Defense Advisory Committee
On Military Personnel Testing

2018 Biennial Report

Office of the Under Secretary of Defense
(Personnel and Readiness)

July 2019

Contents

Executive Summary

The Defense Advisory Committee on Military Personnel Testing (DACMPT) met four times during the last two years to discuss progress on new developments as well as revisions to existing tests and processes. The current report summarizes the four DACMPT meetings occurring from July 2017 to March 2019. The DACMPT applauds the efforts of the Defense Manpower Data Center (DMDC) – now the Defense Personnel Assessment Center (DPAC) – for maintaining an active program of research and development related to the Armed Services Vocational Aptitude Battery (ASVAB) during a period of extreme budgetary limitations. The DACMPT also acknowledges the increases in funding of these efforts in the last two years, as use of the ASVAB is important in the maintenance of a high-quality military enlistment testing program. Full funding directly supports the ability to complete projects central to the continued quality and security of the testing program, and the DACMPT is very much encouraged that, since 2017, the budget funds these efforts at a more adequate level.

The DACMPT continues to review the long-term security and validity of the ASVAB tests and is encouraged by the fact that a number of ongoing projects are exploring ways to develop new items on a regular basis and tests of new constructs. In the following paragraphs, we describe what we consider to be significant achievements followed by areas that need continued special attention. Continued research in these areas is vital to maintaining and increasing the value of an enlistment and testing program, such as the ASVAB.

The most significant achievements of the past two years include:

1. The computerized-adaptive test (CAT)-ASVAB is performing well with new item development and piloting (i.e., seeding) strategies underway. This is important for the quality and security of ongoing ASVAB administrations. Several other new developments are described in the summaries of briefings on these topics.
2. The launch of the Internet CAT-ASVAB (iCAT) platform in the Career Exploration Program (CEP) will reduce the need for paper-and-pencil testing sessions. Growing awareness and use of the CEP provides an important source of military recruiting leads. This growth recognizes the increasing national interest in the importance of career exploration and decision making.
3. The automated (computer) item generation program is now available for operational use. The program has been used to develop approximately 5000 items for the new Assembling Objects test and approximately 3000 items for the Word Knowledge test. Similar methodologies are being explored for other tests.
4. Participation in the Pending iCAT (PiCAT), the unproctored version of the iCAT, is growing. Upon completion of this test, applicants are asked to take a Verification Test (VTest) and, depending on the outcome of the VTest, may be asked to take the full length ASVAB. Careful evaluation of the results of the VTest and the time taken by recruits in various aspects of this procedure suggests that it can shorten testing times significantly and provide quality evaluations of applicants.

5. Research continues regarding the usefulness of the Tailored Adaptive Personality Assessment System (TAPAS), a CAT version assessment of personality. The Army has spearheaded research on this instrument, and other Services have provided encouraging evidence of its validity and incremental validity over the Armed Forces Qualifying Test (AFQT). This report suggests continued research on several issues related to the TAPAS.
6. To support the high-quality, technical rigor, and ongoing success of the ASVAB testing programs, two efforts have made substantial progress in the past two years. A psychometric checklist has been refined, providing a consistent and coherent basis on which all tests will be evaluated. In addition, the ASVAB philosophy and interpretative argument have undergone substantial development, providing a coherent basis for validity and validation of the ASVAB and associated tests.

Several areas of concern, or areas in which development of procedures and assessments must continue, include:

1. Measures of personality characteristics are of interest to the DACMPT because they add substantially new information to the personnel testing programs, measuring something quite different than aspects of intelligence and aptitudes. The TAPAS is an effort currently being used and evaluated across military Services. Some validation efforts have produced interesting results with some uncertainty because of limitations in study design and inability to control relevant features of test administration and personnel decisions (e.g., lack of true randomized control trials). The DACMPT encourages more comprehensive documentation of the technical qualities of TAPAS design, scaling, and scoring. It will be useful to estimate TAPAS validities with respect to other commonly used performance measures across Services.
2. Although not extensively addressed in this report, interest inventories are being developed and evaluated by several military Services, but progress has been slow. Part of this is due to prior resource constraints, so we look forward to further development in the future given the successful funding of the program budget. The DACMPT agrees that interest inventories may be good predictors of personnel satisfaction and retention and could be valuable adjuncts to the current ASVAB measures. It may also be useful to determine if a general interest inventory (as opposed to different inventories across Services) might be achievable.
3. The Assembling Objects and Mental Counters tests, which address specific aspects of cognitive ability, continue to be of concern, and the DACMPT acknowledges that work on these measures continues.
4. The DACMPT continues to be concerned with the progress being made in adopting procedures for changing ASVAB content. However, recent progress on developing a consistent argument-based approach to validation will move the program in the right direction. Validation efforts will be supported by having good and consistent measurement of outcome variables (e.g., performance, satisfaction, retention, reenlistment).

Significant progress and continued development in all areas of achievement and areas of concern are dependent on continued full funding of these efforts. Many of these projects have made significant progress given the recent increases in funding. The continued success of the ASVAB and the security of its item pools depends on a constant replenishment of item pools and consideration of new and changing constructs demanded by changes in the skills required of military personnel.

Introduction

The Defense Advisory Committee on Military Personnel Testing (DACMPT) was established in 1981 following the miscalibration of the Department of Defense (DoD) Joint-Service enlistment test, the Armed Services Vocational Aptitude Battery (ASVAB) Forms 5, 6, and 7, which were in use from January 1976 through September 1980. As a result of the miscalibration, scores of many low-ability applicants were inflated, and a large number of these individuals were erroneously enlisted. DoD advised Congress of this problem in February 1980 in the Defense Manpower Overview Statement and then provided Congress with more detail in a July 1980 report entitled "Aptitude Testing of Recruits." In the Conference Report that accompanied the Fiscal Year 1981 DoD Authorization Act, Congress indicated it was seriously disturbed by the ASVAB calibration problem. To ensure that similar problems did not recur, Congress requested that the Secretary of Defense establish an independent review board composed of professionals in the field of psychological and educational testing.

In April 1981, pursuant to P.L. 92-463, Federal Advisory Committee Act, DoD established the DACMPT. This Committee has three primary objectives: (a) to review procedures used in the development, calibration, and validation of the current DoD selection and classification tests; (b) to review procedures used in the development of future DoD tests; and (c) to provide recommendations regarding applications of state-of-the-art testing technology to DoD personnel testing programs. The Committee reports its findings and recommendations in a Biennial Report, submitted through the Under Secretary of Defense for Personnel and Readiness, to the Secretary of Defense. This is the eighteenth report of the Committee. The first DACMPT report was submitted in June 1983.

This report contains summaries of the major topics reviewed by the Committee during the previous two years (from July 2017 through March 2019), along with the Committee's findings and recommendations. The recommendations summarized here are from those previous meetings, and many have been acted on already. The major topics reviewed are in the areas of: (a) operational procedures, such as ASVAB form development, test content and statistical specifications, and the development of a psychