, Carter, and Dorsey (2003) also predicted GATB scores using O*NET data, but they chose predictors from all O*NET descriptor domains simultaneously rather than separately. Finally, LaPolice, Carter, and Johnson (2005) used a JCV approach to predict mean literacy scores from the National Adult Literacy Survey (NALS) from O*NET ratings across descriptor domains. In all of these studies, mean ability scores were highly predictable from the job analysis data.

Recent Applications of JRM

There have also been several interesting applications of the JRM approach. For example, Hollenbeck and Whitener (1988) extended the traditional synthetic validity paradigm by suggesting a different order of aggregation. In traditional approaches (e.g., Guion, 1965; Primoff, 1959), the relations between tests and job components are determined first and then aggregated into an overall estimate of validity for a single job. In Hollenbeck and Whitener's approach, job component performance scores are weighted by their importance for an individual's job (determined by a job analysis) and aggregated. Thus, each employee in the validation study has an overall performance score based on the weighted sum of the job component scores, with the weights applied to each job component differing according to the employee's job. Test scores are also aggregated, such that each employee has an overall test battery score and an aggregate performance score. Then the correlation between aggregated test scores and aggregated performance scores is computed across all employees resulting in one correlation rather many correlations between individual predictors and individual job components, which would then be aggregated into an overall validity coefficient.

One advantage of the single correlation approach is that the number of employees included in the validation study is not limited by the smallest sample (for a specific job). Rather than providing validity estimates for different test batteries for a number of different jobs in an organization, this approach provides a single estimate of validity for a test battery within the organization as a whole. Although this approach would be useful for small organizations, so far it has had very little impact (Scherbaum, 2005).

In the U.S. Army's Synthetic Validity project (Peterson, Wise, Arabian, & Hoffman, 2001), synthetic validity equations were derived for several Military Occupational Specialties (MOS), and validity estimates using these equations were compared to estimates from traditional criterion-related validation studies conducted as part of Project A (see Chapter 3 for a brief discussion of Project A and references). Industrial-organizational psychologists estimated the magnitude of the relation between each predictor (taken from the Project A test battery) and each job component. These estimates served as the basis for developing a weighting scheme for each MOS for predicting core technical proficiency and overall performance. Results showed very little difference between the validity coefficients developed from the synthetic validation approach and the validity coefficients obtained in the traditional manner. Also, Peterson et al. (2001) examined the extent to which a synthetic validity equation derived for one Army MOS predicted criterion performance for another MOS. There was very little discriminant validity, meaning that mean differences between validity coefficients derived from MOS-specific equations and other-MOS equations were typically very small. The small validity differences were attributed to the test batteries being very cognitively loaded; cognitive ability tends to predict performance in a wide variety of jobs (Scherbaum, 2005). Also more highly intercorrelated tests would also minimize the differential effects of alternative weighting schemes (Wainer, 1978).

In another study, Johnson, Carter, and Tippins (2001) applied a synthetic validation technique to a selection system development project in a large private-sector organization. The organization desired a single selection system for approximately 400 non-management jobs organized into 11 job families. Even when the jobs were grouped into job families, there were still too few employees in some jobs and some job families to conduct a traditional criterion-related validity study. Also, the job title structure changed rapidly, which made it necessary to use the selection system for new jobs that did not exist at the time the validation study was conducted. Faced with this situation, a synthetic validation strategy was considered the most appropriate method for developing the project's selection system. A job analysis identified 27 job components that described the major work behaviors across the job families. Twelve tests were developed to predict supervisor-rated performance on these job components and a concurrent criterion-related validation study was conducted to collect test and job component data (for 1,926 incumbents). A test composite was chosen for each job component based on a combination of psychologist judgments and empirical relations. In other words, both expert judgments of the relations between tests and job components and empirical correlations between test scores and performance ratings on job components were available. A test was determined to be relevant for a job component if expert judgments and/or correlations indicated a strong relation. A test battery was chosen for each job family based on its important job components. Using the equation for computing the correlation between two composites (e.g., Nunnally &

Bernstein, 1994), a validity coefficient for predicting a composite of performance on each job family's important job components was synthesized for each job family's test battery. Because test intercorrelations, job component intercorrelations, and correlations between tests and job components could be computed across job families, an overall validity coefficient could be computed for each job family by using the equation rather than by actually computing composite scores and calculating the correlations within job families. This approach allowed the researchers to take advantage of the large overall study sample size to compute stable validity estimates, even for very small job families.

Because sample sizes within some job families were relatively large, Johnson et al. (2001) were able to compare synthetic validity coefficients to traditional validity coefficients calculated within those job families. The synthesized validity coefficients were very similar to empirical within-family validity coefficients for many job families, indicating that validity coefficients computed from test-job component correlations calculated across job families are reasonable substitutes for traditional validity coefficients calculated within large job families.

Johnson, Paullin, and Hennen (2005) applied a similar synthetic validity procedure to a large public-sector organization for several reasons. First, the five jobs included in the study had many job components in common, allowing validity coefficients to be computed on a larger sample size.[4] Second, the available sample size for one of the five jobs was too small to allow meaningful validity coefficients to be computed for that job alone. Third, the organization expected that test batteries would be needed in the near future for several other jobs similar to those included in the synthetic validation study. The synthetic validation approach allows computation of validity coefficients for jobs not included in the validation study, as long as they comprise job components for which validation data are available. Faced with these issues, the synthetic validation strategy was considered the most appropriate for this project.A noteworthy aspect of the Johnson et al. study is the range of included predictors. Scherbaum (2005) observed that most synthetic validity studies have examined only cognitive, perceptual, and psychomotor ability tests, and so have called for more research including constructs such as personality and vocational interests. Johnson et al. included tests measuring cognitive ability, perceptual speed, biographical data, and six personality constructs (e.g., conscientiousness, interpersonal skill, initiative, and service orientation). Johnson et al. (2005) and Johnson, Carter, and Tippins (2001) also examined alternative predictor and criterion weighting schemes.

In both the Johnson et al. (2005) and Scherbaum (2005 studies), validity coefficients were highest when job components were unit weighted and predictors were weighted according to the number of job components to which they were relevant. Similar to Peterson et al. (2001), Johnson et al. (2005) found that applying a common set of weights to all jobs as opposed to using job-specific weights resulted in very small differences in validity.

[4] It is always beneficial to increase the sample size for computing a validity coefficient. A larger sample size decreases the standard error around the validity coefficient, providing more confidence in the results.

Johnson, Carter, Davison, and Oliver (2001) extended the synthetic validity paradigm to the testing of differential prediction hypotheses. The hypothesis of slope or intercept differences for different subgroups of examinees is typically difficult to test because of small sample size and low power (Aguinis & Stone-Romero, 1997). To increase sample size and therefore power, Johnson et al. showed that the same procedures used to compute correlations between tests and job components across jobs can be used to compute correlations between job component scores and the other variables necessary for differential prediction analyses (i.e., a dummy-coded subgroup variable and the cross-product of the subgroup variable and the predictor score). Equations for computing correlations between a variable and a linear composite or correlations between two linear composites are used to create the matrix of synthetic correlations between overall performance, test scores, subgroup membership, and cross-product terms. This matrix is used to conduct the moderated multiple regression analyses necessary to determine if there is differential prediction across groups. Johnson et al. (2001) illustrated the procedure by showing that the sample size for one job was 149 for a traditional within-job differential prediction analysis, but was increased to 1,361 by using synthetic differential prediction analysis. This analysis increased the power to detect a significant effect from .17 to .97.

Finally, McCloy (1994, 2001) combined a synthetic validity approach with hierarchical linear modeling to create prediction equations for jobs in which criterion data are not available. McCloy created a multilevel regression equation that related (a) individual scores on job-specific hands-on performance tests to characteristics of the individuals (level-one equation) and (b) job-specific, level-one regression parameters to job characteristics determined by a job analysis (level-two equations). Job analysis data for jobs lacking criterion data are entered into the level-two equations, yielding estimated level-one parameters. These level-one parameters are then applied to the individual characteristic variables to yield predicted performance scores for each individual. McCloy (2001) demonstrated that the linkage method provides predictions of job performance for jobs without criterion data that are very similar to predictions obtained from cross-validated least-squares regression equations when criterion data are available.

Legal Implications

There have been two court cases involving selection procedures based on JCV (*McCoy et al. v. Willamette Industries*, 2002; *Taylor v. James River Corporation*, 1989). Although both decisions were in favor of the JCV approach, these were summary judgments that did not carry much legal weight (Scherbaum, 2005). The legal defensibility of other types of synthetic validity has not yet been challenged. The *Uniform Guidelines on Employee Selection Procedures* (1978) do not address synthetic validity directly, but Trattner (1982) argued that the operational definition of construct validity provided by the *Guidelines* is actually a description of a synthetic validity model. The *Guidelines* state:

> "…if a study pertains to a number of jobs having common critical or important work behaviors at a comparable level of complexity, and the evidence satisfies…criterion-related validity evidence for those jobs, the selection procedure may be used for all the jobs to which the study pertains" (p. 38303).

Trattner (1982) interpreted this definition of construct validity as meaning that a selection instrument can be used when work behaviors are important in any occupation within a class of occupations, as long as there is criterion-related validity evidence linking the instrument to the work behaviors for incumbents in the class. Trattner concluded that the synthetic validity approaches of Primoff (1959) and Guion (1965) were consistent with this interpretation of the *Guidelines*. The approaches of Peterson et al. (2001) and Johnson (Johnson, Carter, & Tippins, 2001; Johnson et al., 2005) also appear to meet these requirements of the *Guidelines*.

It is more difficult to infer how a synthetic validity study that is based only on SME judgments would be received in light of the *Guidelines*. The *Guidelines* clearly state that criterion-related validity evidence is required, so there is no direct support in the *Guidelines* for a synthetic validity study in which linkages between tests and job components are provided by SMEs. Of course, the *Guidelines* define an acceptable content validation strategy as demonstrating that the content of a selection procedure is representative of the important aspects of performance on the job. This demonstration is precisely what is done in a synthetic validation study when SMEs link the selection procedure content to job components, and the only content included for a particular job has been linked to the important job components for that job. Thus, it seems likely a synthetic validity study based on a content validity strategy would meet the requirements of the *Guidelines* (Johnson, 2007).

Scherbaum (2005) identified several reasons why synthetic validity should be legally defensible. First, any synthetic validity approach requires a comprehensive job analysis. The first step of a well-constructed selection study is a thorough job analysis, and a selection procedure that is not based on an appropriate job analysis will rarely pass muster with the court. Second, utilizing a synthetic validity approach forces the personnel specialist to create a test battery that is job-relevant. Tests are chosen to measure attributes that have been determined to be relevant to important components of the job. Third, Scherbaum noted that Varca and Pattison (1993) believed that concepts like synthetic validity could more easily be added to the validity frontier based on the trends of recent decisions.

There are also legal implications when using Johnson, Carter, Davison, and Oliver's (2001) synthetic differential prediction analysis. These authors recommended considering the effect size when conducting differential prediction analyses, because the large sample sizes that are possible could lead a meaningless effect to be statistically significant. If a significant slope or intercept difference increases R^2 by less than .01, that could be considered too small an effect to be meaningful. They noted that this argument might not hold up in court, where it could be more difficult to ignore a statistically significant result.

It appears that the synthetic validation approach is consistent with the *Uniform Guidelines on Employee Selection Procedures* (1978), but there is no case law as yet to directly support these procedures. Ultimately, the quality of the job analysis, the appropriateness of the procedures used, and the nature of the inferences made by the users will determine the defensibility of any synthetic validation procedure (Scherbaum, 2005).

Developing a Synthetic Validity Database

A major constraint on the use of synthetic validity is the need to identify and validate predictors for the job components with each new synthetic validity study. There is very little accumulated research on specific predictor measures because most synthetic validity studies have included predictor measures that are not commercially available (Mossholder & Arvey, 1984; Scherbaum, 2005). Hough (2001; Hough & Ones, 2001) called for the creation of a database to be used with synthetic validation models to build prediction equations for specific situations. The idea is to conduct primary studies that report relations between predictor constructs and job components, and then use meta-analysis to cumulate the results of those studies. Such a database would allow us to use synthetic validity techniques to estimate the validity of a battery of predictors for any job that includes job components on which research is available.

When such a database is large enough, practitioners will be able to buy or develop measures of predictor constructs that have been shown to predict performance on job components relevant to any job of interest and to calculate a validity coefficient for that job. The development of this type of database should be the ultimate goal of synthetic validation research. Steel et al. (2006) argue that, although development of a synthetic validity database requires a large amount of resources, larger human resources projects have been conducted in the past and the technology and infrastructure available now makes this type of study more and more feasible. Relatedly, McCloy, Putka, and Gibby (2012) provided one solution to the development of an online tool for estimating and accumulating synthetic validity information.

Concluding Remarks

Because the military can readily conduct most of the criterion related validity studies required to validate and set ASVAB standards, the synthetic validity approaches may not have high utility as a standard procedure. There is a cost over current Navy methods for setting ASVAB standards in unique cases, mainly in the development, updates, and maintenance of a synthetic validity database. These costs might be considered marginal, however, if the focus on the ASVAB for occupational classification is expanded to include other measures such as personality/temperament, biographical/experience data, or simply, new cognitive tests added either to the ASVAB or as adjuncts.

In an expanded predictor set, the natural course of action would be to include job components, not just training as the performance criterion. The Army is leading the way in this regard and the intent of this chapter was to look past the Navy's methods for establishing ASVAB standards for their Ratings purely on the basis of training performance, for which historically the battery has been developed to predict.

To help meld training and job components for the military, the Human Resources Research Organization (HumRRO) demonstrated a promising simplified job analysis method with extensive use of SMEs to link ASVAB constructs, training curriculum, Knowledge, Skills, and Abilities (KSAs), and major job duties (Waters et al., 2009). The method was applied to a subset of diverse occupations across the Services and identified gaps in the ASVAB as part of an effort to expand the ASVAB content. The HumRRO method could be used to support a modified military version of a synthetic validity approach that serves several purposes including (a) the identification of construct gaps in selection/classification instruments, (b) broken linkages between training and job requirements, and (c) to establish cognitive/non-cognitive standards where performance data are not available (or sample sizes are too small for a robust statistical result). The approach may be particularly useful for Ratings with a computer-based training (CBT) format, where ASVAB validity coefficients appear high because of a large academic component.

It is interesting to note that with all of the legal requirements related to conducting validity analyses in industry, there is no legal requirement for the military. The Introductory and Technical Manuals on conducting ASVAB validation/standards studies may be a catalyst for developing a Service-level policy and, ultimately, a DoD policy for the requirement. We now encourage those who have a technical background and who conduct test validation studies to read the accompanying document, "Technical Guidance for Conducting ASVAB Validation/Standards Studies in the U.S. Navy."

Chapter 7. References

American Educational Research Association/American Psychological Association/ National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 82*, 192-206.

Brown, R. D., & Harvey, R. J. (1996, April). *Job-component validation using the MBTI and the Common-Metric Questionnaire (CMQ)*. Poster presented at the 11th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Campbell, J. P., McCloy, R. A., McPhail, S. M., Pearlman, K., Peterson, N. G., Rounds, & J., Ingerick, M. (May 2007). *U.S. Army Classification Research Panel: Conclusions and recommendations on a classification research strategy (*Study Report 2007-05). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Cascio, W. F. (1987). *Applied psychology in personnel management* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justification of test use. *Psychological Methods, 17*, 31-43.

D'Egidio, E. L. (2001). *Building a job component validity model using job analysis data from the Occupational Information Network*. Unpublished doctoral dissertation, University of Houston, Houston, TX.

Guion, R. M. (1965). Synthetic validity in a small company: A demonstration. *Personnel Psychology, 18,* 49-63.

Hamilton, J. W., & Dickinson, T. L. (1987). Comparison of several procedures for generating J-coefficients. *Journal of Applied Psychology, 72*, 49-54.

Hoffman, C. C., Holden, L. M., & Gale, K. (2000). So many jobs, so little "N": Applying expanded validation models to support generalization of cognitive test validity. *Personnel Psychology, 53*, 955-991.

Hoffman, C. C., & McPhail, S. M. (1998). Exploring options for supporting test use in situations precluding local validation. *Personnel Psychology, 51*, 987-1003.

Hollenbeck, J. R., & Whitener, E. M. (1988). Criterion-related validation for small sample contexts: An integrated approach to synthetic validity. *Journal of Applied Psychology, 73,* 536-544.

Hough, L. M. (2001). *I/Owes its advances to personality*. In B. W. Roberts & R. T. Hogan (Eds.), *Applied personality psychology: The intersection of personality and industrial/organizational psychology* (pp. 19-44). Washington, DC: American Psychological Association.

Hough, L. M., & Ones, D. S. (2001). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In N. R. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work & organizational psychology* (Vol. 1, pp. 233-277). NY: Sage.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of Meta-Analysis: Correcting error and bias in research findings*. Beverly Hills, Russell Sage Foundations.

Jeanneret, P. R. (1992). Applications of job component/synthetic validity to construct validity. *Human Performance, 5,* 81-96.

Jeanneret, P. R., & Strong, M. H. (2003). Linking O*NET job analysis information to job requirement predictors: An O*NET application. *Personnel Psychology, 56*, 465-492.

Johnson, J. W. (2007). Synthetic validity: A technique of use (finally). In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. 122-158). San Francisco: Jossey-Bass.

Johnson, J. W., & Carter, G. W. (2010). Validating synthetic validation: Comparing traditional and synthetic validity coefficients. *Personnel Psychology, 63*, 755-795.

Johnson, J. W., Carter, G. W., Davison, H. K., & Oliver, D. (2001). A synthetic validity approach to testing differential prediction hypotheses. *Journal of Applied Psychology, 86*, 774-780.

Johnson, J. W., & Carter, G. W., & Dorsey, D. W. (2003, April). *Linking O*NET descriptors to occupational aptitudes using job component validation*. Poster presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Johnson, J. W., Carter, G. W., & Tippins, N. T. (2001, April). *A synthetic validation approach to the development of a selection system for multiple job families*. In J. W. Johnson & G. W. Carter (Chairs), Advances in the application of synthetic validity. Symposium conducted at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Johnson, J. W., Paullin, C., & Hennen, M. (2005). Validation and development of an operational version of Exam 473. In C. Paullin & J. W. Johnson (Eds.), *Development and validation of Exam 473 for the United States Postal Service* (Institute Report #496). Minneapolis: Personnel Decisions Research Institutes, Inc.

LaPolice, C. C., Carter, G. W., & Johnson, J. W. (2005, April). *Linking O*NET to occupational literacy requirements using job component validation*. Poster presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles.

Lawshe, C. H. (1985). Inferences from personnel tests and their validity. *Journal of Applied Psychology, 70,* 237-238.

McCloy, R. A. (1994). Predicting job performance scores without performance data. In B. F. Green & A. S. Mavor (Eds.), *Modeling cost and performance for military enlistment: Report of a workshop*. Washington, DC: National Academy Press.

McCloy, R. A. (2001, April). *Predicting job performance scores in jobs lacking criterion data*. In J. W. Johnson & G. W. Carter (Chairs), Advances in the application of synthetic validity. Symposium conducted at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

McCloy, R. A., Putka, D. J., & Gibby, R. E. (2010). Developing an online synthetic validation tool. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*, 366-370.

McCoy et al. v. Willamette Industries (2002). United States District Court for the Southern District of Georgia, Savannah Division.

McCormick, E. J. (1959). The development of processes for indirect or synthetic validity: III. Application of job analysis to indirect validity. A symposium. *Personnel Psychology, 12,* 402-413.

McCormick, E. J., DeNisi, A. S., & Shaw, J. B. (1979). Use of the Position Analysis Questionnaire for establishing the job component validity of tests. *Journal of Applied Psychology, 64*, 51-56.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1972). A study of job characteristics and job dimensions based on the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology, 56*, 347-368. (Monograph)

Mossholder, K. W., & Arvey, R. D. (1984). Synthetic validity: A conceptual and comparative review. *Journal of Applied Psychology, 69*, 322-333.

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

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65*, 373-406.

Peterson, N. J., Wise, L. L., Arabian, J., & Hoffman, G. (2001). Synthetic validation and validity generalization: When empirical validation is not possible. In J. P. Campbell & D. Knapp (Eds.), *Exploring the limits of personnel selection and classification* (pp. 411-451). Mahwah, NJ: Erlbaum.

Primoff, E. S. (1957). The J-coefficient approach to jobs and tests. *Personnel Administrator, 20*, 31-40.

Primoff, E. S. (1959). Empirical validation of the J-coefficient. *Personnel Psychology, 12*, 413-418.

Scherbaum, C. A. (2005). Synthetic validity: Past, present, and future. *Personnel Psychology, 58,* 481-515.

Schmidt, F. L., Hunter, J. E., Croll, P. R., & McKenzie, R. C. (1983). Estimation of employment test validities by expert judgment. *Journal of Applied Psychology, 68,* 590-601.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1982). Progress in validity generalization: Comments on Callender and Osburn and further developments. *Journal of Applied Psychology, 67,* 835-845.

Society for Industrial and Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures.* Bowling Green, OH: SIOP.

Steel, P. D. G., Huffcutt, A. I., & Kammeyer-Mueller, J. (2006). From the work one knows the worker: A systematic review of the challenges, solutions, and steps to creating synthetic validity. *International Journal of Selection and Assessment, 14*, 16-36.

Taylor v. James River Corporation (1989). CA 88-0818-T-C (TC) (S.D. AL, 1989).

Trattner, M. H. (1982). Synthetic validity and its application to the Uniform Guidelines validation requirements. *Personnel Psychology, 35,* 383-397.

Varca, P. E., & Pattison, P. (1993). Evidentiary standards in employment discrimination: A view toward the future. *Personnel Psychology, 46,* 239-258.

Wainer, H. (1978). Sensitivity of regression and regressors. *Psychological Bulletin*, *85*, 267-273.

Waters, S. D., Russell, T. L., Shaw, M. N., Allen, M. T., Sellman, W. S., & Geimer, J. L. (2009). *Development of a methodology for linking ASVAB content to military job information and training curricula* (HumRRO Final Report FR-09-64). Alexandria, VA: Human Resources Research Organization.

Wilk, S. L., Desmarais, L., & Sackett, P. R. (1995). Gravitation to jobs commensurate with ability: Longitudinal and cross-sectional tests. *Journal of Applied Psychology, 80,* 79-85.

Appendix A
Armed Services Vocational Aptitude Battery (ASVAB) Standards: Special Warfare Operations (SO) SEAL Program to Rating

Armed Services Vocational Aptitude Battery (ASVAB) Standards: Special Warfare Operations (SO) SEAL Program to Rating

Janet D. Held
Navy Personnel Research, Studies, and Technology
NPRST (BUPERS-1) Millington, TN

Reviewed and Released by
David M. Cashbaugh
August 11, 2011

Armed Services Vocational Aptitude Battery (ASVAB) Standards: Special Warfare Operations (SO/SEAL), from Program to Rating

Janet D. Held
NPRST (BUPERS-1) Millington, TN

Introduction

The Chief of Naval Operations (N132G), at the request of the Director, Naval Special Warfare Recruiting Directorate within Navy Special Warfare Center (NAVSPECWARCEN), tasked Navy Personnel Research, Studies, and Technology (NPRST/BUPERS-1) to review the Armed Services Vocational Aptitude Battery (ASVAB) standards for adequacy in screening Navy SEAL (Sea, Air, and Land Special Warfare Combat Forces) candidates on cognitive ability. The last SEAL ASVAB standards study was conducted in 2004 when SEAL was a program open principally to specified Navy source ratings. In 2006 SEALs became one of three Special Operations (SO) ratings open principally to new accessions.[1] Some characteristics of SEAL candidates before and after the program to rating change, such as work and educational experiences, may have impacted the effectiveness of the current SEAL ASVAB standards.

This study, conducted to evaluate the current (operational) SEAL standards, also addressed a major SEAL community concern. The concern is with the former ASVAB test, Coding Speed (CS), now a Navy special test that is used in one of the two SEAL alternative ASVAB standards (and several other Navy rating ASVAB standards). A study commissioned by the SEAL community in 2010, using a large 10+ year data set, supported the ASVAB alternative composite that contains CS, but not the other.[2] The SEAL community is concerned that not all Navy applicants are administered CS, therefore the SEALs must rely on a suboptimal composite as their primary cognitive screen.[3]

The ASVAB, a measure of cognitive ability, should, theoretically, have little if any validity in predicting performance outcomes that are purely physical and stamina based. BUD/S 1st Phase, which incorporates the notoriously challenging "Hell Week", is mainly physically/stamina based but must contain a cognitive component because it is supported by at least two studies (the 2004 and 2010 referenced studies). It is difficult to determine what that cognitive component is because there was only one performance variable (criterion) available in each study - BUD/S 1st Phase "completion rate". Ideally, there should be gradations of performance adequacy measured not only in BUD/S training, but post BUD/s on the critical SEAL job related dimensions. At best, this study assumes that a relatively high degree of critical thinking and problem solving skills are required for every SEAL and that complicated field decisions and actions occur on a day to day basis. This study, therefore, had a narrow focus, conducted outside the framework of the multiple SEAL screening hurdles, to fulfill the immediate key study objectives.

---

[1] Held, Janet D. & Farmer, William L. (2004). *Armed Services Vocational Aptitude Battery (ASVAB) Standards: Basic Underwater Demolition /SEALs* (NPRST Letter Report Ser 3900, PERS-13/000056 9 Sep 2004).

[2] *"Follow on Research Findings"* submitted by the Gallup Consulting, Inc. in September 2011 to Director, Naval Special Warfare Recruiting Directorate, NAVSPECWARCEN, San Diego, CA.

[3] CS is only ministered at a Military Entrance Processing Stations (MEPS) directly after the computerized version of the ASVAB (CAT-ASVAB). About 65% of all Navy applicants test at the MEPS, while the other 35% test at an outlying Military Entrance Testing Site (METS) that administers paper and pencil versions of the ASVAB.

The objectives of this study for the SEAL community were to (a) provide a floor on cognitive ability recognizing that other uncorrelated SEAL screening measures (physical/stamina and personality/temperament) could be as important, if not more,[4] (b) resolve the dilemma regarding Coding Speed's limited administration, and (c) establish appropriate ASVAB standards that improve the BUD/S 1st Phase completion rates, if possible.

The ASVAB is the enlisted selection and primary classification instrument used by all of the military services. Table 1 gives a brief description of the nine ASVAB tests and also the former ASVAB Coding Speed (CS) test that is now an official Navy special classification test.

Table 1
Description of the ASVAB and Coding Speed Tests

Test Name and Abbreviation	Test Description
General Science (GS)	Knowledge of physical and biological sciences
Arithmetic Reasoning (AR)	Ability to solve arithmetic word problems
Word Knowledge (WK)[a]	Ability to select the correct meaning of words presented in context and correct synonyms
Paragraph Comprehension (PC)[a]	Ability to obtain information from written passages
Mathematics Knowledge (MK)	Knowledge of high school mathematics principles
Electronics Information (EI)	Knowledge of electricity and electronics
Auto and Shop Information (AS)	Knowledge of automobile and shop technologies, tools, and practices
Mechanical Comprehension (MC)	Knowledge of mechanical and physical principles
Assembling Objects (AO)[b]	Ability to determine correct spatial forms from their separate parts and connection points
Coding Speed (CS)[b]	Ability to quickly identify correct word/number pairings from a key with many options

[a]WK and PC are combined to form the Verbal (VE) composite that is a component of the AFQT and several Navy ASVAB classification composites. [b]Not all recruits enter the Navy with AO and CS test scores. CS is only given at the MEPS at the end of the CAT-ASVAB. AO is not given to high school students taking the paper and pencil ASVAB.

Each ASVAB and CS test have scores referenced to the ASVAB normative youth population (Profile of American Youth, 1997, or PAY97) and standardized to have a mean score of 50 and standard deviation (SD) of 10.[5] The bulk of ASVAB test scores typically are in the range of 20 to 80. The WK and PC tests are combined to form the Verbal (VE) composite (also with mean 50 and SD 10).VE is part of the Armed Forces Qualification Test (AFQT) used to qualify military applicants for service (2VE+AR+MK) and is scaled as a uniform percentile distribution with scores ranging from 1-99. The PAY97 youth population is considered the reference population for this study, for which, theoretically, future selection and classification decisions would be made.

[4] Since establishment of the SO SEAL rating, the SEAL community has developed, fielded, and operationalized multiple screening stages that include physical/stamina and personality/temperament measures, fitness preparation, mentoring, and full package reviews, all of which have improved the SEAL selection process.
[5] Segall, D. O. (2004). *Development and Evaluation of the 1997 ASVAB Score Scale* (Technical Report 2004-02. Seaside, CA: Defense Manpower Data Center. www.official-asvab.com/docs/asvab_techbulletin_2/pdf.

Coding Speed (CS), the former ASVAB test, was retained by the Navy as a special classification test because it demonstrated classification utility by (a) increasing the predictive validity of the ASVAB in determining who passes or fails training, (b) increasing the proportion of a recruit population qualified for Navy ratings, and (c) increasing qualification rates for women and some minor groups helping to improve diversity. Recently it was hypothesized that CS test has motivational underpinnings, which would be highly relevant for SEALs.[6] CS is currently being revised by Defense Manpower Date Center (DMDC) – the Personnel Testing Division.[7]

Table 2 lists the Navy's operational ASVAB composites and the two that contain CS.

Table 2
ASVAB and ASVAB/CS Classification Composites

Composite Tests	Composite Names
General Technical	VE+AR
Administration	VE+MK
Hospitalman	VE+MK+GS
Electronics	AR+MK+EI+GS
Basic Electricity & Electronics	AR+2MK+GS
Nuclear Field	VE+AR+MK+MC
Engineering	VE+AR+MK+AS
Special Operations	GS+MC+EI
Mechanical	AR+MC+AS
Mechanical_2	MK+AS+AO
Operations	VE+AR+MK+AO
Business/Clerical	VE+MK+CS
Air Traffic Control	VE+MK+MC+CS

Note. A composites in operational use that contains AO or CS is considered an alternative to a primary composite that does not contain either test because not all Navy recruits have AO and CS scores.

The SEAL ASVAB standard prior to the 2004 NPRST study was VE+AR ≥ 104"plus" MC ≥ 50. This "multiple cutscore" ASVAB standard was replaced in 2004 with the "alternative standards", GS+MK+EI ≥ 165 "or" VE+MK+MC+CS ≥ 220. The GS+MK+EI composite is considered the primary composite because not all Navy recruits have CS scores to derive a VE+MK+MC+CS composite score. The GS+MC+EI composite demonstrated validity in the 2004 study whereas VE+AR did not, which was conjectured to have occurred because, to some extent, the technical tests measure underlying hands on experience dimensions related to SEAL success.

[6] Segal, Carmit (2010). *Motivation, Test Scores, and Economic Success*. Department of Economics and Business paper, Universitat Pompeu Fabra. http://www.econ.upf.edu/~segal/SegalMotivationTestScoresJuly2010.pdf.

[7] DMDC, executive agency for ASVAB development and maintenance, is reworking the CS test to address speeded test issues that relate to computer hardware component replacements that then require special score equating efforts. The revised test, called Processing Speed (PS), intends to be a purer measure of processing speed, will be scored by Item Response Theory (IRT) to allow more accurate measurement of ability, and will incorporate automated item generation eliminating the need for resource intensive new form development procedures.

Methods, Analyses, and Results

The study is organized into the following sections: (1) Data collection; (2) Composite validation; (3) Cutscore development; (4) Theoretical based cutscore analysis; (5) Empirical/Theoretical cutscore comparison; (6) ASVAB waiver analysis; and (7) ASVAB/CS standards models. Separate summary/conclusion and recommendations sections follow.

Data Collection

The study data were obtained from the Navy Integrated Training Resources and Administration System (NITRAS) database and were for six BUD/S 1st Phase classes that started and finished in the 2010 calendar year (CY10). Data were retained only if SEAL candidates (a) had active duty start dates in the CY08 or later, (b) reported to BUD/S 1st Phase without having Fleet duty, and (c) either completed the class or dropped on request (DOR). That is, cases with medical, administrative, or other reasons for dropping were not included. The data curtailment measures established a somewhat homogenous data set without many of the extraneous factors (but not all) that could impact a candidate's likelihood of completing the training.[8]

The data were grouped into three loosely defined seasons (Jan/Mar, May/Jun, and Aug/Sep) to study seasonality effects most likely to occur from harsh winter months, compared to more moderate spring and summer months. However, the grouping did not completely accomplish the intent because the highest drop points occur during "Hell Week", which involves vigorous ocean and beach exercises, occurs in about the 5th week of BUD/S 1st Phase. The position of "Hell Week" would categorize a September class start date, for all intents and purposes, as a winter month. However the seasons were retained due to the constraint of the FY10 time frame and the unavailability of the September final class outcomes. Table 3 gives key data characteristics for the three seasonal groups.

Table 3
Data Characteristics for Six CY10 BUD/S 1st Phase Classes Grouped by Seasonality
(Total N = 633 including setbacks; completion rate = 33.3%)

BUD/S 1st Phase Class Start Date	Sample Size	AFQT Mean/ Median	Non-Academic Setback Rate	Completion Rate
January & March classes (01/25/2010 & 03/15/2010)	199	77.1 / 80	18.6%	39.7%
May & June classes (05/03/2010 & 06/23/2010)	240	77.0 / 79	6.2%	31.3%
August & September classes (08/11/2010 & 09/21/2010)	194	77.1 / 80	6.7%	29.4%

Note. A statistical test was performed to determine if completion rates differed for the two most extreme seasons (Jan/Mar vs. Aug/Sep). The results were statistically significant at the .05 probability level, (Pearson Chi-Square = 4.62, p =.032, df = 1).

[8] For the data ultimately used to establish ASVAB/CS validity, marital status, age, and years of education were not statistically significant in predicting BUD/S 1st Phase completion; however setback rate was significant with a higher incidence negatively related to completion. A positive correlation between age and years of education was statistically significant, which was expected given older candidates have more time to obtain higher education.

Table 3 shows two points of interest. First, the 39.7% winter class completion rate (Jan/Mar) is higher than the 31.3% for spring classes (May/Jun) and 29.4% for summer classes (Aug/Sep), an unexpected outcome if the theory for winter months is for a negative impact on performance. A possible explanation could be that instructors give extra time and attention to prepare candidates in the winter, which boosts the propensity to succeed throughout the "Hell Week" period. The more plausible explanation given by community members at the training site is that candidates in the winter month classes were extraordinarily physically fit. The influence of physical fitness is a reasonable explanation over ASVAB differences given the AFQT averages across the three seasonal groups were essentially the same (77.1, 77.0, and 77.1 respectively) and age and other factors evaluated were not significant, except setback incidence (see note to Table 3 and footnote 8).

Regarding setback incidence, the second point of interest in Table 3 is the approximately three times higher setback rate for the winter classes than for the more moderate spring and summer classes (18.6% setback rate for Jan/Mar compared to 6.2% for May/Jun, and 6.7% for Aug/Sep). No explanation was conjectured for the higher winter class start setback rate incidence other than extraordinary levels of persistence may be related to both overcoming a setback and attaining high levels of physical fitness.

Table 4 gives the recalculated BUD/S 1st Phase characteristics for the Table 3 data with the setback cases removed.

Table 4
Data Characteristics for Six CY10 BUD/S 1st Phase Classes by Seasonality
(Total N = 568 excluding setbacks; completion rate = 32.6%)

Class Start Date	Sample Size	AFQT Average/Median	Completion Rate
January & March classes (01/25/2010 & 03/15/2010)	162	77.5 / 81	37.0%
May & June classes (05/03/2010 & 06/23/2010)	225	76.7 / 79	31.1%
August & September classes (08/11/2010 & 09/21/2010)	181	78.8 / 81	30.4%

Note. A statistical test was performed to determine if completion rates differed for the two most extreme seasons (Jan/Mar vs. Aug/Sep. The results were not statistically significant at the .05 probability level (Pearson Chi-Square = 3.60, p =.058, df = 1).

Table 4 shows a slightly lower 37.0% completion rate for the winter classes with setbacks removed than the 39.7% (Table 3) with setbacks included. The 2.7% completion rate difference appears significant, but was not statistically significant at the .05 probability level. The test for completion rate differences between winter and summer data (the most extreme seasons) was considered marginally significant (see Note in Table 4).

The conclusion from examining the data in this section was that, for the study's validity analyses, only the spring and summer months should be included, and with setback cases removed. The final data set contained 335 cases.

Composite Validation

The objective of an ASVAB validity analysis is to estimate which of a set of ASVAB tests (composite) is most predictive of performance. Validity as it applies to Navy ASVAB validation/standards studies refers to the correlation between scores on a particular ASVAB composite with scores on the school performance measure, which is typically A-School final course grade. Validity coefficients range from -1 to +1, for perfect negative and positive relationships, respectively. A zero validity coefficient means there is no predictive relationship. Some correlation coefficients may appear to have magnitude greater than zero, but due to sampling error that relates to sample size (and other factors), they do not. The sample size for this SEAL study (N = 335) was considered adequate for correlation/validity analysis.

Among the many factors that impact the magnitude of the validity coefficient is the one of most concern for this study - the restriction of range in ASVAB scores that occurs from applying an ASVAB composite cutscore. Restriction in range of ASVAB scores lowers score variance, and because score variance is necessary to derive a correlation, it is suppressed from what would be observed for a full ASVAB range population. And, it is the applicant population (not the school sample) for which selection and classification utility will be assessed and cutscore decisions made. The average ASVAB composite validity coefficient across Navy rating training, corrected only for range restriction, is about .55. The smallest ASVAB composite validity coefficient is about .25 for SEALs and Navy Divers, which both have large physical components. The largest validity is about 0.85 for the Nuclear Field ratings, which have a large academic training component.

For illustrating full range validity (taken as the Navy's .55 average for the example), Figure 1 displays a notional bivariate normal distribution with selection instrument scores on the *x*-axis plotted against performance instrument scores on the *y*-axis.

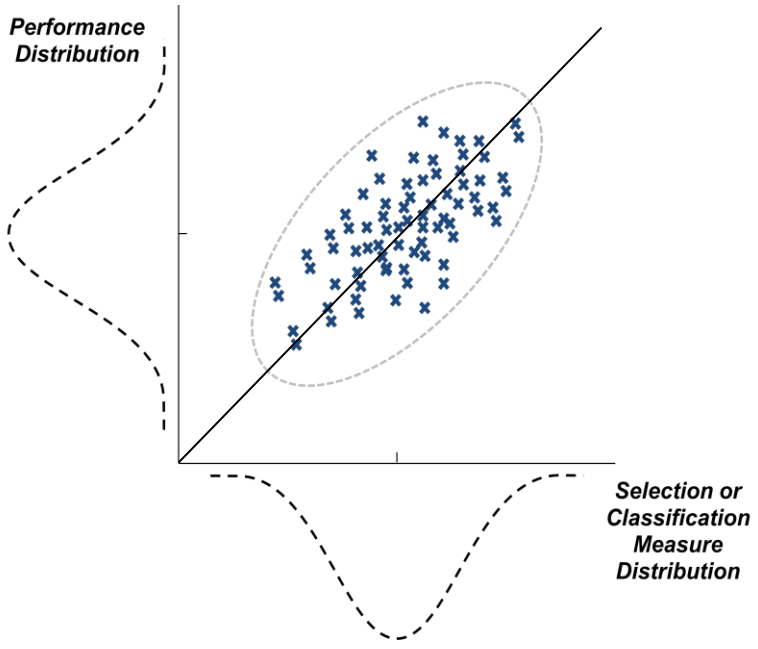

Figure 1
Hypothetical Bivariate Normal Distribution with a Correlation (Validity) of .55

Figure 1 also displays a notional least squares regression line that minimizes the average errors in predicting performance scores across the total ASVAB score range. The assumption for estimating the "population" validity (correcting for range restriction) is that this line extends through the total range of ASVAB scores even though it is developed only in an ASVAB range restricted sample for which performance scores are available. Given certain assumptions hold about the observed sample data and the unobserved population data, it is appropriate to "extrapolate" the sample relationship (validity) to the population.[9] (Reference 10 explains the multivariate correction for range restriction for an ASVAB application.)[10]

Figure 2 can be used to illustrate the utility of a validity coefficient that applies to an applicant population (assumed to be the PAY97 ASVAB normative population from here on out) that would have a 30% qualification rate (somewhat reflective of the SEAL ASVAB standard).

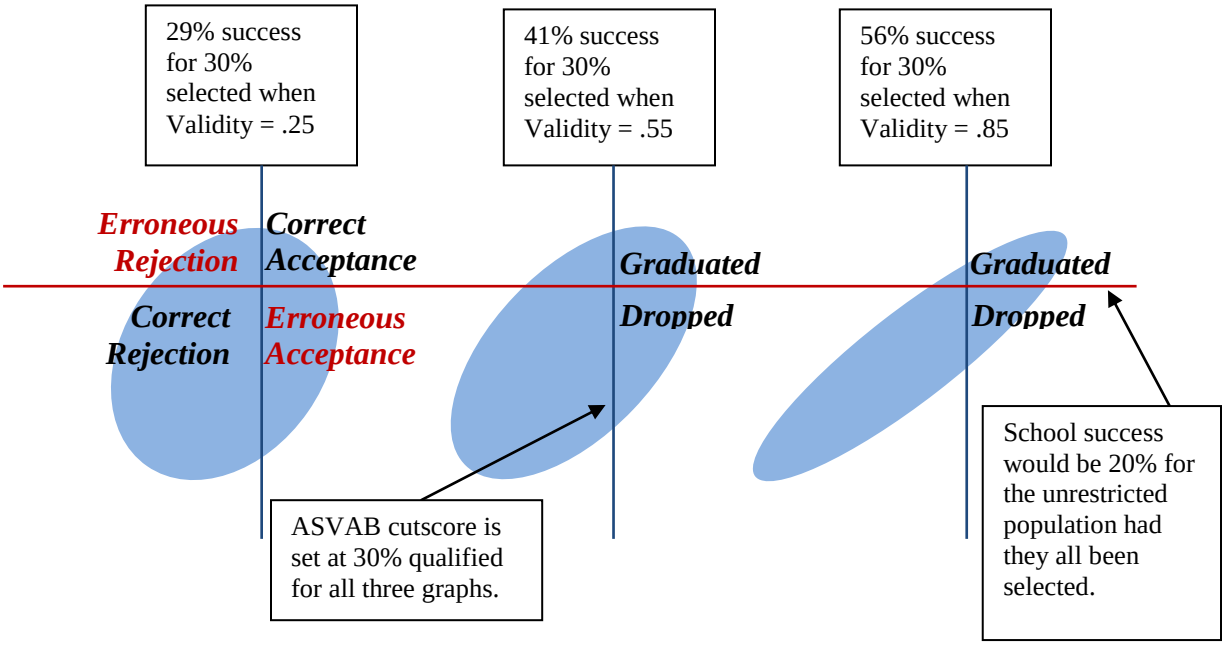

Figure 2
Notionally, Navy School Graduation Rates as a Function of Validity

The three graphs in Figure 2 represent three validities, from left to right, .25 (historical for SEALs), .55 (average for Navy), and .85 (highest for Nuclear Field). The ASVAB cutscore (ASVAB scores on the *x*-axis) for all three graphs are notionally set at 30% qualified (the dark blue vertical lines). The performance bars for graduation (scores are on the *y*-axis) are all set at 20% (the red horizontal line). The text in the box above each graph provides the 30% qualification rate, which is a constant, the varying validities (.25, .55. and .85), and the graduation rates (for those selected for a school with ASVAB scores to the right of the vertical bar) that result from the validity magnitude (29%, 41%, and 56%, respectively). All other things being equal, validity magnitude alone determines the success rates (discussed later with regards to a Taylor-Russell table analysis).

[9] Lawley, D. (1943). A note on Karl Pearson's selection formulae. *Royal Society of Edinburgh, Proceedings, Section A, 62*, pp. 28-30.
[10] Held, J. D. & Foley, P. P. (1994). Explanations for Accuracy of the General Multivariate Formulas in Correcting for Range Restriction. *Applied Psychological Measurement, 18(4)*, pp. 355-367.

In Figure 2, raising the cutscore for each graph (say, notionally, 10 ASVAB composite score points) results in different success rate improvements, the largest improvement associated with the highest validity. For zero validity (visualized as a circle), there would be no success rate improvement no matter how high the cutscore, but a significant reduction in qualified applicants.[11]

The classification decision accuracy regions from raising or lowering the ASVAB cutscore are shown for the .25 validity graph in Figure 2. Classification decision accuracy, however, is not taken solely as the percentage of applicants in each of the four quadrants, but is viewed in the context of a rating's parameters and context. For Nuclear Field, the ASVAB validity is so high, the curriculum so difficult, and the yearly goal so high (about 2,500), that lowering the ASVAB cutscore 10 points in response to a difficult recruiting environment will result in a much higher non-graduation rate, resulting in the need to increase recruiting resources to replaced failed students. On the other hand, raising the Nuclear Field ASVAB cutscore by 10 points will result in a substantial improvement in the training graduation rates but with the risk of not being able to meet yearly recruiting goals and thus Fleet requirements. Obviously there are many interactions related to ASVAB standards.

When Navy training graduation rates are low, policy makers can advocate improving or extending the training so that more students will graduate. But, because modifications to training are costly and impose many negative ripple effects (hiring more instructors, rescheduling classes that impact incoming students, modifying supply chain databases, and shortening Sailors periods of productivity in first term), the usual first of several steps is to conduct an ASVAB standards study. The major objective of the study is to identify a single most valid ASVAB composite in predicting training performance, and thus potentially to solve any academically related graduation rate problems, possibly without a cutscore raise if the composite is substantially more valid than the one that it replaces. ASVAB composites, however, that are relevant to the training curriculum can be highly correlated, so replacing one with another that demonstrates higher validity may only result in limited graduation rate improvement. Within the Navy context, a .05 validity increment is considered meaningful and typically results in about a 2% increase in the graduation rate, but may not be very meaningful enough by itself to solve a specific graduation rate issue.

The ASVAB composites compared in the validity analyses in this study were the operational GS+MC+EI and VE+MK+MC+CS composites and the former SEAL composite, VE+AR. The MC (Mechanical Comprehension) test was also validated because it was used as an adjunct to the VE+AR composite in the former SEAL ASVAB standard. The Navy VE+AR+MK+AO composite that contains Assembling Object (AO) was also included in the validity analyses because spatial ability appears, at face value, to be an aptitude/ ability relevant in field maneuvers.

The performance measure for the validity analyses was BUD/S 1st Phase completion, which was expressed as a binary 1/0 variable. The appropriate validity analysis for a binary outcome variable is logistic regression. BUD/S "Completion", however, was not viewed strictly as a dichotomous variable, rather as a continuous multivariate distributed variable with scores influenced by such factors as personal goals, emotional stability, physical and mental stamina, family support systems, and access to other employment opportunities, none of which were captured in this study.[12]

[11] A school sample can be conceptualized as a random sample taken from qualified applicants to the right of the vertical cutscore bar.

[12] May 2011 communication with Norman M. Abrahams, Ph.D., former NPRDC principal investigator for the Navy's Annapolis Academy selection system.

Table 5 provides the logistic regression results for the targeted ASVAB composites and the ASVAB MC test.

Table 5
Logistic Regression Results for ASVAB/CS Composites:
Spring & Summer Classes Combined without Setbacks (N = 335)

Composite	Slope & Intercept Probability of Significance	Chi Square Overall Significance	Nagelkerke R-Square & R
GS+MC+EI	.008/.001	.008	.030 (R=.17)
VE+MK+MC+CS	.001/.000	.001	.049 (R=.22)
VE+AR	.001/.000	.001	.047 (R=.22)
VE+AR+MK+AO	.001/.000	.001	.050 (R=.22)
MC	.009/.003	.018	.023 (R=.15)

Note. All statistical tests were significant at the .05 probability level.

Table 5 shows statistically significant results for the logistic regression equation slopes and intercepts for all ASVAB composites and the single MC test. The Chi-Square test was significant for rejecting the null hypothesis that there is no differences in performance outcome (completion rate) associated ASVAB composite scores. The validities, as taken from the square root of the Nagelkerke R-Square, were of comparable magnitude to those reported in the 2004 SEAL ASVAB study, and the .17 validity for GS+MC+EI was an exact replication (reference 1). The .17 validity was .05 lower than the .22 validity for the three other composites. The one difference in this study compared to the 2004 study was that VE+AR in 2004 had zero (.00) validity.

Table 6 gives the validity results from the correction for range restriction procedure treating the dichotomous SEAL outcome variable as if it were appropriate and also, for comparison, the logistic regression procedure.[13]

Table 6
Validities Corrected for Range Restriction
Compared to the Logistic Counterpart

Composite	Validities Corrected for Range Restriction using the N = 335 Sample	Logistic Regression Nagelkerke R-square using the N = 335 Sample
GS+MC+EI	.29	.17
VE+MK+MC+CS	.34	.22
VE+AR	.34	.22
VE+AR+MK+AO	.34	.22
MC	.26	.15

Note. No statistical test was performed for the validities corrected for range restriction.

[13] A correction for range restriction was unknown for logistic regression but conceptualized as appropriate for correlating ASVAB with the binary outcome as it was expected that all composite validities would be impacted in the same manner (thus not biasing their rankings).

Table 6 shows, as expected; the validities derived for the correction for range restriction procedure were higher than those derived by logistic regression, but that both procedures identified GS+MC+EI as having the lowest validity (.29 for the correction for range restriction procedure and .17 for logistic regression). Also, both procedures identified the other three composites, VE+MK+MC+CS, VE+AR, and VE+AR+MK+AO as having equal validity (.34 for the correction for range restriction and .22 for logistic regression), and also .05 incremental validity when compared to GS+MC+EI validity (.34 - .29 = .05 for the correction procedure and .22 - .17 = .05 for logistic regression). The VE+AR+MK+AO composite, of high interest for a follow-on study, was dropped for further consideration in this study because scores, like VE+MK+MC+CS, are not available for all recruits.

Approximations for the composite validities were taken as convenient values (.25 and .30) close to those derived from the two validity estimation procedures but that maintained the .05 validity difference (found from each procedure).

Cutscore Development

In order to conduct a cutscore analysis for candidate ASVAB composites, the cutscores for each composite must be equivalent. There are several ways to define "equivalent" and for this study it was defined as the same level of aptitude/ability as partitioned on the normal curve. For illustration purposes, Figure 3 depicts a single ASVAB test score distribution with scores standardized to have a mean (average) of 50 and standard deviation of 10, again, referenced to the PAY97 population. The mean of 50 is displayed along with three standard deviation (SD) points above it (60, 70, and 80) and below (40, 30 and 20). The SD is a measure of departure from the mean in test score units and has meaning for comparing individuals in terms of aptitude/ability levels.

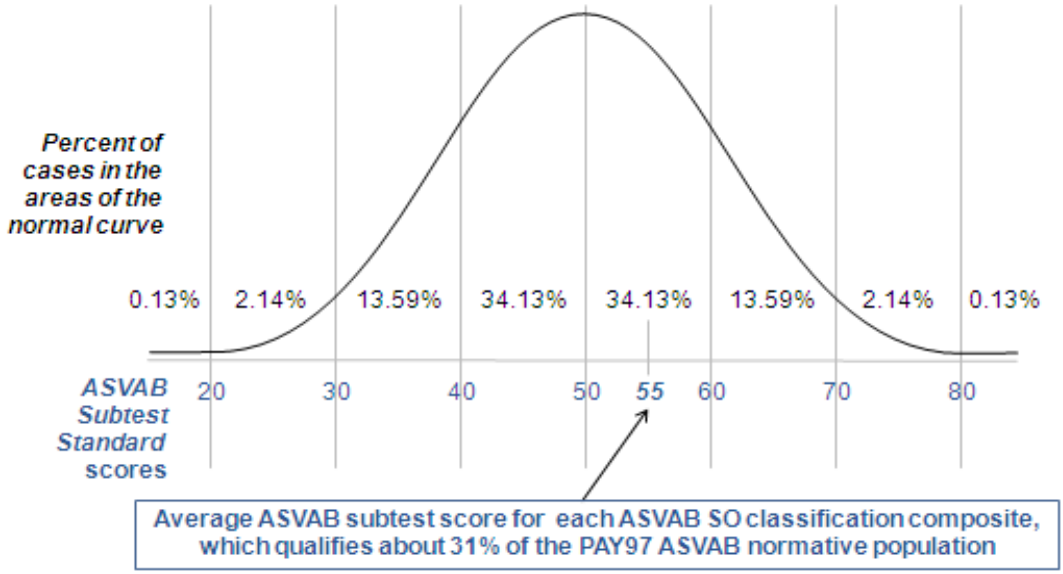

Figure 3
Notional ASVAB PAY97 Test Score Distribution and Areas within Standard Deviation Partitions

The arrow in Figure 3 points to the position of the 55 ASVAB test score on the normal curve *x*-axis. The 55 score multiplied by 3 equals the operational 165 cutscore for the SEAL's primary operational GS+MC+EI composite. The 55 score is 5 score points higher than the mean of 50, which in the standard deviation (SD) metric is 5/10, or .50 SDs. A normal curve table (in the appendices of many statistics text books) shows that .50 SDs above the mean marks the point at which the density in the upper end of the curve is .3085. Or, for the study's purpose, 31% of PAY97 ASVAB youth scoring at or above a "cutscore" of 55 would be considered ASVAB qualified for SEAL, if indeed a single ASVAB test were used for Navy rating classification.[14]

Navy composite scores are computed as the sum of the ASVAB composite test scores. Given 50 is the PAY97 ASVAB normative mean (average) for all ASVAB tests, the mean for a 2-test composite is simply 2 X 50 = 100; a 3-test composite mean is 150, and a 4-test composite mean is 200. However, an ASVAB composite's SD (which is used in this study to establish comparable cutscores) is not simply the sum of that composite's individual tests' standard deviations. The ASVAB composites correlate to varying extents (to the degree that they overlap in measured abilities) so there is a reduction in the composite standard deviation to account for the correlation. The composite SDs for the PAY97 ASVAB youth population, estimated by the same correction for range restriction procedure used to estimate the composite validities, but that are actually known, are: (a) 26.96 for GS+MC+EI; (b) 31.79 for VE+MK+MC+CS; and (c) 18.64 for VE+AR.

The three cutscore levels developed for each composite applying their standard deviations (above the mean) reported in the previous paragraph are: (a) .56 SD qualifying 28.8% of PAY97 youth; (b) .75 SD qualifying 22.7%; and (c) 1.00 SD qualifying 15.9%. The .56 SD applies to the operational 165 cutscore for the GS+MC+EI composite whereas the .75 and 1.00 SDs were merely subjective estimates of potentially relevant SEAL ASVAB aptitude/ability requirements.

Table 7 lists the three PAY97 qualification rates (replicated for each composite), the actual number/percentage of the FY10 accession population qualified, and also the number/percentage of the 1,141 FY10 SEAL accessions disqualified, which is only informative for the SEAL community in that those disqualified would have had to have been replaced. A major consideration in evaluating the contents of Table 7 is that, given not all accessions have VE+MK+MC+CS scores, a lower number/percentage of FY10 accessions qualified on this composite at any particular cutscore in no way indicates the higher aptitude/ability normally associated with "selection stringency". Another consideration is that even though the composites reflect a degree of unique constructs, they are all highly correlated and therefore there is overlap, to a large extent, of qualified individuals across composites at any particular cutscore level.[15]

Finally, Table 7 includes the three AFQT scores that are comparable to the composites' three cutscores so that the SEAL community can baseline their candidates each year to the current exceptional SEAL recruiting environment. (Providing an AFQT analysis is not meant to suggest that AFQT be used for SEAL classification as it is reserved for military eligibility as a general measure of trainability.)

---

[14] As discussed earlier for the MC test, the Navy does not develop individual ASVAB subtest cutscores although several ratings have them, including the Explosive Ordnance Disposal (EOD) rating, a Special Operations rating, and , in the past, SEAL. MC was not carried further for cutscore analysis but was considered for a floor cutscore in conjunction with VE+AR only to provide balance to a purely academic measure.

[15] The PAY97 correlations between the three composites were .84 between GS+MC+EI and VE+MK+MC+CS, .81 between GS+MC+EI and VE+AR, and .91 between VE+AR and VE+MK+MC+CS.

Table 7
Qualification Rates for Comparable Composite Cutscores Applied to the PAY97 ASVAB Youth, FY10 Accessions, and Disqualified for FY10 SEAL Accessions

Composites & Cutscores	PAY97 ASVAB Population Qualification Rate (of 20+ Million)	FY10 Accession Population Qualification Rate (of N = 37,084)	FY10 SEAL Accessions Disqualified (of N = 1,141)
GS+MC+EI (Operational)			
165	28.8%	16,946 = 45.7%	202 = 17.7%
170	22.7%	13,771 = 37.1%	354 = 31.0%
177	15.9%	9,594 = 25.9%	570 = 50.0%
VE+MK+MC+CS (Operational – not applied to Accessions without CS scores)			
218	28.8%	13,305 = 35.9%	414 = 36.3%
224	22.7%	10,365 = 28.0%	523 = 45.8%
232	15.9%	6,883 = 18.6%	682 = 59.8%
VE+AR (Candidate)			
110	28.8%	17,014 = 45.9%	300 = 26.3%
114	22.7%	12,355 = 33.3%	464 = 40.7%
119	15.9%	7,592 = 20.5%	665 = 58.3%
AFQT (for benchmarking)			
70	29.6%	Non-applicable for qualification	Non-applicable for qualification
77	22.7%		
84	16.2%		

Notes. (1) All 37,084 FY10 accessions had GS+MC+EI and VE+AR scores and 25,006 had VE+MK+MC+CS scores. (2) The standard deviations corresponding to 28.8%, 22.7%, and 15.9% qualified are .56, .75, and 1.00. (3) The three AFQT scores are comparable to the three levels of aptitude/ability used to derive the composite cutscores.

Table 7 shows, in the top row of the far right hand column, that the primary operational SEAL ASVAB standard, GS+MC+EI ≥ 165, disqualified, logically, the smallest number of FY10 SEAL accessions (202, or 17.7% of 1,141) compared to the higher cutscores). Some of those 202 unqualified SEAL candidates might have qualified on the alternative standard, VE+AR+MK+CS ≥ 220. Table 7 also shows that, with a 218 cutscore for VE+AR+MC+CS (comparable to the other composites' lowest level cutscores, but 2-score points lower than the operational 220 cutscore), 414, or 36.3% of the 1,141 FY10 SEAL accessions were disqualified, which is over twice as many as the 202 disqualified by the GS+MC+EI ≥ 165 standard. The larger number of disqualified by VE+AR+MC+CS ≥ 220 is consistent with GS+MC+EI ≥ 165 used by the SEALs as their primary ASVAB selection standard because not all Navy recruits have CS scores.

Table 7 also shows that 2,002 more FY10 accessions qualified on the most stringent 177 cutscore for GS+MC+EI (9,594 qualified) compared to most stringent 119 cutscore for VE+AR (7,592 qualified). The fewer VE+AR qualified at this high score level might mean that more academically inclined youth take the college option, compared to youth who are more interested and inclined to pursue the more mechanical and technical career options available in the Navy.

Finally, Table 7 shows, in the bottom block of rows, three AFQT scores (70, 77, and 84) that are comparable (as possible) to the three levels of composite cutscores developed for this study.

The AFQT scores qualified 29.7%, 22.7%, and 16.2% compared to the composites' 28.8%, 22.7%, and 15.9%, respectively. As already noted, the AFQT scores may be useful only for benchmarking (not classification) and therefore the analysis stopped at this point.

Theoretical Based Cutscore Analysis

One of the standard methods for evaluating the utility of a selection instrument involves use of the Taylor-Russell tables (TR-tables).[16] The theoretical based TR-tables are developed from mathematical functions that can generate any number of bivariate normal distributions. These distribution vary in their four parameters, briefly discussed earlier as (a) the population qualification rate (selection ratio) established by the particular selection instrument's cutscore, (b) the selection instrument's validity that applies to this population, (c) the observed success rate in the data analyzed (school success rate in our case), and (d) the success rate that would apply to the population had individuals been selected without having applied a valid selection instrument (not actually known but estimated).[17]

The fourth TR-table parameter is called the "base rate" and is the level of performance that is unknown for Navy because ASVAB standards (composites with cutscores) apply to all ratings. The preliminary objective in conducting a TR-table analysis is therefore to identify which of the base rate tables applies. Given the study produces three of the four TR-table parameters, the fourth, the base rate, is fixed so identifying the applicable base rate table simply involves finding which (10 tables are published) produces the study's observed success rate (internal entries in each table) at the intersection of the study's estimated population validity (listed in the column to the far left of each table) and qualification rate (listed as selection ratios across the top row heading).[18]

Table 8 on the next page shows a portion of the best fitting .20 base rate TR-table given the SEAL study's three parameters. Table 8 shows a 29% success rate (closest of all tables to the about 32% observed BUD/S graduation rate) at the intersection of (a) the about .25 validity coefficient estimated to apply to the primary operational GS+MC+EI composite and (b) a .30 selection ratio (closest to the 28.8% PAY97 ASVAB population qualification rate associated with the GS+MC+EI composite's 165 cutscore).

Table 8 shows that the 29% success rate applying a composite with .25 validity (i.e., GS+MC+EI) and a cutscore that qualifies 30% of the population (.30 selection ratio) can be expected to improve to a 30% success rate (only 1-percentage point) if the composite is replaced with one with .30 validity (i.e., VE+AR or VE+MK+MC+CS). The marginal improvement in the SEAL completion rate from replacing GS+MC+EI with a composite with .05 higher validity is due mainly to the ASVAB's low validity overall in predicting performance in a physically and mentally challenging training course, but also, a function of the parameters of the study that determine the Taylor-Russell base rate table.

[16] Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* pp. 565-587.
[17] The last parameter implies random selection.
[18] Absent the TR-table mathematical functions, the appropriate bate table in most cases is simply the table that best fits the study parameters.

Table 8
Portion of the Taylor Russell .20 Base Rate Table
that Best Fits the Study's Parameters

	Selection Ratio										
Validity	0.05	0.10	0.15	0.20	0.25	**0.30**	0.35	0.40	0.45	0.50	0.55
0.00	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
0.05	0.23	0.23	0.22	0.22	0.22	0.22	0.21	0.21	0.21	0.21	0.21
0.10	0.26	0.25	0.25	0.24	0.24	0.23	0.23	0.23	0.22	0.22	0.22
0.15	0.30	0.28	0.27	0.26	0.26	0.25	0.25	0.24	0.24	0.23	0.23
0.20	0.33	0.31	0.29	0.28	0.28	0.27	0.26	0.26	0.25	0.24	0.24
0.25	0.37	0.34	0.32	0.31	0.30	**0.29**	0.28	0.27	0.26	0.26	0.25
0.30	0.41	0.37	0.35	0.33	0.32	**0.30**	0.29	0.28	0.28	0.27	0.26
0.35	0.45	0.40	0.38	0.36	0.34	0.32	0.31	0.30	0.29	0.28	0.27
0.40	0.49	0.44	0.41	0.38	0.36	0.34	0.33	0.31	0.30	0.29	0.28
0.45	0.54	0.48	0.44	0.41	0.38	0.36	0.35	0.33	0.31	0.30	0.29
0.50	0.58	0.51	0.47	0.44	0.41	0.38	0.36	0.34	0.33	0.31	0.30
0.55	0.63	0.56	0.50	0.47	0.43	0.41	0.38	0.36	0.34	0.32	0.31
0.60	0.68	0.60	0.54	0.50	0.46	0.43	0.40	0.38	0.36	0.34	0.32
0.65	0.73	0.64	0.58	0.53	0.49	0.45	0.42	0.39	0.37	0.35	0.33
0.70	0.79	0.69	0.62	0.56	0.52	0.48	0.44	0.41	0.38	0.36	0.34
0.75	0.84	0.74	0.66	0.60	0.55	0.50	0.46	0.43	0.40	0.37	0.34
0.80	0.89	0.79	0.71	0.65	0.59	0.53	0.49	0.45	0.41	0.38	0.35
0.85	0.94	0.85	0.77	0.69	0.63	0.56	0.51	0.47	0.42	0.39	0.36
0.90	0.98	0.91	0.83	0.75	0.67	0.60	0.54	0.48	0.44	0.40	0.36
0.95	1.00	0.97	0.91	0.82	0.73	0.64	0.56	0.50	0.44	0.40	0.36
1.00	1.00	1.00	1.00	1.00	0.80	0.67	0.57	0.50	0.44	0.40	0.36

Note. The .25 validity for GS+MC+EI is in bold and also the .30 validities that pertain to the three other composites evaluated in this study, VE+MK+MC+CS, VE+AR, and VE+AR+MK+AO. The .29 and .30 (29% and 30%) success rates associated with the .25 and .30 validities are also in bold.

Table 8 can also be used to estimate expected improvements in the success rate by simply increasing the stringency of the cutscore. For our study we would want to increase the cutscore of the lower validity composite, GS+MK+EI, to achieve at least the same success rate as the alternative standard that uses VE+MK+MC+CS. Table 8 shows that in order to achieve a 30% success rate that is associated with the .30 validity, a composite with a.25 validity coefficient would require a cutscore that decreased the qualification rate (selection ratio) from 30% to 25%. In the case of the SEALs, becoming more stringent in selection based upon the ASVAB might well eliminate those candidates who have other more highly valued attributes that, in constellation with ASVAB and other factors, would succeed in becoming a SEAL. Table 8 shows that setting a cutscore to qualify only 10% of the population rather than 30% when the ASVAB composite validity is .30 results in a 37% success rate, a 7% improvement compared to 30% (at 30% qualified). A 7% improvement may not be enough to offset what might be a large increase in the classification error depicted in the upper left hand quadrant of Figure 2 (erroneous rejections). [19]

[19] Coincidentally, 37% is the BUD/S 1st Phase completion rate observed for the highly physically fit SEAL candidates that went through the classes in the winter months.

Empirical/Theoretical Cutscore Comparison

Table 9 contains the results of a GS+MC+EI cutscore analysis applying two methods. The first method was empirical based and used the study data. The second method was theory based and used the Taylor-Russell (1939) tables (reference 16).[20]

Table 9
BUD/S 1st Phase Completion Rates Applying GS+MC+EI
Cutscores in an Empirical and Theoretical Based Cutscore Analysis

Composite/ Cutscore	Grad Rates For the Study N = 335 Sample	Success Rates from Taylor Russell Table
GS+MC+EI ≥ 165	32.8% (85/259)	29.3%
GS+MC+EI ≥ 170	33.8% (76/225)	30.5%
GS+MC+EI ≥ 177	35.2% (58/165)	31.8%

Three cutscores, 165, 170, and 177 were evaluated for Table 9. Table 9 shows, consistent with a low magnitude validity coefficient, that BUD/S 1st Phase completion rates increase only marginally as the cutscores become more stringent, but also by about the same degree for each analysis (empirical versus theoretical). Figure 4 is a plot of the Table 9 data to illustrate the findings.

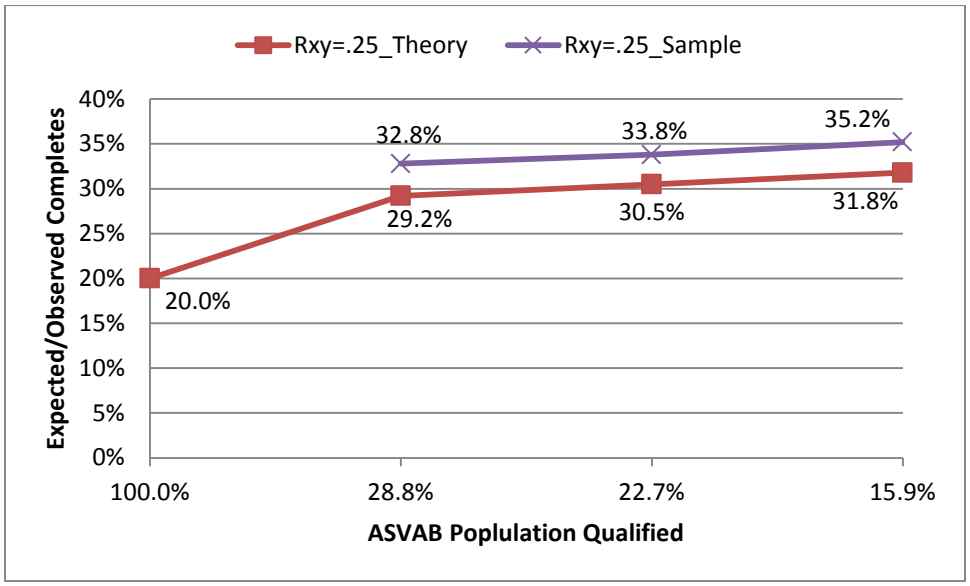

Figure 4
Observed BUD/S 1st Phase Complete Rates and Theory Based Rates
for .25 ASVAB Composite Validity

[20] An empirical ASVAB cutscore analysis is only appropriate for the operational selection instrument, not a correlated candidate replacement composite, because a floor aptitude/ability level has been set for individuals in the dataset by an applied operational cutscore.

Figure 4 shows that the two lines applying to the empirical based completion rates and theoretical based success rates for the GS+MC+EI composite with an estimated .25 validity are parallel across increasing stringent cutscores. The slightly higher observed completion rates for the empirical data compared to the TR-table success rates could be due to (a) inordinately dense scores across the higher GS+MC+EI score range relative to what appears in the normal curve distribution, most likely due to the positive recruiting environment or (b) a non-trivial correlation of the ASVAB with other factors that influence performance but are not accounted for in this study. The point is that improved completion rates due to increases in the GS+MC+EI cutscore are uniformly consistent over the cutscore levels for both the empirical and theoretical cutscore analyses (implying utility of the TR-tables for resetting the GS+MC+EI cutscore if the composite is to be retained for SEAL classification). Figure 5 applies strictly to the theoretical based TR-table cutscore analysis and is intended to provide an unbiased comparison of GS+MC+EI with a .25 validity and the other composites with the higher .30 validity.

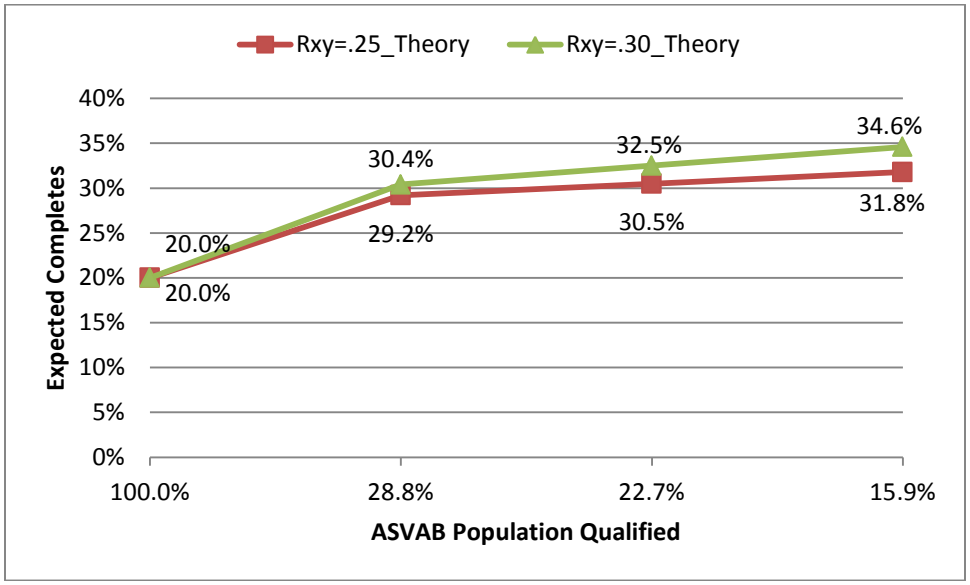

Figure 5
Theory Based BUD/S 1st Phase Complete Rates
for .25 and .30 ASVAB Composite Validities

Figure 5 shows that, when comparing ASVAB composites with .25 and .30 validity, the two Taylor-Russell (1939) table derived success rate lines are not parallel across increasingly stringent cutscore levels. The .30 validity line has a stepper slope and therefore shows increasingly higher completion rates relative to the .25 validity line at the same cutscore points. For example, at a cutscore that qualifies the top 15.9% of the population, the .30 validity line shows a 34.6% complete rate compared to 31.8% for the .25 validity line for a 2.8% difference. In comparison, a cutscore that is less stringent qualifying 28.8% of the population shows a lower 1.2% complete rate difference (30.4% - 29.2%). There are two practical implications. First, higher ASVAB scores along with larger validity coefficients return larger success rate gains as selection becomes more stringent. Second, we can set the cutscores on two composites to attain the same desired success rate (e.g., a 22.7% qualification rate for a .30 validity ASVAB composite achieves about the same success rate as a more lenient 28.8% qualification rate for a .25 validity ASVAB composite - 30.5% compared to 30.4%, respectively).

ASVAB Waiver Analysis

This section examines BUD/S 1st Phase completion rates for ASVAB qualified and ASVAB waivered candidates from the N = 335 data set. Table 10 provides these rates and also the ASVAB waiver rates for the FY10 SEAL accession population.

Table 10
BUD/S 1st Phase Completion Rates by ASVAB Status and Waiver Rate for FY10 SEAL Accessions

Composites/Cutscores Qualified or Not	BUD/S 1st Phase Completion Rate (N = 335 Sample)	FY10 SEAL Accessions (N = 1,141)
Qualified on Alternative Standards		
GS+MC+EI ≥ 165 "or" VE+MK+MC+CS ≥ 220	31.8% (95/299)	1,008 (88.3% Qualified)
Waivered on Alternative Standards		
GS+MC+EI < 165 "and" VE+MK+MC+CS < 220	25.0% (9/36)	133 (11.7% Waived)

Note. The 6.8% difference in the completion rate between the ASVAB qualified and waivered groups was not statistically significant (Pearson Chi-Square = .668, p=.407, df = 1) even though in the expected direction.

Table 10 shows that 31.8% of the 335 candidates completed BUD/S 1st Phase training for the ASVAB qualified group, compared to 25.0% for the ASVAB waivered group. The 6.8% difference (31.8% - 25.0%) appears meaningful but was not statistically significant (see the Table note). Table 10 also shows that 11.7% (133) of the 1,141 FY10 SEAL accessions were ASVAB waivered, which may indicate (a) some stress in filling the SEAL goal for the study time frame that may be related to unavailability of Coding Speed scores or (b) that other highly relevant SEAL strengths were demonstrated justifying ASVAB exception to policy waivers.

Given that the VE+AR composite was found to have the same validity as VE+MK+MC+CS, and that GS+MC+EI appears to be inordinately relied upon, a reasonable strategy is to add VE+AR as a third alterative composite for SEAL selection. There are benefits, but also risks, to a three alternative selection standards model. The benefits are (a) opening the aperture for SEAL selection with three different cognitive profiles (although correlated) that may all be relevant for SEAL field performance, (b) reducing the necessity for an ASVAB waiver, and (c) accessing more SEAL candidates in the outlying more rural METS areas where the market may not be fully tapped (by having to qualify solely on the GS+MC+EI composite because CS is not administered at the METS).[21] The risk in the three alternative ASVAB standards model is that candidates qualifying at the margin on only one standard may do so merely due to test measurement error. Psychological tests are not 100% reliable and adding ways to qualify provides opportunities to do so merely by chance. Strategies to reduce qualifying by chance could be to (a) instate an ASVAB no waiver policy requiring candidates to study harder/retest (demonstrating perseverance) if they do not meet one of the standards or (b) require a candidate to meet two out of the three standards (suggested by the SO ECM during a positive recruiting environment).

[21] METS and MEPS applicants may differ somewhat in experiences growing up due to geographical, but also, cultural differences. These differences may include variations in approaching problem solving, differences in problems, and the context for the problem, which should increase SEAL diversity and enhance team capabilities.

ASVAB/CS Standards Models

Table 11 lists three ASVAB/CS models (one, the operational) along with the AFQT benchmark model to assess qualification rates for the total FY10 accession population and disqualification rates for the FY10 SEAL accessions.

Table 11
ASVAB/CS Standards Models and AFQT for Benchmarking

Alternative ASVAB/CS Models	FY10 Accessions Qualified (of N = 37,084) and Average AFQT	FY10 SEAL Accessions Disqualified (of N = 1,141)
Operational Standards (1 required)		
GS+MC+EI ≥ 165 "or" VE+MK+MC+CS ≥ 220	N = 20,020 (54.0%) AFQT average = 77.0	N = 106 (9.3%)
Model 1-Alternative Standards (1 required)		
GS+MC+EI ≥ 170 "or" VE+MK+MC+CS ≥ 220 "or" VE+AR ≥110 + MC ≥ 50	N = 20,461 (55.2%) AFQT average = 78.1	N = 116 (10.2%)
Model 2-Alternative Standards (2 out of 3 required)		
GS+MC+EI ≥ 170 "and" VE+MK+MC+CS ≥ 220 "or" GS+MC+EI ≥ 170 "and" VE+AR ≥ 110 + MC ≥ 50 "or" VE+AR ≥110+ MC ≥ 50 "and" VE+MK+MC+CS ≥ 220	N = 14,018 (37.8%) AFQT average = 83.2	N = 340 (29.8%)
AFQT (for benchmarking to past and future recruiting environments)		
AFQT ≥ 70 (approximates the composite cutscores – lowest level)	N = 17,113 (46.1%) AFQT average = 83.0	The AFQT is not used for military job classification but could be used to benchmark the current favorable Navy and SEAL recruiting environment.
AFQT ≥ 77 (approximates the composite cutscores - mid-level	N = 12,105 (32.6%) AFQT average = 87.3	
AFQT ≥ 84 (approximates the composite cutscores – highest level)	N = 7,995 (21.6%) AFQT average = 91.0	

Notes. (1) All 37,084 FY10 accessions had GS+MC+EI and VE+AR scores; 25,006 had VE+MK+MC+CS scores.
(2) The average AFQT for FY10 Navy accessions was 67.3; and for the FY10 1,141 SEAL accessions, 77.5.

Table 11, as with Table 7, should be evaluated from the standpoint that fewer than the 37,084 FY10 accessions have CS scores (25,006). Therefore, a comparison of qualification rates should be interpreted as availability of accessions, not level of aptitude/ability. The average AFQT of FY10 accessions is more informative for benchmarking average academic math and verbal skill of those qualified for SEAL selection in addition to the Navy's yearly accession population from which some SEAL candidates would be recruited.

Addressing first FY10 accession population qualification rate, Table 11 lists the operational alternative standards first (GS+MC+EI ≥ 165 "or" VE+MK+MC+CS ≥ 220) and shows that 54.0% (20,020) of the FY10 accessions qualified for SEAL classification, which is much higher than the 28.8% PAY97 ASVAB population qualified applying GS+MC+EI ≥ 165 as a single standard (see Table 7) (again indicating the current positive recruiting environment). The operational standards resulted in an average AFQT of 77.0 compared to 67.3 for the total FY10 Navy accession population (Note in Table 11).

Model 1 adds VE+AR ≥ 110 plus MC ≥ 50 as a third alternative standard but also raises the 165 cutscore for GS+MC+EI to 170 (to compensate for the .05 lower validity). Table 11 shows that 55.2% (20,461) of FY10 accessions qualified, which is just slightly higher than the 54.0% for the operational standards. The three alternative standards resulted in an average AFQT of 78.1 which is just slightly higher than 77.0 for the operational standards. Model 1 adds flexibility to the ability/aptitude profile assessment of a candidate SEAL and eliminates the reliance on GS+MC+EI as the primary classification composite. Model 1 would have a suggested a no exception to policy ASVAB waiver policy due to the study's apparent lower completion rate for those ASVAB waivered, and also because of the non-zero ASVAB validity. Also, candidate motivated to become a SEAL who did not meet one of the three standards would have an opportunity to study harder and achieve more education (potentially) and return for an ASVAB retest, a behavior that demonstrates not only motivation but perseverance. On the other hand, an exception to policy ASVAB waiver could be issued upon the candidate demonstrating other SEAL relevant strengths.

Model 2 is a variation of Model 1 but requires meeting two out of the three alternative standards. Table 11 shows that 37.8% (14,018) of FY10 accessions qualified to become a SEAL on ASVAB, which is much lower than the 55.2% qualified for the Model 1 standards. The two out of three requirement resulted in an average AFQT of 83.2. Model 2 would have a suggested flexible exception to policy waiver guidance where candidates demonstrated other than ASVAB strengths.

The last block in Table 11 provides three AFQT scores (70, 77, and 84) that approximate the three levels of ASVAB cutscores evaluated in this study (composite cutscores that produced 28.8%, 22.7%, and 15.9%, respectively, of the PAY97 ASVAB normative population SEAL qualified). (Again, these AFQT cutscores are meant only to establish benchmarks for the current positive recruiting environment that can be compared with future Navy accession and SEAL populations.)The AFQT can be used as one tool for deciding from year to year, all other things being equal, whether the more or less stringent ASVAB standards should be applied (providing an actual metric for deciding ASVAB stringency flexibility).

Addressing the FY10 SEAL accession disqualification rates (last column), Table 11 shows that Model 2, the two out of three standard, disqualified the largest percentage of already classified SEALs (29.8%), about three times that of the operational standard (9.3%). Clearly, this model would only be used as an "option" in stellar recruiting environments that would be defined by not just the ASVAB, but by multiple SEAL potential dimensions.

Summary and Conclusions

This study was narrowly focused to address an immediate SEAL community need regarding the ASVAB standards and so does not address the total SEAL selection/classification system. A summary of the findings and conclusions follow:

1) The validity of the ASVAB/CS is low (estimated at .25 - .30) for predicting BUD/S 1st Phase completion rates; nevertheless, the ASVAB has practical value in SEAL selection. It is unknown at this time which of the underlying constructs measured by the ASVAB map to the ability to complete BUD/S 1st Phase training, or to perform well in the field.
2) Because of the low ASVAB/CS validity, there are limits to improving the about 30% observed BUD/S 1st Phase completion rate by simply raising cutscores. The upper limit appears to be about 37% evaluated in two independent samples. One sample was comprised of extremely motivated, physically fit candidates (without a higher ASVAB cutscore applied) and the other sample, not at the same motivation/physical levels, but with a very high ASVAB composite cutscore applied.
3) The GS+MC+EI composite is not recommended for replacement at this time even though it has an estimated .05 lower validity than the other composites evaluated in this study. The composite is retained because it (a) measures a technical factor that may relate to job performance (not yet measured), (b) is not devoid of measuring problem solving skills, and (c) may complement other team skills that result from diversity of experiences. A higher cutscore for GS+MC+EI compensates for the slightly lower validity.
4) The VE+MK+MC+CS composite was found, consistent with past studies, to have value for SEAL selection most likely due to multiple relevant constructs (academic, technical, and speed/accuracy with a motivational/perseverance component). However, the prior SEAL ASVAB composite, VE+AR, had equally high validity and was evaluated in conjunction with a separate MC cutscore, as was operational for SEALs prior to 2004.
5) The VE+AR+MK+AO composite had the same validity as VE+AR and VE+MK+MC+CS and will be a candidate for future SEAL selection research. As with CS, the AO test (measuring spatial ability) is less academically linked and thus can reduce subgroup differences that occur from use of academically based measures. However, as with CS, AO is not administered to all Navy applicants, an issue that may be addressed in the future.
6) Two ASVAB/CS alternative standards models were considered appropriate in providing the SEAL community with flexibility in adapting to varying SEAL and Navy recruiting environments. The model that appears appropriate for SEAL selection in a moderate to difficult recruiting environment appears to provide an adequate floor on cognitive ability and has three options for qualification. The three alternative standards are (GS+MC+EI $\geq$ 170) "or" (VE+MK+MC+CS $\geq$ 220) "or" (VE+AR $\geq$ 110 plus MC $\geq$ 50). This model has a suggested no waiver policy because of the apparent lower BUD/S 1st Phase completion rates by candidates ASVAB waivered (consistent with non-zero ASVAB validity), but could be open for an exception to policy waiver upon demonstration of importance SEAL strengths that do not correlate with the ASVAB. The second model is a variation of the first that requires meeting two out of the three alternative standards. The second model appears appropriate for SEAL selection in extraordinarily positive recruiting environments where all candidates exceed the physical and non-cognitive requirements to the same degree. Both models reduce over-dependence on the CS test.

Recommendations

The following recommendations regarding the ASVAB standards for the SEAL rating are addressed to CNO-13, Naval Special Warfare Command, Navy Special Warfare Recruiting Directorate, Navy Recruiting Command, the Center for SEAL/SWCC, and the SO Enlisted Community Manager and Technical Advisor.

1) Replace the operational ASVAB alternative standards with the following alternative standards: GS+MC+EI ≥ 170 "or" VE+MK+MC+CS ≥220 "or" (VE+AR ≥110 plus MC ≥ 50). These standards are appropriate for moderate to difficult recruiting environments and should not be waivered unless candidates demonstrate exceptional SEAL related strengths.
2) Adopt a two out of three version of the above ASVAB standards, as appropriate, for exceptionally positive recruiting environments. These standards should be subject to ASVAB waivers when candidates demonstrate other than ASVAB SEAL related strengths.
3) The AFQT scores, 70, 77, and 84, should be used to benchmark Navy and SEAL accessions during the current exceptional recruiting environment to inform future recruiting efforts. A periodic AFQT assessment will help the SEAL community and N132G (S&C office) about the suitability of applying the more or less stringent ASVAB standards for SEAL selection/classification.
4) Formalize an evaluation/validation study of the complete SEAL multiple hurdle selection system when job performance measures are available. This study should include the new version of Coding Speed and other ASVAB candidate tests that are currently scheduled for MEPS administration on the computer version of the ASVAB (a working memory test and possibly a non-verbal reasoning test).

The SEAL annual goals in the future will be reduced from the historical 1,000 + to be in line with the Navy's drawdown but also because of improved screening processes that result in higher SEAL training success rates. At the same time the annual requirement for successfully trained SEALs is increasing. The SEAL community is making every effort to identify, at the earliest point possible, those candidates likely to pass all of the SEAL selection and training hurdles so that every candidate assessed has a higher likelihood of becoming a SEAL. One additional screening strategy is to identify ASVAB qualified SEAL candidates earlier than formal ASVAB administration. A possible approach is to administer the Navy's Enlistment Screening Test (EST), which resides on the Recruiters' laptops, to SEAL prospects. The EST, which correlates about .84 with the AFQT, does, within error bands, give the Recruiter a rough estimate of whether he or she should invest the time in taking the prospect to the MEPS for ASVAB testing. The EST could be used by the SEAL community in the same way, which is as an initial tool to decide whether or not to invest recruiting resources when an interested SEAL prospect does not exhibit many SEAL potential attributes.

Another screening strategy is to assess SEAL prospects early on with a dynamic situational judgment test (SJT) developed to depict critical field scenarios that require the candidate to demonstrate the autonomous and team skills necessary to successfully perform the evolving SEAL missions. A future study could be planned to determine the feasibility of these two approaches for enhancing and streamlining the SEAL selection/classification system.

Appendix B
Armed Service Vocational Aptitude Battery (ASVAB) and Navy Advanced Placement Test (NAPT) Standards: Nuclear Field Ratings

Armed Services Vocational Aptitude Battery (ASVAB) and Navy Advanced Placement Test (NAPT) Standards: Nuclear Field (NF) Ratings

Janet D. Held, David L. Alderton, and LCDR Don Britton
Navy Personnel Research, Studies, and Technology
NPRST (BUPERS-1) Millington, TN

Reviewed and Released by
David M. Cashbaugh
August 24, 2010

Armed Services Vocational Aptitude Battery (ASVAB) and Navy Advanced Placement Test (NAPT) Standards: Nuclear Field (NF) Ratings

Janet D. Held, David L. Alderton, and LCDR Don Britton
NPRST (BUPERS-1) Millington, TN

Introduction

The Chief of Naval Operations (N-132G), at the request of the Naval Reactors (NAVSEA-08), tasked the Navy Personnel Research, Studies, and Technology (NPRST/BUPERS-1) to review the Armed Services Vocational Aptitude Battery (ASVAB) and Navy Advanced Placement Test (NAPT) standards for the Nuclear Field (NF) ratings as well as non-ASVAB waiver impact on training performance. The NF ratings are Machinist's Mate (MM), Electrician's Mate (EM), and Electronics Technician (ET). The current NF ASVAB and NAPT standards were set in 1998 and so a review of the standards is warranted.[22] The major objective for the study was to re-estimate the validity of the ASVAB and the NAPT in predicting NF course grades and review the effectiveness of the operational cutscores in producing acceptable training graduation rates. ASVAB and NAPT score waivers are not given so an additional objective was to consider whether there is latitude for issuing test score waivers under specific conditions during difficult recruiting environments.

Four questions submitted by various entities were addressed in this study. First, at the request of the NAVSEA-08 and the Enlisted Community Manager (N133D), should there be a revision in non-test score waiver policy for both academic and non-academic reasons, or guidance for not issuing a specific combination of waiver reasons. Second, should the MM rating have a different aptitude standard than the EM and ET ratings given that deep knowledge of electrical and electronics information is not essential for MMs. Third, at the request of Naval Nuclear Power Training Command (NNPTC), would NF rating classification be more effective if conducted at the schoolhouse where student performance could be observed in a short prep-course developed to measure in depth electrical and mechanical skills, since the MM students were not thought to be doing as well in the follow-on to A-School Power School as the EM and ET students. Fourth, at the request of Navy Recruiting Command (NRC) we considered the question of whether women do as well as men in the training, but after speaking with school officials and conducting the analysis for the small sample of women, we concluded they were and that any further investigation should be addressed in a future study that would also address the performance of various ethnic groups.

The ASVAB is the enlisted selection and classification instrument for all military services. The battery consists of nine tests listed and briefly described in Table 1. Each test has a standard mean score of 50 and a standard deviation of 10; scores typically range from 25 to 75. Assembling Objects (AO) is the newest ASVAB test.[23] Coding Speed (CS), a former ASVAB test, is now a Navy special test and is used in two Navy classification composites (not addressed in this study).[24]

---

[22] Held, J., Johns, C., & McMahan, M. (1998). *Validation of the Armed Services Vocational Aptitude Battery (ASVAB) and the Nuclear Field Qualification Test (NFQT) for the Nuclear Field (NF) Class "A"- and Power Schools* (Ltr Rep NPRDC-12: ser 3900/12/279 of 4 Jun 98). San Diego: Navy Personnel Research and Development Center.

[23] AO is administered in both the computerized version of the ASVAB (CAT-ASVAB) and in the enlisted paper-and-pencil forms. AO is not administered in the Student Testing Program (STP) paper-and-pencil forms.

[24] CS is administered only at the Military Entrance Processing Stations (MEPS) at the end of CAT-ASVAB.

WK and PC are combined and rescaled to form the ASVAB Verbal (VE) composite. VE is included in the Armed Forces Qualification Test (AFQT) used to qualify applicants for military service, and in several Navy rating classification composites. The AFQT, a composite equal to 2VE+MK+AR, is rescaled to represent the cumulative percentile scores for the national population of 18-23 year old non-institutionalized youth (scores range from 1 to 99). [25]

Table 1 provides a brief description of the ASVAB tests and Coding Speed, a Navy special classification test.

Table 1
Description of the ASVAB and Coding Speed Tests

Test Name and Abbreviation	Test Description
General Science (GS)	Knowledge of physical and biological sciences
Arithmetic Reasoning (AR)	Ability to solve arithmetic word problems
Word Knowledge (WK)[a]	Ability to select the correct meaning of words presented in context and correct synonyms
Paragraph Comprehension (PC)[a]	Ability to obtain information from written passages
Mathematics Knowledge (MK)	Knowledge of high school mathematics principles
Electronics Information (EI)	Knowledge of electricity and electronics
Auto and Shop Information (AS)	Knowledge of automobile and shop technologies, tools, and practices
Mechanical Comprehension (MC)	Knowledge of mechanical and physical principles
Assembling Objects (AO)[b]	Ability to determine correct spatial forms from their separate parts and connection points
Coding Speed (CS)[b]	Ability to quickly identify correct word/number pairings from a key with many options

[a]WK and PC are combined to form the Verbal (VE) composite that is a component of the AFQT and several Navy ASVAB classification composites. [b]Not all recruits enter the Navy with AO and CS test scores. CS is only given at the MEPS at the end of the CAT-ASVAB. AO is not given to high school students taking the paper and pencil ASVAB.

Various combinations of the ASVAB tests, called composites, are used by each service to classify recruits into their military occupations. Validation of ASVAB composites is conducted periodically for each Navy rating's initial technical training school to ensure that the ASVAB composite most predictive of training performance is used to classify recruits. A minimum qualifying score (cutscore) is set for each rating's ASVAB composite to manage academic non-graduation and setback rates, while at the same time considering that a certain amount of recruit talent must be made available for all Navy schools.

The NAPT is an approximately 2-hour test that measures chemistry (6% of the test), math (61%), and physics (33%) constructs. The math questions are categorized as arithmetic, geometry, trigonometry, algebra, and probability. NAPT questions were developed to measure knowledge and comprehension (20% of the test) and application (80% of the test). There are purportedly multiple NAPT parallel forms.

[25] Segall, D. O. (2004). *Development and Evaluation of the 1997 ASVAB Score Scale* (Technical Report 2004-02. Seaside, CA: Defense Manpower Data Center. www.official-asvab.com/docs/asvab_techbulletin_2/pdf.

There is some content overlap between the NAPT and the ASVAB GS, MK, and AR tests used for NF classification; however, the NAPT was developed to measure higher math and science ability than the ASVAB and so is a better and more reliable measure in the NF relevant aptitude/ability range. (The ASVAB is developed for the general applicant population and does not focus on measuring aptitude/ability in the very high and very low ranges. The NAPT therefore fills the ASVAB measurement gap as well as content gap.)

The NF ratings have two methods for ASVAB use, one of which requires the NAPT. Both methods apply two ASVAB composites: Electronics composite (AR+MK+EI+GS) or the Nuclear Field composite (VE+AR+MK+MC) as alternatives with cutscores that were set to give equivalent classification standards within each method. Meeting a very high 252 cutscore on either composite eliminates the requirement to take the NAPT and this group is referred to as "Automatically Qualified", or for the purposes of this study, Method A selected. If the score is lower than 252 on both ASVAB composites, the NAPT is required and a 290 cutscore must be met for one of the ASVAB composites + NAPT score combinations; for the purposes of this study, this group is referred to as "NAPT Required", or Method B selected. Table 2 contains a summary of the two enlisted NF selection methods.

Table 2
Nuclear Field Rating Selection Methods

Method A Selection		
Electronics Composite	AR+MK+EI+GS	≥ 252
OR		
Nuclear Field Composite	VE+AR+MK+MC	≥ 252
Method B Selection		
Electronics Composite + NAPT	AR+MK+EI+GS+NAPT[a]	≥ 290
OR		
Nuclear Field Composite + NAPT	VE+AR+MK+MC+NAPT[a]	≥ 290

[a]Requires a minimum of 50 on the NAPT or a 55 if the NAPT score is a retest.

Not shown in Table 2 is an inadvertent "standard" for a small number of students who qualified with the 252 score, Method A selected, but were administered the NAPT anyway, possibly because the recruiter identified them as at risk for failure. For the purposes of this study this third group is referred to as Method C selected and it is involved in only one analysis discussed later.

The 1998 NF study found that the two operational ASVAB composites, AR+MK+EI+GS and VE+AR+MK+MC had higher validity in predicting "A" School and Power School performance than the NAPT, but that the NAPT added incremental validity (all the measures were correlated). The incremental validity that the NAPT provided meant that classification decisions would be more accurate if the ASVAB and the NAPT were used in combination, without setting an explicit cutscore on each. However, as a conservative strategy the 290 total cutscore included a 50 cutscore for the NAPT. An objective of this study was to evaluate the adequacy of the NAPT 50 cutscore, the total ASVAB + NAPT 290 cutscore, and whether there should be an explicit cutscore for each. The current thought is that in recent years the general student population has benefited from exposure to a higher level of mathematics and science courses, and at an earlier age, and that NAPT scores may have risen since the 1998 NF study, but that marginally higher NAPT scores may not reflect the true ability required to successfully complete the difficult NF courses.

Apart from potential changes in the knowledge of applicants, there have been substantial changes in the curriculum and delivery methods at NNPTC, which may have affected the validity of the current ASVAB and NAPT instruments and also the NF selection/classification methods. For example, there have been many improvements in training and training technology. Notably, the introduction of high fidelity computer graphics and increased opportunities for hands on experience has helped contextualize and solidify learning. The many opportunities to learn from different sources and curriculum delivery methods also foster learning. The impressions taken from the school visit for this study are that the NNPTC staff and the deliberate cultivation of a structured, disciplined learning environment have produced an excellent climate for producing first rate NF professionals, which may be quite different from the training circumstances in 1998.

Methods, Analyses, and Results

The study is organized into the following sections: (1) Data collection, (2) Waiver analysis, (3) Regression analysis, (4) ASVAB composite validation, (5) Empirical cutscore analysis, (6) Theoretical cutscore analysis, (7) ASVAB and NAPT cutscore combinations, and (8) "A" School GPA relationship to Power School performance. A summary and conclusions section follows, and finally, the recommendations.

Data Collection

A large multiyear data file was obtained from NNPTC (over 8,000 cases) that spanned from 2004 to 2009. The NNPTC dataset contained grade point average (GPA) for both the "A" School and the Power School, graduation status from each school, waiver reason upon acceptance as an NF candidate, and reason for disenrollment. The GPA cut-off point to graduate from each school appeared to be 2.50. Another file was developed from the NRC accession files for fiscal year (FY) 2006– 2009 (FY06 – FY09). Both files were merged on a code variable so that ASVAB test scores, calculated ASVAB composites scores, NAPT scores, and other variable in the NNPTC file could be in one file and to check the NAPT scores for accuracy. A third file was developed from the Navy Integrated Training Resources and Administration System (NITRAS) database to compare reasons for disenrollment with those listed in the NNPTC file.

The merged file contained no NAPT score below 50 and the highest score observed was 79. Of the 6,031 cases retained, 2,261 (37.5%) had NAPT scores. The NAPT mean score was computed for each of four fiscal years (FY2006 – FY2009) and they appeared stable for the last three years (mean of about 59 and standard deviation of slightly lower than 5.5). The ASVAB composite mean scores were also similar for this 3-year time frame.

Table 3 gives the ASVAB composite means and standard deviations, and the GPA (where present) means and standard deviations for "A" and Power School for the final 4-year NNPTC and NRC matched dataset broken out by the Auto-qualified and NAPT required groups (Methods A and B).

Table 3
Means, Standard Deviations, and Quartiles for the ASVAB Composites and "A" and-Power School GPAs for the Matched Dataset (ASVAB Auto-qualified and NAPT Required Students)

Measure	Mean	Standard Deviation	Quartile Values (.25, .50, .75)
Method A: ASVAB Auto-qualified Students (N=3,181)			
VE+AR+MK+MC	257.4	10.44	252, 257, 263
AR+MK+EI+GS	259.2	10.69	253, 258, 265
"A" School GPA	3.27	.323	3.07, 3.31, 3.52
Power School GPA	3.22	.372	2.98, 3.26, 3.51
Method B: NAPT Required Students (N=2,739)			
VE+AR+MK+MC	239.3	7.64	234, 240, 245
AR+MK+EI+GS	238.7	8.08	234, 240, 245
"A" School GPA	3.03	.360	2.82, 3.06, 3.28
Power School GPA	3.04	.376	2.79, 3.06, 3.31

Note. Some cases were excluded in the analysis due to missing GPAs.

Table 3 shows that the two ASVAB composites' mean scores for the Auto-qualified students are about 20-score points higher than for the NAPT required students. Table 3 also shows that the "A" School and Power School GPA means are also higher for the Auto-qualified (for example, a Power School GPA of 3.22 for the Auto-qualified students compared to 3.04 for the NAPT required students). Obviously, if a 2.50 GPA is required to graduate from "A" School; fewer NAPT required students would be expected to graduate from the training pipeline than Auto-qualified students, all other things being equal.

Table 4 considers that some of the students with NAPT scores were ASVAB Auto-qualified, but were required to take the NAPT as an extra screen (Method C selected). Table 4 gives, for FY06-07, a breakout of the "A" and Power School drop rates and the total pipeline graduation rates by rating and by the A, B, and C selection methods. Table 5 applies to FY08-09, which had fewer students than the FY06-07 dataset because many students who accessed in FY09 (from the NRC file that was used in the NNPTC data match) had not received a final pipeline disposition.

Table 4 shows for FY06-07 that students who qualified by having to take the NAPT (Method B selected) had lower pipeline graduation rates than those Auto-qualified by the ASVAB (Method A selected). For the NAPT required students the Power School graduation rates for the MM, EM, and ET ratings were 75.1%, 74.8%, and 77.4%, respectively; for the Auto-qualified students they were 86.6%, 86.1%, and 88.1%, respectively. The about 10% Power School graduation difference between the two selection methods is not surprising given there is no ASVAB cutscore applied in the ASVAB + NAPT standard, but for even for the NAPT required students the total pipeline graduation rates are high when compared to those reported in the 1998 NF study, which supports the decisions made at that time to change the NF aptitude/ability standards.

Table 4
Drop Rates for FY06-07 NRC Matched Accession Students ASVAB Auto-qualified, Qualified at 290 on ASVAB+NAPT, and ASVAB Auto-qualified but Took the NAPT (N=3,770)

Drop Variables	MM	EM	ET
Method A: ASVAB Auto-qualified on 252 (N=2,025)			
"A"Drop	3.5%	6.2%	6.8%
Power Drop	9.9%	7.7%	5.1%
Power Graduate	86.6%	86.1%	88.1%
Total Number of Students	1,063	534	428
Method B: ASVAB with NAPT qualified on 290 (N=1,666)			
"A"Drop	9.0%	14.0%	10.8%
Power Drop	15.9%	11.2%	11.8%
Power Graduate	75.1%	74.8%	77.4%
Total Number of Students	820	457	389
"Method C"[a]: ASVAB Auto-qualified on 252 but who took the NAPT (N=79)			
"A"Drop	4.4%	7.7%	14.3%
Power Drop	20.0%	0.0%	4.8%
Power Graduate	75.6%	92.2%	81.0%
Total Number of Students	45	13	21

[a]Method C is not really a selection method for but merely a category of interest for analysis.

Table 5
Drop Rates for FY08-09 NRC Matched Accession Students ASVAB Auto-qualified, Qualified at 290 on ASVAB+NAPT, and ASVAB Auto-qualified but Took the NAPT (N =2,261)

Drop Variables	MM	EM	ET
Method A: ASVAB Auto-qualified on 252 (N=1,156)			
"A"Drop	3.4%	6.6%	8.0%
Power Drop	12.9%	7.8%	7.6%
Power Graduate	83.7%	85.7%	84.4%
Total Number of Students	673	258	225
Method B: ASVAB with NAPT qualified on 290 (N=1,073)			
"A"Drop	10.9%	15.5%	12.5%
Power Drop	20.3%	10.6%	12.0%
Power Graduate	68.8%	73.9%	75.5%
Total Number of Students	644	245	184
"**Method C"[a]: ASVAB Auto-qualified on 252 but who took the NAPT (N=32)**			
"A"Drop	10.0%	20.0%	14.3%
Power Drop	0.0%	0.0%	0.0%
Power Graduate	90.0%	80.0%	85.7%
Total Number of Students	20	5	7

[a]Method C is not really a Selection Method for recruiters but a category for analysis.

Table 4 also shows that graduation rates across the ratings within each of the selection methods appeared uniform. However, the NAPT required MM students had a much higher Power School non-graduation rate than the EM and ET students (15.9% for MM compared to 11.2% and 11.8% for EM and ET, respectively), suggesting that the MM "A" School does not contain some critically difficult content that the EM and ET students receive, which penalizes the MM students at the Power School.

Table 5 for FY08-09 also shows that students who qualified by having to take the NAPT (Method B selected) had lower pipeline graduation rates than those Auto-qualified by the ASVAB (Method A selected). For the NAPT required students, the Power School graduation rates for MM, EM, and ET were 68.8%, 73.9%, and 75.5%, respectively; for the Auto-qualified students they were 83.7%, 85.7%, and 84.4%, respectively.[26] Graduation rates across the ratings within the Auto-qualified group appeared uniform, as in Table 4. However, the 68.8% Power School graduation rate for the MM NAPT required students was noticeably lower than for their EM and ET counterparts, suggesting, consistent with Table 4, that the MM "A" School does not contain some critical difficult content that the EM and ET students receive, but also that there is more of an impact for the MM students in the more recent years.

There were not many Method C selected students in Tables 4 and 5 (Auto-qualified but NAPT required) and that group, with the exception of MMs in the earlier Table 4 time frame, appeared to perform comparably to the Method A selected group. The Method C sample size was too small (in either table) to be of consequence in the study and because small samples produce statistically unstable results, Method C students were dropped from subsequent analyses.

The overall conclusions from inspection of the contents of Tables 4 and 5 were that (a) the ASVAB Method A selected Auto-qualified students were doing better in training than the Method B NAPT required students and (b) MMs are lacking curriculum early on that may have helped in their Power School performance.

The next section examines waiver reasons, their incidence, and whether they relate differentially to graduation rates between the two NF selection methods.

Waiver Analysis

The NNPTC dataset contained five data fields that could be used to log the code that pertained to the category of waiver a student might have (No ASVAB score-point waivers are given so no waiver category pertained to the ASVAB), and also five data fields that could be used to describe, for the particular waiver code, the particular reasons. There were ten waiver category codes: ACA = Academic, CIV = Civil, DEP = Dependents, DIS = Discipline, DRU = Drugs, MED = Medical, OTH = Other, PRI = Prior Service, TRA = Non-traditional education, WEI = Physical Standards (Height or Weight). The level of analysis for this study was contained to the ten waiver categories and not the particular reasons (explained in sometimes open ended statements, which could be entered in the future in a standardized coded format). Mutually exclusive waiver groupings were formed from the waiver categories and applied to the more recent FY08-09 time frame of interest. Table 6 shows the percentage of students in each group for each of the two selection methods.

[26] There are two occurrences, one in Table 4 and one Table 5, where addition of the "A" Drop, Power Drop, and Pipeline Graduate rates do not total 100%. These occurrences are due to rounding error and they occur in several other tables.

Table 6
Description and Percent of Mutually Exclusive Waiver Groups Formed from Waiver Categories Broken out for FY08-09 Auto-qualified and NAPT Required Students

Waiver Group Category	Group Waiver Description	Auto-qualified (N=1,156)	NAPT Required (N=1,073)
Group_1	No waivers	48.8%	48.9%
Group_2	Waiver1 Academic, no others	22.8%	22.4%
Group_3	Waiver1 Academic, Waiver2 Civil, no others	2.8%	3.4%
Group_4	Waiver1 Academic, Waiver2 Drug, no others	.9%	1.5%
Group_5	Waiver1 Academic, Waiver2 Other, no others	2.0%	2.6%
Group_6	Waiver1 Academic, Waiver2 Civil, Waiver3 any reason	1.5%	.9%
Group_7	Waiver1 Civil/Drug/Medical, Waiver2 any entry except Academic	17.6%	17.3%
No Group	Did not meet any group definition	3.8%	3.0%
Total Percent		100.0	100.0

Note. None of the waiver categories pertained to the ASVAB (no ASVAB waivers are issued).

Table 6 shows essentially the same percentage of waiver categories across the two selection methods, consistent with the random rating assignment approach taken at Great Lakes.[27] Table 6 also shows that about 49% of Nuclear Field candidates reporting to the schoolhouse had no waivers and the balance had at least one waiver.

Table 6 shows that the waiver category with the highest incidence was one academic waiver and no other (22.8% and 22.4% for the Auto-qualified and NAPT required groups, respectively). The next highest incidence was one academic waiver plus a civil waiver (17.6% and 17.3% for the Auto-qualified and NAPT required groups, respectively). The results from Table 6 were used to cross-tabulate waiver category and reason for disenrollment, but the results were not supportive of a strong relationship that could be used for predictive, diagnostic, or constructive intervention purposes. (For example, many students with only an academic waiver and were dropped in training were not dropped for an academic reason.)

The next analysis was a contrast of the graduation rate differences in each waiver category for each selection system. Because there were very few cases in some of the waiver categories they were further consolidated, waiver group 1 (no waivers) and group 2 (one academic waiver only) were kept intact. Waiver groups 3, 4, 5, and 6 that involved an academic waiver and some other category of waiver were collapsed into waiver group 3. Waiver group 7 (any waiver except academic) became waiver group 4. The "any other" waiver group was dropped due to insufficient sample size.

Table 7 gives the breakout of the graduation rates for the reformulated waiver groups (as NewGroup_1 though NewGroup_4) for each of the two selection methods.

[27]Recruits are apportioned into four aptitude tiers and then randomly assigned into a rating with personal preference considered after some rating duty information is provided. However, Navy manpower requirements in some cases may overturn preferences.

Table 7
Pipeline Graduation Rates by Consolidated Waiver Groups for FY08-09 Auto-qualified and NAPT Required Students

Consolidated Waiver Group Category	Group Waiver Description	Pipeline Graduation for Auto-Qualified (N=1,089)	Pipeline Graduation for NAPT Required (N=1,013)
NewGroup_1	No waivers	86.7% (N=564)	72.6% (N=525)
NewGroup_2	Waiver1 Academic, no others	78.7% (N=263)	65.4% (N=240)
NewGroup_3	Waiver1 Academic, others	67.8% (N=59)	69.4% (N=62)
NewGroup_4	Any waiver type except Academic	86.2% (N=203)	73.7% (N=186)

Note. 67 Auto-qualified students and 60 NAPT required students are not included because they did not fit into the newly defined consolidated waiver category groups.

Table 7 shows, mostly consistent with performance in general for the ASVAB Auto-qualified students, higher NF pipeline graduation rates across consolidated waiver groups compared to the NAPT required students, with the exception of the small sample NewGroup_3 (academic waiver and a least one other). Also, for both groups, an academic waiver only (based on education/ classes) by itself resulted in a substantial drop in the graduation rate compared to having no waivers. The graduation rate for NewGroup_3 that included at least one other type of waiver in addition to academic was in the high 60% range for both selection methods, but considered unstable due to the small sample sizes. Nevertheless, it appears that this category of waiver combination might be considered high risk, at least for the Auto-qualified students who might not receive as much scrutiny as the NAPT required students (due to a lower ASVAB requirement). The next section applies regression analysis and various prediction models to explore how the ASVAB and NAPT factor into the waiver impact.

Regression Analysis

Hierarchical regression analysis was used to assess the contribution of the NAPT and the NF waiver information in predicting "A" School GPA for each NF rating above the ASVAB composites. The analysis is termed "hierarchical" because it allows an assessment of the contribution of each variable (or set of similar variables) to the prediction model at every step. Variables input in the first steps are generally those that are easy to collect without expending extra resources (e.g., administration and maintenance of the NAPT is cost above the cost of the ASVAB). The analysis was not intended to evaluate the relative merits of the ASVAB composites, which is done in the validity analysis reported in the following section. The VE+AR+MK+MC composite was entered first into the equation only because of the interest in determining if AR+MK+EI+GS, entered second, added incremental validity. The two composites, while correlated, do provide utility in identifying slightly different aptitude/ability/knowledge profiles that are both relevant for NF training success.

Table 8 gives the results of the hierarchical regression analysis for the NAPT required students (FY2006 through FY2009) broken out for each rating. Table 9 gives the results for the Auto-qualified students. All students with an "A" School GPA were included in the analyses.

Table 8
Model Summary for Hierarchical Regression Analysis Predicting "A" School GPA for FY06-09 NAPT Required Students

Model	Multiple Correlation (R)	Adjusted R Square	Statistical Significance for R Square Change	Standard Error of the Estimate
MM "A" School (N=1,462)				
1. NF Composite	.200	.039	.000	.338
2. Model 1 + El Composite	.224	.049	.000	.337
3. Model 2 + NAPT	.257	.064	.000	.334
4. Model 3 + Academic only	.264	.067	.019	.333
5. Model 4 + Academic plus	.285	.078	.000	.331
6. Model 5 + Any except Academic	.287	.079	.225	.331
EM "A" School (N=701)				
1. NF Composite	.194	.036	.000	.368
2. Model 1 + El Composite	.199	.037	.265	.368
3. Model 2 + NAPT	.307	.090	.000	.358
4. Model 3 + Academic only	.317	.095	.026	.357
5. Model 4 + Academic plus	.318	.095	.459	.357
6. Model 5 + Any except Academic	.319	.094	.563	.357
ET "A" School (N=573)				
1. NF Composite	.137	.017	.019	.375
2. Model 1 + El Composite	.164	.023	.008	.374
3. Model 2 + NAPT	.261	.063	.041	.366
4. Model 3 + Academic only	.273	.068	.007	.365
5. Model 4 + Academic plus	.273	.067	.000	.366
6. Model 5 + Any except Academic	.275	.066	.001	.366

Note. Drops with GPAs were included in the analysis. At the most only 2 cases per rating did not have an "A" School GPA.

Table 8 shows for the NAPT required students that the VE+AR+MK+MC composite entered first into the equation was statistically significant in predicting "A" School GPA for all three NF ratings (at lower than the .05 probability level typically used for statistical analysis) with the standard error of estimating GPA not highly dissimilar (.338 for MM, .368 for EM, and .375 for ET). The AR+MK+EI+GS composite, entered second into the equation contributed incrementally to the multiple correlation, R, for the MM and ET ratings (but not the EM rating), but the practical significance of the contribution was marginal and the reduction in the standard error of prediction was trivial .001 reduction for both MM and ET from their original .338 and .375, respectively, and no reduction in .368 for EM. Not until the NAPT was entered (Model 3) did appreciable gain in the multiple R occur for each rating, but even then the standard error of estimate was only marginally reduced (in the 2nd decimal place, at best).

Table 8 also shows that having an academic waiver and no other (Model 4) contributed significantly to the prediction of "A" School GPA above the ASVAB and NAPT for all three ratings. For MM and ET, having an academic waiver and at least one other (Model 5) contributed significantly to the prediction above having just an academic waiver, but not for EM. No solid explanation can be given for the EM results (non-significant validity increment for the AR+MK+EI+GS composite or the combination of academic and other waivers) – the results may be due to the quality of the EM curriculum and concentration on electronics/electricity (thereby negating the need to have a large knowledge base in this area) plus teaching/classroom engagement that diminishes risks relative to MM and ET courses (speculation at best).

Table 9 shows the hierarchical regression analysis for the Auto-qualified students with stronger multiple correlations to start off with (The NF ASVAB composite) and statistical significance through the first four models for all three ratings. NAPT values are missing (gapped) because these Auto-qualified students were not required to take the test.

Table 9
Model Summary for Hierarchical Regression Analysis Predicting "A" School GPA for FY06-09 ASVAB Auto-qualified Students

Model	Multiple Correlation (R)	Adjusted R Square	Statistical Significance for R Square Change	Standard Error of the Estimate
MM "A" School (N=1,734)				
1. NF Composite	.336	.113	.000	.299
2. Model 1 + El Composite	.373	.138	.000	.294
3. Gap - no NAPT scores				
4. Model 3 + Academic only	.404	.162	.000	.291
5. Model 4 + Academic plus	.405	.162	.237	.291
6. Model 5 + Any except Academic	.405	.162	.882	.291
EM "A" School (N=792)				
1. NF Composite	.335	.111	.000	.319
2. Model 1 + El Composite	.364	.130	.000	.316
3. Gap - no NAPT scores				
4. Model 3 + Academic only	.379	.140	.001	.314
5. Model 4 + Academic plus	.390	.148	.004	.313
6. Model 5 + Any except Academic	.392	.148	.294	.313
ET "A" School (N=651)				
1. NF Composite	.300	.088	.000	.300
2. Model 1 + El Composite	.342	.114	.000	.296
3. Gap – no NAPT scores				
4. Model 3 + Academic only	.371	.133	.000	.293
5. Model 4 + Academic plus	.378	.138	.036	.292
6. Model 5 + Any except Academic	.380	.138	.354	.292

Note. Drops with GPAs were included in the analysis. At the most only 2 cases per rating did not have an "A" School GPA.

Table 9, like Table 8, shows that having an academic waiver and no other (Model 4) contributed significantly to the prediction of "A" School GPA over and above the two ASVAB composites for all three ratings. For EM and ET, having an academic waiver and at least one other (Model 5) contributed significantly to the prediction over and above having only an academic waiver. Again, although statistically significant, the reduction of the standard error of estimate was marginal at best adding each successive model.

A contrast of interest for Tables 8 and 9 is that the standard errors of estimate in Table 9 for the Auto-qualified students are substantially lower than in Table 8 for the NAPT required students. The ETs show the largest difference. For Model 1 in Table 8, entering the VE+AR+MK+MC (NF) composite into the ET equation in the first step results in a .375 standard error of estimating "A" School GPA compared to .300 for ET in Table 9. The larger standard error of estimate for the NAPT required students may simple mean that a score in the lower ASVAB score range for these students are not as predictive of extremely difficult coursework as are higher ASVAB scores. However, the smaller R values for the NAPT required students are most likely due to the narrow range of ASVAB scores (below 252 but not too much lower than 230) which attenuates the correlation coefficient (labeled a Multiple R in the regression analysis). The ASVAB score range is not as narrow (restricted in range) for the Auto-qualified students with its floor of 252 but without a cap. Restriction in range in the validity coefficient is addressed in the next section.

ASVAB Composite Validation

The two operational ASVAB composites used for NF classification were considered the most appropriate for validity analysis from a rational perspective given their underlying constructs linked well to the curricula, although AR+MK+EI+GS mapped more closely to the EM and ET "A" School curricula while VE+AR+MK+MC mapped more closely to the MM "A" School curriculum. For comparison, several other official Navy composites that were in place for the 1998 study were included. The analysis was conducted using "A" School GPA as the performance criterion but not Power School GPA because it was determined from regression analysis (not reported here) that the ratings' "A" School GPAs were more predictive of Power School GPA than either of the ASVAB composites. Therefore, we oriented our objectives towards predicting "A" School GPA and establishing ASVAB standards that would result in improved "A" School graduation rates, which would then translate to improved Power School graduation rates.

The objective of a validity analysis is to determine (with some statistical error) which ASVAB composites are most predictive of school performance. Validity in this report refers to the correlation of scores on a particular ASVAB composite with scores on the school performance measure ("A" School GPA). Validity coefficients range from -1 to +1. A zero validity coefficient means there is no predictive power at all and that course outcomes are completely unrelated to ASVAB scores whereas a +1 or -1 validity coefficient means there is perfect (linear) prediction (positive and negative, respectively). The average validity coefficient across Navy ratings is around .55. The lowest validity is about +.25 for the physically-focused Basic Underwater Demolition/SEAL course (BUD/S) and the highest is about +.80 to +.85 for the academically-oriented Nuclear Field courses and other highly technical Navy courses.

Figure 1 is a display of three graphs that depict different validity magnitudes for the bivariate relationship between the ASVAB (as the predictor conceptually scaled on the *x*-axis) and final course grade (as the criterion conceptually scaled on the *y*-axis).

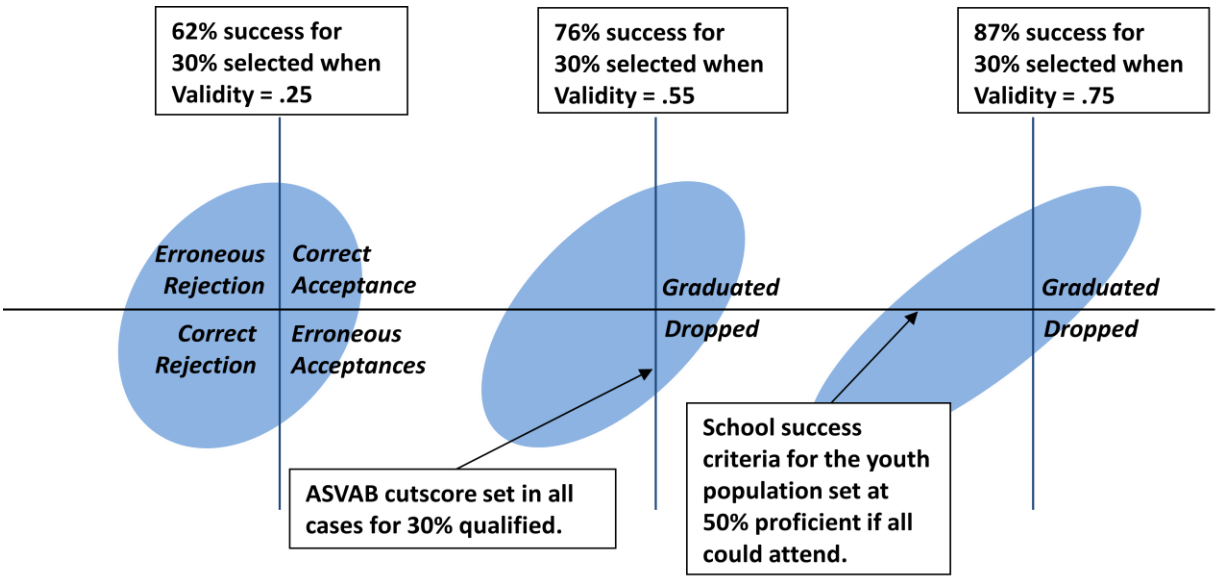

Figure 1
School Success Rate as a Function of Validity and Cutscore

The graphs in Figure 1 show that the ASVAB validity applies to the total applicant population from which future candidates would be selected - not to candidates actually selected (cases that would be to the area to the right of the vertical line depicting the ASVAB cutscore). The objective in the ASVAB validation analysis is to determine the graph's form and associated validity for each of the study's candidate ASVAB composites. All other things equal, the larger the validity coefficient, the greater the success rate (assuming there is a performance deficiency to begin with). An estimate of the unrestricted validity coefficient is also a gauge of the how effective a cutscore adjustment will be in reducing non-graduation rate and is obtained using a multivariate statistical correction procedure[28] (explanation provided for an ASVAB context).[29]

Figure 1 as an illustration does not reflect any of the NF study parameters. The NF cutscore is much more stringent than the displayed 30% qualified, and the success rate had no ASVAB standard (or other cognitive test) been applied (50% shown). Also, the ASVAB estimated population validities for NF discussed in the next section were shown to be higher than the far right graph's .75.

Finally, Figure 1 shows classification decision errors (far left graph) associated with validity magnitude (that generally diminish as validity magnitude increases).

[28] Lawley, D. (1943). A Note on Karl Pearson's Selection Formulae. *Royal Society of Edinburgh, Proceedings, Section A, 62*, 28-30.
[29] Held, J. D. & Foley, P. P. (1994). Explanations for Accuracy of the General Multivariate Formulas in Correcting for Range Restriction. *Applied Psychological Measurement, 18(4)*, 355-367.

The ASVAB composite validities were developed only for the Auto-qualified students because of the score variance curtailment for the NAPT required students due to a score cap/ceiling on ASVAB (252). Table 10 gives, for each rating, the validities corrected for range restriction using "A" School GPA as the criterion.

Table 10
ASVAB Composite Corrected Validities for Auto-qualified Students using "A" School GPA as the Criterion
(FY2006-2009 Accession data)

ASVAB/NAPT	MM-A (N=1,484)	EM-A (N=681)	ET-A (N=567)
VE+AR+MK+MC*	.84	.82	.84
AR+MK+EI+GS*	.84	.81	.83
VE+AR	.82	.81	.82
VE+MK+GS	.83	.81	.82
AR+2MK+GS	.82	.80	.82

*Operational NF composite.

Table 10 shows higher ASVAB validities than were found in the 1998 Nuclear Field study (that were generally in the mid-to high .70 range). Compared to the other ASVAB composites the two operational NF composites, VE+AR+MK+MC and AR+MK+EI+GS, still have the highest (and equal) validities for the MM rating and near equal for the EM and ET rating (although a .01 difference is trivial). To answer a study question about whether VE+AR+MK+MC should be used exclusively for the MM rating because of the MC test, the validity of the two operational composites for MM were the same (.84) lending no evidence for such a decision.

Empirical Cutscore Analysis

Cutscore analysis is conducted for the Navy considering the following factors: (1) academic non-graduation and setback rates, (2) waiver rates, (3) yearly school input requirement, (4) intellectual complexity of the rating, and (5) training time and cost.

The yearly school input requirement and complexity of the NF jobs are both considered high. The ASVAB and NAPT waiver rates are nonexistent; however, the academic waiver rate, solely or in combination with other categories of waivers, is considered substantial (but perhaps unavoidable). The NF training pipeline is lengthy and expensive so it is not cost effective for the Navy to send candidates to NNPTC if they have a high likelihood of failure (recognizing that the ASVAB and other information used to classify NF candidates are not perfect predictors and so classification decisions cannot be 100% accurate).

We acknowledged earlier in the study the difficulty in resolving the student drop codes with disenrollment codes in our dataset formulation and thus the decision to include all cases in our analyses, which includes the empirical cutscore analyses that follow. The reported results are considered conservative as we would expect a cutscore analysis that included only graduates and drops for purely academic reasons to show greater improvements in the graduation rate as a result of raising the ASVAB cutscore. We limited the cutscore analysis to the MM rating, given the challenges relative to the EM and ET ratings in graduating from the Power School and the shorter MM "A" School.

The first set of empirical cutscore analyses for the MM rating applied to the NAPT for the NAPT required students. The ASVAB was not considered in this cutscore analysis because of the ASVAB cap/ceiling effect (252 score). The analysis was conducted separately for students who received an academic waiver solely or in combination with others (Table 11), and for students who did not receive any waivers (Table 12) to get a further understanding of the impact of issuing academic waivers.

Table 11
NAPT Cutscore Analysis for MM NAPT Required Students who Graduated or Dropped for Any Reason and had at Least One Academic Waiver (FY06-09, N=380)

NAPT Score	Graduates (#)	Non-Graduates (#)	Graduation Rate (%)	Students at or Above the Score (#)	Students of Total Above the Score (%)
64	50	12	80.6	62	16.3
62	72	21	77.4	93	24.5
60	103	32	76.3	135	35.6
58	**140**	**51**	**73.3**	**191**	**50.3**
56	182	67	73.1	249	65.5
54	220	90	71.0	310	81.6
52	249	109	69.6	358	94.2
50	266	114	70.0	380	100.0

Table 12
NAPT Cutscore Analysis for MM NAPT Required Students who Graduated or Dropped for Any Reason and had No Waivers (FY06-09, N=734)

NAPT Score	Graduates (#)	Non-Graduates (#)	Graduation Rate (%)	Students at or Above the Score (#)	Students of Total Above the Score (%)
64	120	34	77.9	154	21.0
62	190	56	77.2	246	33.5
60	241	80	75.1	321	43.7
58	**321**	**109**	**74.7**	**430**	**58.6**
56	385	144	72.8	529	72.0
54	447	181	71.2	628	85.6
52	485	199	70.9	684	93.2
50	523	211	71.3	734	100.0

Tables 11 and 12 show that MM students with at least an academic waiver had initial pipeline graduation rates that were not much different than MM students without any type of waiver (70.0% and 71.3%, respectively). Further, the graduation rates for both groups were similarly improved by raising the operational NAPT cutscore of 50 to 58 (73.3% - 70.0% = 3.3% for waivered compared to 74.7% - 71.3% = 3.4% for non-waivered).

Tables 11 and 12 also show that further raising the NAPT score of 58 to 62 resulted in a higher graduate rate gain for the students with at least an academic waiver than for those without any waivers. For example, raising the NAPT score from 58 to 62 in Table 11 results in a 4.1% gain in the graduation rate (77.4% - 73.3%) compared to a 2.5% gain in Table 12 (77.2% - 74.7%). The graduation rate gain differences, however, may merely be due to the small number of students remaining qualified in Table 11 at these high aptitude levels and thus small sample instability. At an NAPT cutscore of 62, 246 students remained qualified in Table 12 compared to only 93 students in Table 11 (indicating a recruiting stress issue at this high of a NAPT cutscore).

We recognize that the similar initial pipeline graduation rates for the Table 11 and Table 12 contrasts for the MM rating (academically waivered versus non-waivered students) is not consistent with the disparate graduation rates reported in Table 7. However, Table 7 combines all of the ratings and applies to the more recent FY08-09 time frame whereas the cutscore analysis extracts just the MMs, and for the broader FY06-09 time frame (to obtain larger sample sizes). The results of the MM cutscore analysis are encouraging in that it is expected that MM pipeline graduation rates would improve for both academically waivered students and non-waivered students if the NF community raised the NAPT cutscore (not necessarily as high as 58).

The second set of empirical cutscore analyses was applied to the VE+AR+MK+MC composite for the students who were Auto-qualified, realizing that the 252 cutscore (required for only one of the two NF ASVAB composites) was exceedingly high and that even a modest cutscore raise would deplete the pool of NF qualified Navy recruits. Thus, the analysis was limited to a comparison of the academically waivered and non-waivered groups merely to confirm the high validity results. That is, with high validity we would expect substantial graduation rate improvements from raising the ASVAB cutscore, given there was substantial non-graduation rates to begin with.

Tables 13 and 14 show that MM students with at least an academic waiver and those without any waivers had, contrary to the Table 11 and 12 comparisons, dissimilar initial pipeline graduation rates (78.0% versus 88.9%, respectively) (but consistent with Table 7). The substantial negative impact of academic waivers on Auto-qualified student performance may be due to an interaction with actual reasons for the academic waiver and the MM "A" School lack of content. However, analyzing the logged academic reasons was beyond the scope of this study (we understand that NNPTC is looking at ways to efficiently code the data for future studies).

Because of the different initial graduation rates in Tables 13 and 14, comparisons of graduation rate improvements from raising the VE+AR+MK+MC cutscore is not appropriate, especially given the already high graduation rate for the non-waivered students (88.9% for Table 14 - not much room for improvement). Table 13, however, supports the validity analysis reported in Table 10 – raising the VE+AR+MK+MC cutscore of 252 by just 4-score points (to 256) results in a 4.5% gain in the graduation rate (85.8% - 81.3%), which is substantial. We note that the lowest VE+AR+MK+MC scores (228) in both tables are not an indication of an issued ASVAB waiver, but occurred because the two ASVAB composites are not perfectly correlated. All students with VE+AR+MK+MC scores lower than 252 qualified at 252 on AR+MK+EI+GS (an alternative standard).

Table 13
VE+AR+MK+MC Cutscore Analysis for MM ASVAB Auto-qualified Students who Graduated or Dropped for Any Reason and had at Least One Academic Waiver (FY06-09, N=460)

VE+AR+MK+MC Score	Graduates (#)	Non-Graduates (#)	Graduation Rate (%)	Students at or Above the Score (#)	Students of Total Above the Score (%)
260	130	21	86.1	151	32.8
256	**212**	**35**	**85.8**	**247**	**53.7**
252	283	65	81.3	348	75.7
250	302	74	80.3	376	81.7
246	328	83	79.8	411	89.3
238	350	93	79.0	443	96.3
.	.	.	.	.	.
228	359	101	78.0	460	100.0

Table 14
VE+AR+MK+MC Cutscore Analysis for MM ASVAB Auto-qualified Students who Graduated or Dropped for Any Reason and had No Waivers (FY06-09, N=839)

VE+AR+MK+MC Score	Graduates (#)	Non-Graduates (#)	Graduation Rate (%)	Students at or Above the Score (#)	Students of Total Above the Score (%)
260	281	23	92.4	304	36.2
256	**413**	**41**	**91.0**	**454**	**54.1**
252	563	68	89.2	631	75.2
250	596	73	89.1	669	79.7
246	646	81	88.9	727	86.7
238	722	92	88.7	814	97.0
.	.	.	.	.	.
228	746	93	88.9	839	100.0

Theoretical Cutscore Analysis

The two empirical cutscore analyses conducted in the last section for the ASVAB Auto-qualified groups (Table 13 for waivered; Table14 for no waivers) showed some VE+AR+MK+MC scores below the 252 cutscore, but with those students qualifying on the alternative ASVAB standard, 252 or above on AR+MK+EI+GS. Thus, there is no empirical way to evaluate the impact of lowering the VE+AR+MK+MC 252 cutscore given there is an aptitude/ability floor imposed by the alternative standard.

The fact that there is no empirical method for evaluating the impact of lowering the 252 cutscore on either standard due to the alternative standard's cutscore would make it impossible to estimate NF training performance decrements if, at some point in time, there were to be a severe downturn in the military recruiting environment. That is, there might be a time when the NF community has to, by exception to policy, issue an ASVAB score point waiver specifically for those ASVAB Auto-qualified. There is, however, a theoretically based cutscore analysis procedure that is based upon the mathematical and statistical relationships used to form the graphs in Figure 1. There are an unlimited number of combinations of parameter values that can be used to form the graphs because the parameters can vary in value to the nth degree (with values reaching infinity). These parameters are: (1) the validity of the aptitude measurement instrument, (2) the selection ratio, or qualification rate in our case resulting from a cutscore applied to the ASVAB composites, (3) the observed success rate for those selected – or graduation rate observed in our schools, in our case, and (4) the base rate, or the graduation rate that would apply for the total applicant population had all members been allowed to attend the NF school without adhering to any aptitude standard (or a smaller randomly drawn sample). The graphs are merely visual representations of the selection situation and the applied parameter whereas the quadrant values associated with those parameters are taken from known bivariate relationships published for convenience as an abbreviated set of tables called the Taylor- Russell (1939) tables.[30]

There are ten published Taylor Russell Tables, each referencing a different base rate. The tables can be used for several purposes. One purpose is to assess the expected improvement in the current success rate by replacing an operational selection instrument with one that has higher validity. Another purpose is to assess the impact on the current success rate by either raising or lowering an existing cutscore on the operational selection instrument. We used the Taylor Russell Tables to assess the impact of lowering the 252 cutscore on VE+AR+MK+MC making several practical assumptions including that the ASVAB is normally distributed in the population and that all of the non-graduation that occurred was due to academic issues (albeit we know that this is not the case), and also that there is an underlying continuum of performance scores (GPA) that has guided the decision to graduate or drop a student. Finally, we picked several base rates that we thought reflected the difficulty of the NF training, which, of course, cannot be verified.

The following parameter estimates were used in the Taylor Russell analysis: (1) a conservative ASVAB validity estimate of .80 from our correction for range restriction procedure (Table 10); (2) the observed 89% graduation rate (Table 14) for the non-waivered Auto-qualified group; and (3) a qualification rate of about 7% that results from the 252 cutscore, not in the recruit population, but in the ASVAB normative population from which we, theoretically, will be selecting future NF candidates – our base rate population. Table 15 gives the results of the Taylor Russell analysis.

Table 15 shows, internal to the table, four rows of entries taken directly from four different Taylor Russell base rate tables (.15, .20, .25, and .30 listed in the first column). The base rate that most likely portrays the NF situation, and that best fitted table from our study parameters appears to be .20 (in bold). That is, at a base rate of .20 our observed 89% success rate is found at the intersection of a validity of .80 (column to the right of the base rate column) and a .05 selection ratio (qualification proportion – "Qual(SR)" in the header above the table entries). The ASVAB cutscore that is associated with the .05 proportion, or 5% qualification rate, is 257, which is 5-score points higher than our 252 cutscore, not a concern for this analysis.

[30] Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565-587.

Table 15
Compilation of Taylor Russell Table Values for a Theoretical Cutscore Analysis

Taylor Russell Base Rate Table	Validity	VE+AR+MK+MC Cutscore Applied to the ASVAB Normative Youth Population				
		257	245	236	229	223
		Qual(SR) = .05	Qual(SR) = .10	Qual(SR) = .15	Qual(SR) = .20	Qual(SR) = .25
.15	.85	.88	.76	.66	.58	.51
.20	**.80**	**.89**	**.79**	**.71**	**.65**	**.59**
.25	.75	.89	.81	.74	.69	.64
.30	.70	.89	.82	.76	.72	.67

Note. The base rate is the success rate that would be observed if all individuals in the population of interest were selected for a job (or training, in our case) without applying an aptitude standard. Each line in our table references a different base rate Taylor-Russell table. Qual(SR) refers to the qualification rate, or selection ratio associated with a particular cutscore.

The 257 cutscore in Table 15 is 5-score points higher than the operational 252 cutscore, but for all intents and purposes Table 15 shows the main point of the analysis.[31] Lowering the 257 cutscore to 245 (12 score-points) to qualify the top 10% of the applicant population, rather than the more restrictive top 5% lowers the expected graduation rate from 89% to 79%, which is 10% lower. Table 15 shows that the expected loss in graduation rates is not the same for fixed qualification rate increments (.05) - from most restrictive to least restriction (.05 to .10 to .15, etc.), so interpolating the expected graduation rate loss from lowering the 252 cutscore is not a linear process. Roughly, a 4-score point reduction (which amounts to the current restrictive Navy ASVAB waiver policy guidance of 1-score point per subtest) could translate into about a 3% reduction in the graduation rate (1/3 of the 12 point difference between 257 and 245 translates into, roughly, about 1/3 of the 10% lower expected graduation rate – from 89% to 79%).

Although the analysis conducted from Table 15 makes many assumptions, we conclude that there could be real consequence in lowering the 252 cutscore, but depending upon the waiver status of the group the consequences may be tolerable. That is, a 4-score point waiver on the 252 cutscore might be tolerable only for NF candidates who have no academic waivers (or others to be safe).

Table 16 was constructed to show NF qualification rates for four different cutscores in two Navy accession populations recruited in relatively difficult (FY98) and easy (FY09) recruiting environments.

[31] The ASVAB cutscores and associated proportion of qualified youth were derived from normal curve deviates commonly found in the appendices of statistical textbooks. For calculating the qualification rate resulting from a 252 cutscore we must know the mean and standard deviation of the composites in the population, which are 200 and 35, respectively (for both composites). The 252 score is about 1.485 standard scores (Z-Score) above the mean, which has a Z-Score mean of zero. Referencing the normal curve table we see that about 7% of the population scores at or above a Z-Score of 1.485. Likewise, scores of 257, 245, 236, 229, and 223 were calculated as "Z" scores and placed above the Taylor-Russell selection ratio, or qualification rate (Qual(SR)) in Table 15.

Table 16
Qualification Rates for VE+AR+MK+MC and AR+MK+EI+GS for FY98 and FY09 Navy Accessions

	FY98 (N=47,938)		FY09 (N=38,694)	
ASVAB Composite	**Qualified (#)**	**Qualified (%)**	**Qualified (#)**	**Qualified (%)**
VE+AR+MK+MC				
257	1,007	2.1	1,741	4.5
252*	2,109	4.4	2,747	7.1
245	4,410	9.2	4,643	12.0
236	8,533	17.8	8,164	21.1
AR+MK+EI+GS				
257	1,007	2.1	1,817	4.7
252*	2,013	4.2	2,747	7.1
245	4,027	8.4	4,605	11.9
236	7,718	16.1	8,010	20.7
Either Composite:				
257	1,456	3.0	2,387	6.2
252*	2,787	5.8	3,555	9.2
245	5,374	11.2	5,777	14.9
236	9,817	21.5	9,744	25.2

*Operational NF ASVAB cutscore for the Auto-qualified selection method.

Table 16 shows that in FY98 only 2.1% of recruits would have qualified at the very stringent 257 cutscore, compared to 4.5% in FY09. The 4.5% qualification rate for FY09 is very similar to the 5.0% (.05 in Taylor Russell Table terminology) in Table 15, which suggests that recruiters in the current timeframe are successfully tapping into the high ability levels of the available youth population. All qualification rate comparisons at the four listed cutscores (257, 252, 245, and 236) were higher for FY09 than for FY98.

Also in Table 16, qualification rates were higher applying the alternative ASVAB standards model (qualified on either composite at a given cutscore) than with each individual ASVAB composite with cutscore applied separately. For example, at the operational 252 cutscore for FY09, 7.1% of the recruits qualified on both the VE+AR+MK+MC and AR+MK+EI+GS composites, but used as alternatives, 9.2% qualified, for an improved qualification rate of 2.1% (9.2% - 7.1%). The 2.1% improvement amounted to 808 more recruits NF qualified (3,555 – 2,747), a substantial opening of the recruiting aperture but with a substantial improvement in graduation rates compared to the earlier 1998/1999 evaluation of the previous ineffective NF ASVAB/NAPT standards.

ASVAB and NAPT Cutscores Combinations

Table 17 was developed to observe the impact of various combinations of separate ASVAB and NAPT cutscores on students remaining NF qualified, their performance (pipeline graduation rates) constraining the total ASVAB + NAPT cutscore at the current 290.

Table 17
Pipeline Graduation Rates at Various ASVAB and NAPT Cutscores for NAPT Required Students (FY06-09 Accession/NNPTC Matches, N=2,739)

ASVAB/NAPT Standard	Pipeline Graduation Rate	Qualified on the Standard	Qualified on the ASVAB
VE+AR+MK+MC or AR+MK+EI+GS = 225 and NAPT = 65	81.1%	445, (16.2%)	2,739, (100%)
VE+AR+MK+MC or AR+MK+EI+GS = 230 and NAPT = 60	79.2%	1,088 (39.7%)	2,644 (96.5%)
VE+AR+MK+MC or AR+MK+EI+GS = 235 and NAPT = 55	75.7%	1,839 (67.1%)	2,431 (88.8%)
VE+AR+MK+MC or AR+MK+EI+GS = 240 and NAPT = 50	76.3%	1,929 (70.4%)	1,929 (70.4%)
VE+AR+MK+MC or AR+MK+EI+GS, no cutscore, and NAPT = 50	73.8%	2,739 (100%)	2,739 (100%)

Note. The dataset has limitations because all students had to score at least 290 on ASVAB+NAPT.

Table 17 shows the current operational standard in the bottom row (no cutscore for the ASVAB and a 50 cutscore for the NAPT). The four combinations of ASVAB and NAPT scores that total 290 are shown in the rows above: (1) 240 for either ASVAB composite and 50 for NAPT, (2) 235 for ASVAB and 55 for the NAPT, (3) 230 for ASVAB and 60 for NAPT, and (4) 225 for ASVAB and 65 for NAPT.

Table 17 shows the 73.8% pipeline graduation rate for the total sample of 2,739 NAPT required students. The pipeline graduation rate for the imposed (now explicit) 240/50 cutscore combination was 76.3%, which is 2.5% higher than the original 73.8% graduation rate where higher than 50 NAPT scores were allowed to compensate for relatively lower ASVAB scores. The tradeoff of imposing the 240 ASVAB cutscore is that 1,929 (70.4%) of the original sample (N=2,739) remained qualified, which means 29.6%, or 571 students, did not. Notionally, recruiters would have to fill the 571 student gap by either recruiting more high scoring ASVAB youth, or negotiating with other rating communities to increase the NF share of very high ASVAB scoring recruits (all students already had an NAPT score of 50 or higher, so recruiting more with high NAPT scores would not fill the gap).

The pipeline graduation rate for the 235/55 combination was a slightly lower 75.7% when compared to 76.3% with the 240/50 score combination; however, the 75.7% graduation rate is still 1.9% higher than the original 73.8%. Compared to the 240/50 cutscores, the 235/55 cutscores retained a lower percentage of students qualified (67.1% for 235/55 versus 70.4% for 240/50), but the difference in the number of students is only 90 (1,929 – 1,839). In contrast, when considering the number of students who remained ASVAB qualified, the 235/55 cutscores retained a much larger 502 students qualified than the 240/50 cutscores (2,431 – 1,929), which means recruiters would not have as a high burden in recruiting more high scoring ASVAB youth with that 235/55 combination.

The pipeline graduation rate for the 230/60 combination was a much higher 79.2% when compared to 75.7% with the 235/55 score combination, and is 5.4% higher than original 73.8%. The tradeoff, however, is only 1,088 students (39.7%) from the original sample remained qualified due mainly to the lack students with NAPT scores of 60 or higher. In contrast, 2,644 students (96.5%) from the original sample remained ASVAB qualified at 230. Recruiters would have to fill the large student gap by recruiting more high scoring NAPT youth rather than by negotiating with other rating communities to increase the NF share of very high ASVAB scoring recruits. The pipeline graduation rate for the 225/65 combination is a still higher 81.1%, when compared to 79.2% for the 230/60 score combination, but the tradeoff is severe. Only 445 students (16.2%) from the original sample remaining qualified, meaning NAPT becomes marginally useful to the NF for recruiting purposes despite its demonstration of high validity.

Although Table 17 is highly specific to existing data, we concluded that higher standards result in higher graduation rates, but that there is a cost on the recruiting side for specific cutscore combinations. Even in the current positive recruiting environment, recruiters would have needed to replace many students to make goal if the NAPT score had increased from 50 to 60. The ASVAB and NAPT 235/55 cutscore combination appears a conservative standard in that it is expected to increase NF course graduation rates without a severe impact on recruiting.

"A" School GPA Relationship to Power School Performance

Finally, we examined the relationship between "A" School GPA and Power School GPA for all students who graduated from Power School. The purpose of this analysis was to provide insight into the relative capabilities of graduates from the MM, EM, and ET "A" Schools and how NNPTC might be advised in equalizing the probabilities of all "A" School graduates passing the Power School where we observed these probabilities are lower for MM.

NNPTC currently scales and standardizes the three rating's "A" School GPA so that the means and variation (variance) about them are consistent across the schools. However, this process does not adjust for the inherent difficulty differences in these schools, all other things being equal, or how achieving a particular GPA from each "A" School impacts performance in Power School. To address the latter we applied linear regression, not hierarchical as in Tables 8 and 9. Table 18 gives the results for each rating applying their respective "A" School GPA as a predictor of Power School GPA. (Only Power School graduates were used in the analysis.) No adjustment was made for GPA scaling differences.

Table 18
Linear Regression Model Summary of
"A" School GPA Predicting Power School GPA
(Includes only Graduates from Power School)

Rating	Correlation	R-Squared	Adjusted R-Squared	Standard Error of Estimate
MM (N=2,543)	.792	.627	.627	.195
EM (N=1,204)	.837	.700	.700	.172
ET (N=1,007)	.844	.713	.712	.171

Note. The adjustment to the R-squared is for sample size.

Table 18 shows a high correlation between the "A" and Power School GPAs for EM and ET (.837 and .844, respectively) and slightly lower, but still a high correlation, for MM (.792). The percentage of Power School GPA accounted for by the "A" School GPA is also substantial (62.7% for MM; 70.0% for EM, and 71.2% for ET – adjusted for sample size). The slightly lower MM values are consistent with a content gap between the MM "A" and Power schools.

As briefly mentioned before for the hierarchical regression analysis, the standard error of estimate can be used to predict with a certain degree of confidence what Power School GPA is likely given a specific "A" School GPA. For example, a 3.0 "A" School GPA received by an ET student is plugged into the ET regression equation formulated from the Table 18 data (Estimated Power School GPA = .989 X "A" School GPA - .004) to get the predicted Power School GPA of 2.96. However, there is an error associated with the prediction, and in the ET case the standard error is .171 (for MM it is .195). The standard error is used to build a confidence interval about the predicted value and this confidence interval will contain the observed Power School GPAs for a large proportion of students. The larger MM standard error or estimate means there is lower precision in the MM Power School performance estimation.

Table 19 was developed to observe the linkage between the "A" and Power School GPAs and the resulting pipeline graduation rates for each rating. The data used for Table 19 included all students who graduated from their "A" Schools with a GPA of 2.50 or higher. The table shows that, at an "A" School 2.50 GPA graduation requirement, NAPT required MM students had an 81.2% pipeline graduation rate compared to 89.2% for MMs who were Auto-qualified. In order for MM NAPT required students to achieve a pipeline graduation rate comparable to the counterpart MM Auto-qualified students, the "A" School GPA graduation standard would have had to have been raised to 2.80. The 2.80 GPA requirement for the MM NAPT required students produced an 88.1% pipeline graduation rate.

The difference in the 81.2% and 89.2% pipeline graduation rates between the two MM selected groups is 8%, not inconsistent with the approximate 10% spread we have observed in previous analyses. The tradeoff of obtaining the 8% or so improvement in the MM Power School graduation rate from a raise in the MM "A" School GPA passing requirement (2.50 raised to 2.80) for NAPT required selection is a reduction in the number of students that would have graduated from the "A" School and therefore would not have continued on in the NF training program. There were 1,775 MM students (NAPT required) who had an "A" School GPA of 2.50 or higher whereas there were 307 fewer students (1,775 - 1,468) who had the 2.80 "A" School GPA. Many of those 307 students completed the NF training pipeline. The NF community may conclude that the benefits are enough (possible 8% - 10% increase in the Power School graduation rate) to justify incurring the early student losses (for the MM NAPT required students at the current ASVAB/NAPT standard), or improve the MM "A" School training (including extending time to train) rather than expend the resources required to recruit many more youth with higher ASVAB and/or NAPT scores.

Table 19 should be replicated yearly by NNPTC to gage academic quality in the NF "A" Schools for the two selection methods. For example, to the degree that NNPTC has identified competencies associated with a particular "A" School GPA level (or Power School GPA), the more disparate the performance between the two groups selected by different methods, the more likely that disparity translates into not only disparity in performance in Power School, but performance further out. For Table 19, 90.6% of the MM students who were ASVAB Auto-qualified performed at an MM "A" School GPA of 2.90 or higher, contrasted to 76.5% of the MM students who were required to take the NAPT. The difference may translate into real differences in NF capabilities.

Table 19
Pipeline Graduation Rates by "A" School GPA Level for "A" School Graduates Broken out by Rating and Selection Method (NNPTC Data, N=7,519)

"A" School GPA and Above	Rating	Auto-Qualified Selection Method A: Drawn from 4,210 "A" School Grads with GPA 2.50 or Above		Power School Graduation Rate Gap Between the two Selection Methods	NAPT Required Selection Method B: Drawn from 3,309 "A" School Grads with GPA 2.50 or Above	
		Grad Rate (%)	Students (#)		Grad Rate (%)	Students (#)
2.90	MM	92.9	2,028	2.6	90.3	1,313
	EM	94.5	977	2.3	92.2	696
	ET	94.7	808	0.2	94.5	524
			N=3,813 or 90.6% of total			N=2,533 or 76.5% of total
2.80	MM	91.6	2,130	3.5	88.1	1,468
	EM	93.9	1,027	1.7	92.2	755
	ET	94.5	837	2.5	92.0	589
			N=3,994 or 94.9% of total			N=2,812 or 85.0% of total
2.70	MM	90.4	2,214	4.8	85.6	1,613
	EM	93.2	1,057	2.4	90.8	822
	ET	93.9	851	3.8	90.1	638
			N=4,122 or 97.9% of total			N=3,073 or 92.9% of total
2.60	MM	89.8	2,245	7.0	82.8	1,720
	EM	93.5	1,075	4.3	89.2	854
	ET	90.1	862	1.7	88.4	665
			N=4,182 or 99.3% of total			N=3,239 or 97.9% of total
2.50	**MM**	**89.2**	**2,262**	**8.8**	**81.2**	**1,775**
	EM	92.3	1,084	3.4	88.9	864
	ET	93.4	864	5.5	87.9	670
			N=4,210, or 100.0% of total			N=3,309 or 100.0% of total

Note. ASVAB scores were not necessary for this analysis so all of the NNPTC data were used after selecting only "A" School graduates with an "A" School GPA of 2.50 or higher. The table can be viewed as a course performance quality benchmark tool (updated yearly) for assessing shifts in performance over time or between selection methods.

Summary and Conclusions

This study incorporated a mix of statistical tests and procedures, empirical analyses, and theoretically based analyses applied to data that came from several sources. A summary of the findings and our conclusions follow:

1) The process for assigning Nuclear Field ratings at RTC seems fair and unbiased and students usually receive their first or second rating preference. The process does not seem to cause disparity in representativeness of recruit characteristics or waiver incidence (except for a slightly higher academic waiver rate for MMs).

2) The ASVAB alternative composites (VE+AR+MK+MC and AR+MK+EI+GS) have equally high validity (about +.80 to +.85) in predicting "A" School GPA; and the validity was found to be as high as or higher than the other ASVAB composites evaluated.

3) The NAPT was evaluated as having utility as an alternative standards screen to increase the pool of qualified NF candidates as it measures NF relevant constructs at the upper aptitude/ability range where ASVAB has gaps. However, NAPT scores are weakly related to the ASVAB as a consequence of the 2- tiered selection methods resulting in ASVAB and NAPT score floors and caps (NAPT scores below the 50 cutscore were not captured in the data). Therefore, we could not appropriately estimate the NAPT validity coefficient.

4) Because validity of the ASVAB is high, graduation rates are expected to decline significantly if the operational 252 cutscore for the Auto-qualified NF candidates is lowered. However, in the event of a recruiting downturn, a valid academic risk scale could be used to establish a several point ASVAB waiver policy for those with demonstrated strengths and without an academic waiver requirement.

5) In general, students without academic waivers have higher graduation rates than those who do, and this appears to be true for both selection methods. Yet, by historical standards, the NF graduation rates are high, and very high for the Auto-qualified group, so it appears that recruiters are doing an excellent job identifying capable and motivated NF candidates.

6) MMs have a lower Power School graduation rate than the ETs and EMs. The supposition is that the MM "A" School content lacks some complex technical content taught in the ET and EM "A" Schools. MM failures in Power School may be reduced by adding some additional modules to the MM "A" School curriculum.

7) An alternative or adjunct to 8) for improving the MM Power School graduation rate for the NAPT required students is to raise the "A" School GPA requirement to graduate. The "A" and Power School GPAs are highly correlated and this strategy would result in fairly accurate training progression decisions. However, the strategy also amounts to differential treatment between schools on performance requirements and does not eliminate the need for recruiters to replace the MM students who would just fail the training pipeline earlier.

8) Several ASVAB and NAPT cutscore combinations were assessed for potential in improving graduation rates. A 5-score point raise for the NAPT was considered necessary, as was establishing a 235 cutscore on the ASVAB (where currently no cutscore exists), leaving the 290 combination cutscore in place. In principle, recruiters will benefit from more targeted NAPT testing because potential candidates will be clearly identified by the 235 ASVAB cutscore. There is no way, however, to determine the impact of the higher NAPT requirement at this point in time.

Recommendations

The following recommendations regarding the ASVAB and NAPT standards for the Nuclear Field (NF) ratings are addressed to CNO-13, Naval Reactors (NAVSEA-08), Navy Recruiting Command (NRC), the Enlisted Community Manager and Technical Advisor, and officials at the Naval Nuclear Power Training Command (NNPTC).

7) Retain the 252 cutscore for the ASVAB composites, VE+AR+MK+MC and AR+MK+EI+GS, required on either (not both) to "Auto-qualify" NF candidates.
8) Raise the NAPT cutscore to 55 (a 5-score point increase) and establish a 235 cutscore on either (not both) of the ASVAB composites. Keep the 290 total cutscore explicit in policy documents until circumstances for an ASVAB or NAPT cutscore waiver policy can be evaluated.
9) Follow-through with an NRC initiative in developing an academic waiver risk scale to help guide recruiters in determining whether or not to accept a candidate, and establish the reliability and validity of the scale in the future.
10) Develop a process to accurately log waiver information by recruiters and propagate this information in a reliable manner to the NNPTC database.
11) Consider adding a module to the MM "A" School that provides concentrated study time on difficult content/principles that are covered in the Power School but not addressed in the MM "A" School, which may add to the MM "A" School training time.

These recommendations are expected to result in an increase in the yearly NNPTC pipeline graduation rates, thereby eventually reducing the yearly NF recruiting goal. The recommendations are not expected to significantly impact recruiting; although there could be some impact from establishing a 235 ASVAB cutscore and 5-score point increase in the NAPT cutscore for those who are required to have scores on both instruments. However, if all of the recommendations are implemented and the academic risk scale eventually shows utility, it may be possible to remove the 235 ASVAB cutscore for candidates required to take the NAPT (retaining the 290 total cutscore) and even issue ASVAB waivers (4-score points at most) for the 252 cutscore where academic strengths are identified (and where academic waivers are not required).

By historical standards, the NF graduation rates are high, and there is no evidence that there should be a restriction on the percentage of students accepted with waivers, although the combination of an academic and other category of waiver does increase a student's risk for failure relative to having only an academic waiver. Given the high quality of the NF rating assignment system at Great Lakes, alternative strategies for mitigating the risk for this small group could be to (a) develop policy to exclude this waiver combination or (b) equip the schoolhouse to identify and track course performance early on and provide the necessary support. Given the already stellar training support observed at the schoolhouses during the study, the former strategy may be the most prudent.

Appendix C
Armed Service Vocational Aptitude Battery (ASVAB)
Standards: Mineman Rating

Armed Services Vocational Aptitude Battery (ASVAB) Standards: Mineman (MN) Rating

Janet D. Held
Navy Personnel Research, Studies, and Technology
NPRST (BUPERS-1) Millington, TN

Reviewed and Released by
David M. Cashbaugh
August 17, 2012

Armed Services Vocational Aptitude Battery (ASVAB) Standards: Mineman (MN) Rating

Janet D. Held
NPRST (BUPERS-1) Millington, TN

Introduction

The Chief of Naval Operations (N132G) at the request of the Mineman (MN) community tasked Navy Personnel Research, Studies, and Technology (NPRST/BUPERS-1) to review the Armed Services Vocational Aptitude Battery (ASVAB) standards for the MN rating. The reasons for the study were (a) observed Sailor difficulty in MN A-School training and (b) MN officials' concern that the ASVAB technical knowledge based test, Electronics Information (EI), is relevant for MN training but is not included as one of the MN's two alternative ASVAB composites. The last MN rating ASVAB standards study was conducted in 2005 to address a more difficult curriculum due to increasing electrical and mechanical complexity that also generalized to the MN job.[32] The current study is a review of the ASVAB standards put in place in 2005 and a more in-depth review of the current MN curriculum to determine if the standards adequately address aptitude/ability requirements.

The MN study is complicated due to a significant number of factors. First, there are diverse skill-set requirements for MN rating, both in training and on the job. About one third of MN billets are assigned to Navy Mine Countermeasure (NMC) Detachments (formerly Mobile Mine Assembly Units [MOMAU]). The other two thirds of the billets are assigned at Sea on Mine Counter Measure (MCM) ships. Shore duty appears to be an entry level assignment with skills narrowly applied to the mines themselves (maintenance and repair) whereas Sea duty is more technically diverse. However, even for the Sea billets, there are diverse skill-set requirements due to the large number of MN that man the MCM ships that are relatively small. Sea billets make up about 60% of the MCM ship crews and these billets have largely different responsibilities. For example, some MN may be responsible for assembling, testing, and maintaining the sub-assemblies that comprise the Submarine Launched Mobile Mine (SLMM) unit (that detects and destroys mines). Other MN may be responsible for tracking and classifying mine-like objects. Still other MN may be responsible for detecting and isolating equipment failures using computer aided software and correcting any faults or replacing failed modules and components. Still other MN may be primarily responsible for maintaining the ship's weapons arsenal.

The second complicating factor in the MN ASVAB standards study is that the MN A-School curriculum and advanced C-Schools' curricula are at least as technically complex as for some other Navy ratings with substantially higher ASVAB standards; for example, Missile Technician Submarine (MTS) and Sonar Technician Surface (STG). Not all MN actually perform jobs that are as technically complex; however, each MN student is trained as if he or she could in the event of an unplanned loss on a ship. Less than 200 MN are recruited each year so every MN counts and all must be prepared to take on extra technically complex duties. Further, many MN students graduating from A-School are now required to complete advanced C-Schools right after whereas before, a MN would usually attend C-School only after serving a Sea tour. Completing a Sea tour gives the MN the advantage of familiarization with MN systems but also provides a basis for demonstrating job skills to superiors making it a more objective decision for the community in deciding which MN should take advanced courses.

[32] Held, J. & Fedak, G. (2005). *Armed Services Vocational Aptitude Battery (ASVAB) Standards: Mineman Rating* (NPRST Letter Report Ser 3900, PERS-13/00013 22 Feb 2005).

The MN A-School directly followed by C-School training track was stood up in anticipation of many MN billets migrating to the Littoral Combat Ship (LCS). If LCS MN staffing accelerates with more ships being built, there may be many MN without Sea duty on the MCM ships and therefore without opportunity for gaining Sea duty experience. The only other Navy rating with this A-School directly followed by C-School training path is the Aerographer's Mate (AG). The AG rating requires C-School immediately after A-School in order for AGs to perform their jobs at their first duty assignment. In 2000, the ASVAB standard was raised for the AG rating so that more A-School graduates would graduate from the more difficult AG C-School. [33]

The third complicating factor in the MN ASVAB standards study pertains to the LCS itself. The LCS is a reduced manning ship that has a core crew augmented by mission specific modules, or "packages". Three mission modules (MMs) "plug into" the LCS, one mission at a time. The MMs are Mine Warfare (MIW), Surface Warfare (SUW), and Anti-Submarine Warfare (ASW). MN will fill twelve out of fifteen MIW billets.[34] The MN will have an array of responsibilities and many will require C-School and LCS training. Center for Naval Analysis (CNA) recently addressed the question of the feasibility of cross-training the three MM crews so as to lower the Navy's personnel costs.[35] It appears that the pros do not offset the cons, but in a tight budget environment it may become necessary to consider, at some level, the MM cross-training/assignment model. It should be noted that the skills and training required for crossing over to another MM may be similar (yet to be determined), but the success factors may be fatigue and personnel scheduling. It is also noted that the MN ASVAB standards are lower than the standards for the key ratings that man the SUW and ASW MMs [Sonar Technician Surface (STG) and Fire Control (FC)].

The fourth complicating factor in the MN ASVAB standards study is the diversity in plans for the MN rating. For example, several years ago a merger was proposed for the MN rating and Gunner's Mate (GM) rating; however, the proposal went only so far as the conduct of an occupational standards commonality study as part of a Navy Enlisted Occupational Standards (NEOCS) review [conducted by the Naval Manpower Analysis Command (NAVMAC)]. The ASVAB standards for GM are lower than for MN; and in turn, the MN ASVAB standards are lower than for STG. Although there are currently no plans in the future to conduct a MN rating merger, the MN community itself has been the recipient of conversions from other ratings as well as taking on these ratings' duties in an attempt to grow the MN rating. That is, the MN rating has consolidated the job functions of the Boatswain's Mate (BM), Operations Specialist (OS), Quartermaster (QM), Gunner's Mate (GM) ratings, and some duties of the STG rating. The MN can therefore be viewed as the original "Hybrid" Sailor, even though any one MN would not perform all of these ratings' duties aboard the MCM ships.

The fifth and final complicating factor for the MN ASVAB validation/standards study is the potential for a change in the Navy recruiting environment. The current recruiting environment is positive due to (a) the protracted downturn in the economy and unavailability of private sector jobs, (b) steadily increasing costs of a college education that many youth cannot afford, (c) the perception that the military offers concrete and accredited education and training during and after service, and (d) a rise in patriotism since 9/11. As a result of these factors, many more youth with some college

---

[33] Held, J. & Johns, C. (2000). *Validation of Armed Services Vocational Aptitude Battery (ASVAB) Selection Composites: Aerographer's Mate Class "A" School* (NPRST Letter Report Ser 3900, PERS-12/000094 11 Aug 2000).

[34] Four additional "Apprentice" billets may be added to the fifteen LCS MIW module billets for on the job training purposes.

[35] Sayala, S., Miller, & Stoloff, P. H. (2011). *Investigating the Feasibility of Cross-Training for LCS Mission Modules* (CRM D0026060.A2). CNA Report.

education, if not full degrees, have entered the military so that the annual ASVAB score distributions have shifted to the right. At the same time, the Navy has limited its recruiting of non-high school graduates, all of which means many more Navy recruits are classified into ratings with much higher ASVAB scores than are required (a large delta between Sailors' ASVAB scores and a rating's cutscore). If the economy were to completely recover, these large ASVAB score deltas observed for MN Sailors would shrink, as they would for other Navy ratings. The ASVAB has substantial validity in predicting training grades for technically saturated courses so there would be an overall increase in academically related training issues, including a decline in graduation rates (all other things remaining constant including training resources, curriculum, training time, training standards, and ASVAB standards).

All of these factors considered in context, a conservative approach was adopted with a forward looking strategy to incrementally evaluate and potentially change the MN ASVAB standard in two stages. The first stage is considered an interim period when there is not complete clarity on all factors that will influence the MN in their duties, including all of the training that applies to the LCS. Therefore, the main objective of this particular ASVAB standards study was to validate the ASVAB operational composites relative to other ASVAB composites that are more technically saturated to see which composites are most appropriate for the MN rating. A second stage follow-on tracking effort will occur to evaluate the effectiveness of the initial cutscore established in the first stage having instituted the most valid ASVAB composite(s).

Methods, Analyses, and Results

The study is organized into the following sections: (1) Description of the ASVAB; (2) Data collection; (3) MN major job duties; (4) Curriculum content; (5) ASVAB composite validities; and (6) Cutscore evaluation. The last section contains the recommendations.

Description of the ASVAB

The ASVAB is a mix of aptitude/ability/knowledge based tests that are used by all of the military services as their primary cognitive instrument for selecting military applicants and classifying them into enlisted occupations. The ASVAB was developed to predict training performance, and new tests are now being considered as additions to the ASVAB as technology and military jobs change. Table 1 on the next page gives a brief description of the current nine ASVAB tests and also the former ASVAB Coding Speed (CS) test that is now a Navy special classification test. Table 2 that follows lists the Navy's operational ASVAB composites including two that contain the CS test.

Each ASVAB and CS test listed in Table 1 have their scores referenced to the ASVAB normative youth population and standardized to have a mean score of 50 and standard deviation (SD) of 10.[36] The bulk of ASVAB test scores typically are in the range of 20 to 80. Scores on Word Knowledge (WK) and Paragraph Comprehension (PC) are combined to form the Verbal (VE) composite (also with mean 50 and SD 10).VE is part of the Armed Forces Qualification Test (AFQT) used to qualify military applicants for service (2VE+AR+MK) and is scaled as a uniform percentile distribution with scores ranging from 1-99. The Navy ASVAB composite scores are simply the sum of the scores on the individual ASVAB tests that form the composite.

[36] Segall, D. O. (2004). *Development and Evaluation of the 1997 ASVAB Score Scale* (Technical Report 2004-02). Seaside, CA: Defense Manpower Data Center. (http://official-asvab.com/norming_res.htm)

Table 1
Description of the ASVAB and Coding Speed Tests

Test Name and Abbreviation	Test Description
General Science (GS)	Knowledge of physical and biological sciences
Arithmetic Reasoning (AR)	Ability to solve arithmetic word problems
Word Knowledge (WK)[a]	Ability to select the correct meaning of words presented in context and correct synonyms
Paragraph Comprehension (PC)[a]	Ability to obtain information from written passages
Mathematics Knowledge (MK)	Knowledge of high school mathematics principles
Electronics Information (EI)	Knowledge of electricity and electronics
Auto and Shop Information (AS)	Knowledge of automobile and shop technologies tools and practices
Mechanical Comprehension (MC)	Knowledge of mechanical and physical principles
Assembling Objects (AO)	Ability to determine correct spatial forms from their separate parts and connection points
Coding Speed (CS)[b]	Ability to quickly identify correct word/number pairings from a key with many options

[a]WK and PC are combined to form the Verbal (VE) composite that is a component of the AFQT and several Navy ASVAB classification composites. [b]Coding Speed is a Navy special classification test not part of the ASVAB.

Table 2
ASVAB and ASVAB/CS Classification Composites

Composite Tests	Composite Names
General Technical	VE+AR
Administration	VE+MK
Hospitalman	VE+MK+GS
Electronics	AR+MK+EI+GS
Basic Electricity & Electronics	AR+2MK+GS
Nuclear Field	**VE+AR+MK+MC**
Engineering	**VE+AR+MK+AS**
Special Operations	GS+MC+EI
Mechanical	AR+MC+AS
Mechanical_2	MK+AS+AO
Operations	VE+AR+MK+AO
Business/Clerical	VE+MK+CS
Air Traffic Control	VE+MK+MC+CS

Notes. (1) The MN composites are in bold. (2) A composite in operational use that contains AO or CS is considered an alternative to a primary composite that does not contain either test, which is necessary because not all accessions are administered the AO and CS tests.

The MN ASVAB standard up until the 2005 NPRST study was VE+MC+AS ≥ 158. The current ASVAB standard resulting from the 2005 study follows an "alternative standards" model, VE+AR+MK+MC ≥ 210 "or" VE+AR+MK+AS ≥ 210. The two composites, called the Nuclear Field and Engineering composites, respectively, measure a mix of verbal, arithmetic reasoning, mathematics knowledge, mechanical, and auto/shop content domains. The Auto and Shop Information (AS) test is knowledge based and measures an individual's engagement and possibly interests in those subjects. The Mechanical Comprehension (MC) test is more a measure of mechanical aptitude, although not devoid of knowledge from experiences. The use of the two composites as alternatives tends not to "penalize" individuals without actual AS knowledge.

The ASVAB Electronics Information (EI) test is contained in just one of the Navy's operational ASVAB composites, AR+MK+EI+GS (the Electronics composite). The EI test, like AS, is a knowledge/experience test. The MN community expressed concerns at the time of this ASVAB standards study that neither of the two MN ASVAB composites contains EI; and EI appears to be relevant to the MN training and occupational standards. There are two aspects of EI that should be noted. First, recruits scoring high on the EI test would not necessarily be classified to the MN rating because the AR+MK+EI+GS composite, as with all Navy ASVAB composites, reflects a compensatory model where strengths in one area can offset slight weaknesses in another. As such, recruits with low EI scores could qualify if their other ASVAB scores are high. Recently, Avionics Electronics Technician (AT) rating instructors informed the NPRST ASVAB standards team during a Pensacola schoolhouse visit that students with high math skills can learn electronics and that it is not necessary for them to report to the schoolhouse with any EI knowledge. However, recruits classified into the AT rating have a much higher ASVAB standard to meet than the MN rating (VE+AR+MK+MC ≥ 222 "or" AR+MK+EI+GS ≥ 222 as alternatives) and are thus, in the aggregate, more able to absorb complex technical curriculum. But also, the AT training is longer and more devoted to electronics principles than MN training so there is more time to learn.

The second aspect regarding EI, and that applies to all ASVAB tests, is that the Navy does not, in almost all cases, apply a cutscore to any single test, mainly because (a) a single ASVAB test is a less reliable measure that a composite of ASVAB tests and (b) the composite, as a multiple construct measure, has higher validity than a single test when predicting learning of multidimensionality training concepts.

Data Collection

The MN A-School study data were obtained from two sources. The first data source was the Navy Integrated Training Resources and Administration System (NITRAS) database, whose interface is referred to as the Corporate Enterprise Training Activity Resource Systems (CeTARS) (data pulled for FY09 through partial FY11). The second data source was the MN A-School itself for the collection of performance grades, retest information, Academic Action Reviews (ARBs), failure information, and final school grade (FSG). FSGs, collected for FY09 and FY10, were used in the ASVAB validity analysis (relating ASVAB scores to FSG). The NITRAS/CeTARS data were only used to obtain ASVAB scores and to better understand the reasons for not graduating from the MN A-School. Table 3 provides a frequency distribution of schoolhouse disposition codes (Personal Event Codes) and associated reason obtained from the NITRAS/CeTARS data broken out by Accession and Fleet returnees.

Table 3
FY09-Partial FY11 NITRAS Mineman A-School Disposition

Personnel Event Code	Frequency	Percent
Sailors from the Fleet		
288 – Graduate	31	96.9
203 – ATTR non-academic (legal- misconduct)	1	3.1
Total	32	100.0
Accessions from RTC		
288 – Graduate	169	94.4
148 – ATTR non-academic (admin- unsuitability)	1	.6
198 – ATTR non-academic (legal- arrest by civil authorities)	1	.6
203 – ATTR non-academic (legal- misconduct)	5	2.8
232 – DSNRL non-academic (admin- non prereq security)	1	.6
320 – ATTR non-academic (motiv- neg. military attitude)	2	1.1
Total	179	100.0

Table 3 shows an exceedingly high graduation rate for both accessions from RTC and Fleet Sailors (although the Fleet sample size of N = 32 is small and therefore subject to greater sampling error). Fleet Sailors graduated at 96.9 percent, and similarly, accessions graduated at 94.4 percent. All non-graduation events were logged as due to non-academically related reasons, and so there is no evidence from the NITRAS/CeTARS data source of academic stress at the MN-A School, and thus no reason to hypothesize that the operational ASVAB standards are inadequate.

Another possible indicator of insufficient ASVAB standards is performance in the initial training prior to MN A-School. MN recruits take a core principles course right after basic training called Apprentice Technical Training (ATT). ATT is tailored for some 20 ratings and is attended at Naval Service Training Command (NSTC) in Great Lakes. MN ATT grades are not entered into NITRAS but it was observed that only two cases at most over the study time frame (FY09 through partial FY11) were dropped from ATT. Therefore, the MN ATT course was not considered a significant MN educational hurdle. On the other hand, ATT is does not provide in-depth learning opportunities for the broad based topics (discussed later) and is only about 9 weeks longs. Further, grades are not available in NITRAS to assess how well students do in the course.

MN Major Job Duties

Because there were no obvious indicators from the MN ATT and A-School data that the courses were so difficult so as to question the effectiveness of the ASVAB standard, a more in-depth approach was taken involving examination of the MN major job duties and how well they map to the training curriculum. The MN major job duties were taken from the MN "Hard Card". The Navy Classifiers traditionally use these laminated cards at the Military Entrance Processing Stations (MEPS) during applicant processing to inform applicants about Navy ratings.

Table 4 on the next page lists the MN major job duties (rows) as taken from the Navy MN hard card, and alongside in columns, a check mark for job duties that also apply to other ratings that have similar or more stringent ASVAB standards.

Table 4
Mineman Major Job Duties and Applicability to Selected Navy Ratings

MN Major Job Duties	MN	GM	EM	GSE	STG	MTS
Operate sonar systems for detection & classification of contacts	X				X	
Function in the minesweeping tactical nerve center (CIC) as part of the CIC team	X					
Handle and operate deck-loaded mine neutralization equipment	X					
Perform maintenance on and assemble mines	X					
Perform electrical and electronic checks and tests of circuitry and components	X	X	X	X	X	X
Solve complex electronic problems when tests fail	X	X	X	X	X	X
Work with basic mechanical test equipment	X	X				
Work with basic electronic test equipment	X	X	X	X	X	X
Operate various types of mine handling equipment such as forklifts, cranes and heavy transport trucks	X					
Operate various types of hand held equipment such as sandblasters, grinders and pneumatic torque tools	X	X			X	
Basic Engineering Common Core (BECC)			9	9		
Basic Enlisted Submarine School (BESS)						4
Apprentice Technical Training (ATT)	9.4	5.8	6.4	6.4	11.2	6.8
A-School Training Weeks	13	27			18	8
Strand Technical School				6		
Total Training Weeks	22.4	32.8	15.4	21.4	29.2	18.8

Note. Ratings are Mineman (MN), Gunner's Mate (GM), Electrician's Mate (EM), Gas Turbine Systems Technician (Electrical) (GSE), Sonar Technician Surface (STG), and Missile Technician Submarine (MTS).

Table 4 shows that the Gunner's Mate (GM), Electrician's Mate (EM), Gas Turbine Systems Technician (Electrical) (GSE), Sonar Technician Surface (STG), and Missile Technician Submarine (MTS) ratings have two common duties. These duties are listed as "Perform electrical and electronic checks and tests of circuitry and components" and "Solve complex electronic problems when tests fail". The MN, however are trained in a broad scope of duties compared to the other ratings (their other duties are not shown) that includes mechanical hands on equipment operations and maintenance.

Table 4 also lists the training weeks for each rating through A-School. All of the ratings in Table 4 attend their rating tailored ATT course; post ATT the ratings have largely different tracks. The EM and GSE ratings go through a Basic Engineering Common Core (BECC) course while the MTS rating goes through Basic Enlisted Submarine School (BESS). The GSE rating goes through strand training after BECC and then Sea duty before more advanced training. The STG rating has a slightly longer ATT training than MN (about 2 weeks longer) but also a significantly longer A-School training than MN (18 weeks of A-School for STG compared to 13 weeks for MN).

The length of A-School training is longest for the GM rating (about 33 weeks) followed by STG (about 29 weeks), followed by MN (about 22 weeks). The somewhat lower ASVAB standard for GM relative to MN appears to be compensated for by longer GM training time. The MN and GSE ratings seem comparable in training time and have the same ASVAB cutscore level, although not the same alternative ASVAB composites. The GSE rating (and EM) ASVAB alternative standards are VE+AR+MK+MC ≥ "or" AR+MK+EI+GS ≥ 210 and are considered in this study for adoption by the MN rating.

Curriculum Content

The MN A-School curriculum outline and testing plan were obtained from the schoolhouse for this study. The MN ATT module topics were taken from the course version that applied to the study period. Table 5 lists the MN ATT module titles for the technical portions of the course.

Table 5
Module Titles for the MN ATT Course

Mod 1-Introduction to Electricity	Mod 17-Transitor Amplifiers
Mod 2-Multimeter Measurements	Mod 21-Operational Amplifiers
Mod 3-Basic DC Circuits	Mod 23-Introduction to Digital Circuits
Mod 4-Complex DC Circuits	Mod 24-Digital Logic Functions
Mod 5-Wiring	Mod 25-Combinational Logic Functions
Mod 6-Introduction to AC	Mod 26-Microprocessors
Mod 7-AC Test Equipment	Mod 31-Basic Motors
Mod 12-Transformers	Mod 37-Fiber Optics
Mod 13-Relays and Switches	Mod 39-Hydraulic/Pneumatic Systems
Mod 14-Diodes and Diode Circuits	Mod 43-Basic Mathematics
Mod 15-Transitor Circuits	Mod 44-Algebra (4/10 topics)
Mod 16-Power Supplies	Mod 47-Computer Math (4/11 topics)

Note. The MN ATT course is 9.4 weeks and involves a large portion of CBT delivered trouble shooting/problem solving scenarios.

The MN ATT course topics listed in Table 5 allow, on average, one week of study for two modules. The CBT curriculum presentation is coupled with problem solving exercises, or scenarios, presented on computer "Cards". Instructors circulate in the computer laboratories to assist students who are having difficulties and to answer questions. There is also instructor led classroom time before CBT laboratories (blended solution) and classroom based testing (although, as mentioned earlier, the grades are not uploaded into NITRAS/CeTARS, the Navy's corporate training database).

Table 6 contains the MN A-School curriculum high level module topics.

Table 6
Major Curriculum Topics for MN A-School

MN Common Core and Core
Mod 1- Surface Mine Countermeasures (SMCM), Airborne Mine Countermeasures (AMCM) and Underwater Mine Countermeasures (SMCM) Familiarization
MN Common Core Mod II- Weapons Safety and Security
MN Core Mod II- Underwater Explosive Effects, Ship Vulnerability, Ship Protective Measures, Degaussing and Cathodic Protection
Mod III- Pressure Theory, Acoustic and Seismic Principles, Magnetic Theory, Minefield Types and MCM Planning for integrated operations
Mod IV- Ordinance Certification, Safety, Handling and Stowage
MOMAU Core
Mod I- Mine Types, Actuation Mechanisms and Maintenance
Mod II- MET Familiarization, Mine Maintenance Program and Assembly Technical Administration
Mod III- Mine MK 56, 62, 63 and 65 Familiarization
Shipboard Core
Mod I- SLQ-48 MNS
Mod II- Aids to Navigation, Sound and Distress Signals, Lights and Day Shapes, Rules of the Road
Mod III- Man Overboard, Basic Lookout, Search and Rescue
Deck Operations
Mod I- Minesweeping/ Neutralization Operations and Safety, Mechanical Minesweeping, Magnetic Minesweeping and Combination Influence Minesweeping
Mod II- Small Boat Operations, Anchoring and Astern Refueling
CIC Operations
Mod 1- Passive MCM, Q-Routes and Environmental Effects on MCM
Mod II- Physics of Underwater Sound Propagation, Sonar System Theory and Fundamentals, SQQ-32 Mine Hunting Sonar and UQN-4A Fathometer
Mod III- MIW Messages, SMCM DTE Sequence, Introduction to CIC
Mod IV- Piloting, Introduction to Plotting, Time and Time Zones
Mod V- Relationship of Time/Speed/ Distance, Intro. to Maneuvering Boards, Direction of Relative Motion, Speed of Relative Motion, Closest Point of Approach, Course and Speed, Point Along Track, Revised CPA and True Wind
Mod VI- Signal Book, Formations, Naval Warfare Library and External Communications
Mod VII- Introduction to Radar, SPA-25G Radar Repeater, Radar Scope Interpretation and Log Keeping
Mod VIII- Chart Scales, Types, Use, Symbols, Numbering, Cataloging, Maintenance, Chart No. 1 and Coastal Navigation
Mod IX- Navigation Detail, GCCS-M/ MEDAL, Global Positioning System and SSN-2 (V4) PINS

MN A-School performance assessment on the modules shown in Table 6 involved six exams and a final course exam. Each exam, including the final, is weighted 14.29 percent of the final school grade (FSG). One or two module exam failures results in four hours of remediation. One or two final exam failures results in twenty four hours of remediation. More than two failures results in an Academic Review Board (ARB). The data from the schoolhouse showed there must have been 5 ARBs within the first two modules of the MN Core curriculum for the subset of relevant data (12.5% ARB rate). These two Core modules involve acquiring deep knowledge of technical and electronics principles, which are not meant to be solely addressed in the Apprentice Technical Training (ATT, Table 5).

ASVAB Composite Validities

The objective of a Navy ASVAB validity analysis is to estimate (always with some statistical error) which composite of a set of ASVAB composites is most predictive of training performance, as measured by the final school grade (FSG). Validity as it applies to Navy ASVAB validation/standards studies refers to the correlation between scores on a particular ASVAB composite with scores on the FSG. Validity coefficients range from -1 to +1, for perfect negative and positive relationships, respectively. A zero validity coefficient means there is no ASVAB predictive relationship. The larger the validity coefficient the more accurately a cutscore can be set to reduce school failure rates or setback rates (failures and setbacks due only to academic reasons).

Among the many factors that impact the magnitude of the validity coefficient is the one of most concern for this study - the restriction of range in ASVAB scores that occurs from applying the operational ASVAB standard (composite with cutscore). Restriction in range of ASVAB scores lowers score variance, and because score variance is necessary to derive a correlation, suppresses that correlation from what would be observed for a full ASVAB range applicant population. And, it is the applicant population (not the school sample) for which future recruits must meet the ASVAB standard and for which the ASVAB cutscore is set.

The average ASVAB composite validity coefficient across Navy rating training, corrected for range restriction due to prior selection based upon an ASVAB cutscore, is about .55. The smallest ASVAB composite validity coefficient is about .25 for SEALs, for which the training has large physical and mental stamina components.[37] The largest validity coefficient is about .80 to .85 for the Nuclear Field ratings, Machinists Mate (MM), Electrician's Mate (EM), and Electronics Technician (ET), which all have large academic and technical components in their training. [38]

Figure 1 on the next page shows, notionally, the effect of restriction in range of ASVAB test scores on the ASVAB validity coefficient and how its unrestricted value applies to a full range ASVAB population.

[37]Held, J. (2011). *Armed Services Vocational Aptitude Battery (ASVAB) Standard: SEAL from Program to Rating* (Letter Report Ser 3900 BUPERS-1/00092, 11 Aug 2011). Millington: Navy Personnel Research, Studies, and Technology.

[38] Held, J., Alderton, D. & Britton, D. (2010). *Armed Services Vocational Aptitude Battery (ASVAB) and Navy Advanced Placement Test (NAPT) Standards: Nuclear Field (NF) Ratings* (Letter Report Ser 3900 BUPERS-1/00158 of 4 Jun 98). Millington: Navy Personnel Research, Studies, and Technology.

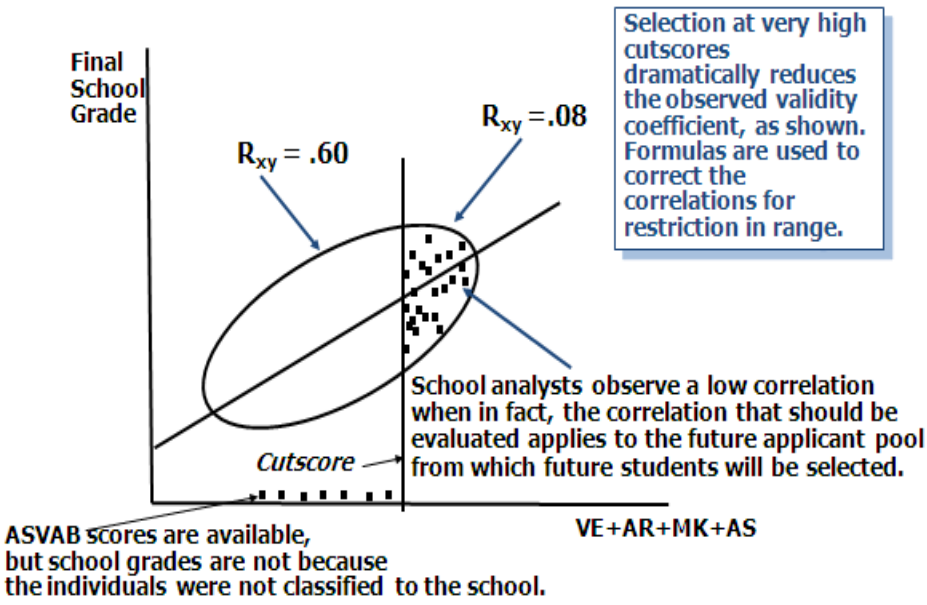

Figure 1
Notionally, the Effects of an Operational ASVAB Standard on the Validity Coefficient

Figure 1 shows what a hypothetical school analyst would observe as the relationship between ASVAB scores and FSG scores, represented by the students (dots) who have both ASVAB composite scores and final school grades (FSG). Because there is a cutscore applied to the operational ASVAB composite, in this case VE+AR+MK+AS, the elliptical sphere that applies, theoretically, to the applicant population from which MN are selected, is truncated to more or less a rounded sphere. The ASVAB score truncation reduces the correlation (validity coefficient) between the ASVAB and FSG scores. The estimated validity for the applicant population is calculated using statistical procedures that correct for ASVAB range restriction. [39]

A horizontal bar could be placed on the graph in Figure 1 to show the performance standard (FSG pass point) that the school places on a successful training status. However, in the case of the MN, the success rate is so high (about 95 %) as not to be instructive. Instead, Figure 2 on the next page is used to discuss the impact of a cutscore on school graduation rates given a substantial non-graduation rate exists, which could, theoretically, be managed by (a) a substantial increase in training resources to progress students through the training pipeline or (b) a higher ASVAB cutscore.

Figure 2 on the next page shows three graphs each depicting a different ASVAB validity coefficient (R_{xy} = .00, .55, and .85). Unlike the graph in Figure 1, the graphs in Figure 2 establish not only an ASVAB cutscore, but a performance score (FSG) that determines pass/fail status. For convenience, the dots that depict ASVAB/FSG score pairs are not included in the three graphs.

[39] Lawley, D. (1943). A Note on Karl Pearson's Selection Formula. *Royal Society of Edinburgh, Proceedings, Section A, 62*, 28-30.

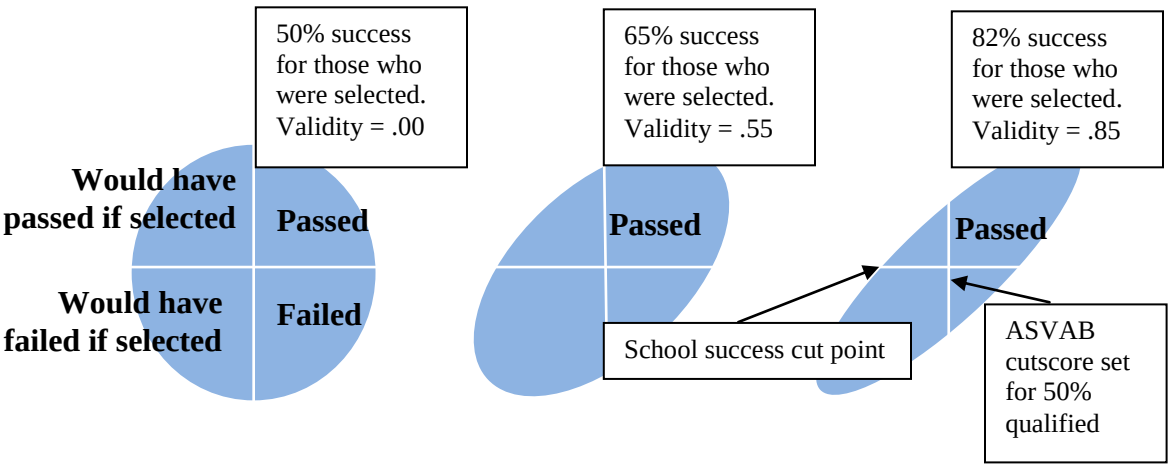

Figure 2
School Success Rate as a Function of ASVAB, all else being Equal

In Figure 2, the cutscore is set in each of the three graphs to qualify 50 percent of the applicants. In the leftmost graph, the ASVAB validity, R_{xy}, is .00 so the pass rate with and without an applied ASVAB cutscore is the same (50 percent). Obviously, raising the ASVAB cutscore to a very stringent level (say 10 % of applicants qualified) would not improve the 50 percent pass rate but would increasingly eliminate more applicants from being selected. This outcome under a zero validity scenario is what would be expected if applicants had been randomly assigned to Navy ratings without regard to a selection standard.

Assuming the same 50 percent qualification rate and 50 percent pass rate before an ASVAB standard cutscore was set, the middle and far right graphs show pass rate improvements when the validity of the ASVAB is non-zero, that is, R_{xy} = .55 and R_{xy} = .85. For R_{xy} = .55, the cutscore improves the pass rate to 65 percent. For R_{xy} = .85, the cutscore improves the pass rate to a much higher 82 percent. These "success rate" improvements were taken from the Taylor-Russell (1939) tables that are used by institutions to assess the value in using personnel selection instruments.[40]

An estimate of the validity for an "unrestricted population" based upon a selected sample is obtained using multivariate statistical correction procedures applied to the observed validity coefficient (reference in footnote 8). The procedure is uniformly used by all of the military services in deriving validity coefficients for ASVAB composites.[41]

The composites evaluated in validity analysis were the two MN operational composites, VE+AR+MK+AS and VE+AR+MK+MC, the Navy's Electronics composite, AR+MK+EI+GS, and also the Navy's Mechanical composite, AR+MC+AS (one test different from the prior VE+MC+AS used by the MN before the 2005 change – that composite has been eliminated for Navy use). Finally EI was included as a single test predictor to satisfy an inquiry by the MN community about EI's direct relevance. Table 7 on the next page shows the ASVAB range corrected validities for both the current study and the last MN ASVAB standards study (2005).

[40] Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565-587.
[41] Held, J. D. & Foley, P. P. (1994). Explanations for accuracy of the general multivariate formulas in correcting for range restriction. *Applied Psychological Measurement, 18,* 355-367.

Table 7
ASVAB Composite Validities Corrected for Range Restriction for the 2005 and 2012 MN Studies

Composite	2005 Validities Corrected for Range Restriction	2012 Validities Corrected for Range Restriction
VE+AR+MK+AS	.60	.71
VE+AR+MK+MC	.60	**.72**
AR+MK+EI+GS	.59	**.72**
AR+MC+AS	.55	.60
EI	-	.51
VE+MC+AS	.55	-

Note. The sample size for the 2005 study was 84; for 2012 the sample size was 225.

Table 7 shows that the two composites with highest validity for the current 2012 study are VE+AR+MK+MC and AR+MK+EI+GS, each with range corrected validities of .72. A validity of .72 magnitude is considered high. The validity for the alternative operational VE+AR+MK+AS composite was .71, which is lower than .72 by only a trivial .01. The validity for AR+MC+AS, the Mechanical composite similar to the prior VE+MC+AS composite that was operational for MN in 2005, was a much lower .60.[42] The about .10 validity difference between AR+MC+AS and the other three composites in this 2012 study compared to the about .05 validity difference for the 2005 study is most likely due to the current computer based training (CBT) that is more knowledge acquisition based than the earlier MN training that incorporated a large hands on laboratory component.[43]

The Mathematics Knowledge (MK) test that appears in the first three composites could be considered sufficient to learn electronics material, but as discussed earlier, compensatory models do no ensure high scores on any particular test. Also, the EI test cannot be discounted because students high in EI scores are sufficiently knowledgeable and with experience to effectively help others in the learning process (Peer to Peer tutoring). Peer to Peer tutoring is a productive model that both helps students in the classroom having difficulty but also further instantiate the tutor's own knowledge (a good model for Navy training in times of training resource shortages).

Finally, the validity of the single EI test was only .51 compared to the .72 and .71 that applied to the first three composites in Table 7. The .20 validity difference is very large and would have real negative classification decision consequences if EI were to be used either as a single selection instrument, or as an additional requirement layered on top of the higher validity ASVAB composites. Classification decision error increases as validity decreases, all other things equal.

[42] The VE+MC+AS composite was not evaluated in this study because it was eliminated from Navy use after demonstrating subpar performance in all of the ratings that used the composite.
[43] Subsequent to the data collection for this study, the MN laboratories were stood up to complement CBT in a blended training format.

Electronics/Electricity, while not predominant in the A-School curriculum, is important for ATT (Table 5) and C-School. And ATT as an apprentice course is not intended to provide in-depth knowledge. Additional support for the Electronics composite for MN use is in the content of the MN C-School (consideration of the LCS MIW courses is beyond the scope of the study). Table 8 shows the MN C-School courses that are attended not only by MN, but by the Sonar Technician Surface (STG) and Electronics Technician (ET) ratings that use AR+MK+EI+GS with a 223 cutscore as their sole ASVAB standard, much higher than would be proposed for MN.

Table 8
Electronically Based C-School Courses Attended by MN and other Ratings that use the AR+MK+EI+GS Composite

C-School	Ratings Attending
Precise Integrated Navigation Systems Maintenance	MN, STG, ET
Minehunting Sonar Set Maintenance Training	MN, STG, ET
Versatile Exercise Mine System Ashore (Operate & Maintain)	MN
Underwater Mine Test Set Maintenance Technician	MN

Table 8 shows substantive technical courses attended by the MN, STG, and ET ratings. The contrast could be made for the Aviation Mechanics ratings that have the same ASVAB standards as the MN rating (VE+AR+MK+MC and VE+AR+MK+AS composites with a 210 cutscore) in that their training does not involve electronics based training or nearly the level of technical content. The Aviation Mechanics ratings using VE+AR+MK+AS as their classification composites are Aviation Machinist' Mate (AD), Aviation Structural Mechanic (AM), Aviation Structural Mechanic – Safety Equipment (AME), Aviation Support Equipment Technician (AS), and Air Crew (formerly, AW).

Cutscore Evaluation

A cutscore analysis considers factors such as (1) the academic failure rate at the schoolhouse, (2) the academic setback rate, (3) the ASVAB waiver rate, (4) yearly school input requirement, (5) cognitive complexity of the rating, and (6) training time and cost.

With regard to academic stress (Factors 1 and 2), by the graduation metric (Table 3), there does not appear to be an academic problem at the MN A-School, However, there were indications of academic stress by incidents of ARBs reported by the MN A-School staff but also retest and low final school grades (FSGs) in the schoolhouse data.

With regard to MN ASVAB waivers (Factor 3), the rates were statistically different between Recruits and Fleet returnees, with Fleet returnees having higher ASVAB waivers. Of the NITRAS data pulled for (FY09 to partial FY11), 16 of the 32 MN Fleet returnees (50%) were not ASVAB qualified, which coincides with a growth in the MN rating that included members from other ratings with lower ASVAB standards [Boatswain's Mate (BM) and Ship's Serviceman (SH) ratings]. In contrast, 33 of the 278 MN students (12%) who came from the accession population (direct from Great Lakes) were not ASVAB qualified. Still, a 12 percent waiver rate appears large and does not coincide with the current positive recruiting environment.

With regard to MN annual throughput (Factor 4), the MN community has a relatively low annual recruiting goal at around 200. A continued positive recruiting environment will afford all Navy ratings, including MN, with high ASVAB scoring recruits; but then again the trend may not last much longer. Figure 3 shows the trend (over 5 years) of higher Navy recruit ASVAB scores on the three composites of interest in this MN study (all at the 210 cutscore).

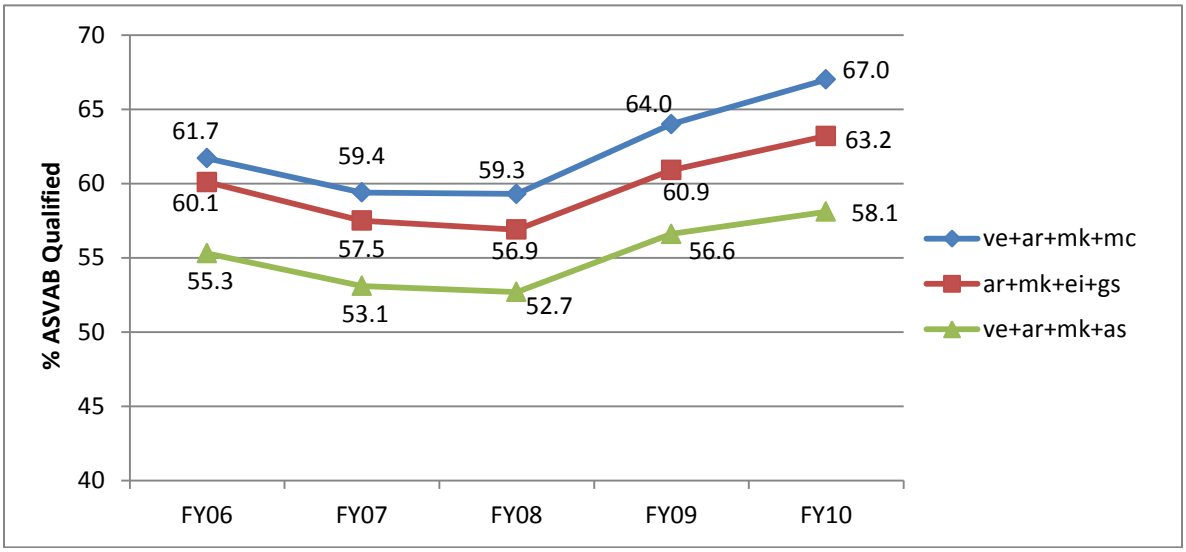

Figure 3
Navy Accession ASVAB Qualification Rates at a 210 Cutscore over Fiscal Years

Figure 3 shows a shift up in the percentage of accessions scoring at and above the 210 cutscore on the three composites starting in FY09 and continuing in FY10 (the period of the study data). Figure 3 also shows that the VE+AR+MK+AS composite had a lower percentage of accessions meeting the 210 cutscore than VE+AR+MK+MC and AR+MK+EI+GS, and, across all years due strictly to the AS test. That is, there are fewer youth accessing to the Navy with high Auto/Shop (AS) ASVAB test scores compared to high Mechanical Comprehension (MC) test scores, and in turn, the line representing the composite with Electronics Information (EI) scores is in the middle. The AS test is valuable as it contributes to differential assignment capability) for the Aviation Mechanics ratings (that use the VE+AR+MK+AS composite), but also for the Machinery Repairman (MR) and Machinist's Mate (MM) ratings (to name a few). The competition among the Navy mechanical based ratings for high AS scoring recruits may be the reason why the MN rating has a substantial ASVAB waiver rate and so eliminating VE+AR+MK+AS for used with MN classification may help both the MN rating and the other mechanically based ratings.

The cognitive complexity of the training (Factor 5) was considered high for many of the curriculum modules and major job duties linked with them. The total time to train for MN ATT through MN A-School was considered fairly short (see Table 4) compared to time allowed for other ratings with similarly difficult electronics based job duties. (A full evaluation of the other ratings' curriculum was out of scope of this study.) Finally, the cost of MN training (Factor 6) is considered moderate when compared to the length of training for the longer GM and STG training pipelines.

Because of the virtually non-existing MN academic failure rates logged at the schoolhouse (a tribute to the dedication of the training staff), there is no way to conduct an empirically based cutscore analysis, that is, to establish performance improvements (in terms of graduation rate improvements) associated with a higher ASVAB cutscore. This study, therefore, refocused on MN training content and major job duties that are technically complex in the context of the current availability of Navy accessions with high enough ASVAB scores. The training complexity was considered comparable to the STG and ET ratings that have more stringent ASVAB cutscores, but that also use appropriately matched ASVAB composites (VE+AR+MK+MC and AR+MK+EI+GS as alternatives). As a result of the study, the following recommendations are made for the MN rating's ASVAB standards.

Recommendations

The following recommendations regarding the ASVAB standards for the Mineman (MN) rating are addressed to CNO-13, the MN Enlisted Community Manager and Technical Advisor, the Mine Warfare Training Center (MWTC) officials, and Navy Recruiting Command (NRC).

1) Replace the operational alternative ASVAB standards (VE+AR+MK+MC = 210 and VE+AR+MK+AS = 210) with VE+AR+MK+MC = 210 and AR+MK+EI+GS = 210, also as alternatives.
2) Consider limiting ASVAB waivers given the annual MN goals are relatively small.
3) Conduct a follow-on ASVAB standards review to assess the adequacy of the 210 cutscore given the recently augmented MN A-School curriculum that incorporates CBT, instructor led classes, and hands on laboratories in a blended training format.
4) Synchronize the follow-on MN study to include performance measures in LCS Mine Warfare (MIW) training and the availability of MN ATT and C-School performance measures.

NPRST will monitor the youth ASVAB scores of future Navy recruits and also the effectiveness of the 210 cutscore in supporting the MN community.

Appendix D
INTERSERVICE Aptitude/Ability Standards Panel Charter

INTERSERVICE Aptitude/Ability Standards Panel

CHARTER

OBJECTIVE

Provide a mechanism for interservice collaboration and the exchange of technical information in the areas of personnel measurement, selection, and occupational classification in both research and applied settings, but primarily to address the setting of aptitude/ability standards as measured by the ASVAB and other special occupational classification instruments.

SCOPE

The Committee will serve as a forum for the development and validation of classification standards. It will review (1) past and ongoing personnel research that may be consolidated or applied in a joint-service setting, (2) endorse and apply best practices test validation and standard setting methodologies, including the development of best practice training and post-training performance criteria upon which predictors will be validated, (3) joint-service training transformation that may impact criterion quality or training outcomes, and (4) joint-service training ASVAB standards that may require alignment.

The Committee also will serve as a forum for sharing technical information about new methods to assess personnel qualification for Service occupations and the evaluation of new predictors. Information regarding these topics may be advanced to the Manpower Accession Policy Working Group (MAPWG). The committee also will consider alternative strategies for administering and operationalizing new predictors that may be reviewed by the MAPWG so as not to over-extend MEPCOM and DMDC time and resources.

MEMBERSHIP POLICY

The committee will be comprised, on a voluntary basis, of members from the Services' research, testing, manpower and personnel communities who are knowledgeable about test development and validation, but with an operational focus of improving occupational classification outcomes (FIT and FILL).

MEETING SCHEDULE

Meetings will take place at appropriate and logical locations (e.g., targeted Manpower, Personnel, and Training sites or the Service labs). Meetings also may be conducted by teleconference.

CHAIR SELECTION and TENURE

The committee Chair will be elected by simple majority of members attending the election meeting. The chair will serve a two-year term and may be reelected for another term. A Co-Chair also will be elected.

GENERAL PROCEDURE

Issues and agenda items will be submitted to the Chair, who will distribute the items to the committee members. The meetings will be informal without requirements for voting. A volunteer will take minutes and distribute them to the committee. The Chair will consolidate the meeting's activity and, along with the members, decide which agenda items will be submitted to the MAPWG. When appropriate, working groups will be established to address particular problems or issues.

July, 2012

Distribution List

AIR FORCE PERSONNEL CENTER, STRATEGIC RESEARCH AND ASSESSMENT BRANCH (HQ AFPC/DSYX)
AIR FORCE – FORCE MANAGEMENT POLICY DIRECTORATE, TRAINING AND EDUCATION REQUIRMENTS AND RESOURCES DIVISION (HQ AF/A1PT)
AIR UNIVERSITY LIBRARY
ARMY RESEARCH INSTITUTE LIBRARY
CENTER FOR NAVAL ANALYSES LIBRARY
CHIEF OF NAVAL PERSONNEL (OPNAV 132G, NAVY SELECTION AND CLASSIFICATION OFFICE)
DEFENSE MANPOWER DATA CENTER (CHIEF, PERSONNEL TESTING DIVISION)
MARINE CORPS MANPOWER AND RESERVE AFFAIRS, MANPOWER PLANS DIVISION, INTEGRATION AND ANALYSIS SECTION (SECTION HEAD)
MILITARY ACCESSION POLICY WORKING GROUP
NAVAL POSTGRADUATE SCHOOL DUDLEY KNOX LIBRARY
NAVAL RESEARCH LABORATORY RUTH HOOKER RESEARCH LIBRARY
NAVY PERSONNEL RESEARCH, STUDIES, AND TECHNOLOGY SPISHOCK LIBRARY
HQ USMEPCOM (DIRECTOR OF TESTING)
OFFICE OF NAVAL RESEARCH (CODE 34)
OFFICE OF THE UNDERSECRETARY OF DEFENSE (PERSONNEL & READINESS) (ASSISTANT DIRECTOR, ACCESSION POLICY)
USAF ACADEMY LIBRARY
US COAST GUARD ACADEMY LIBRARY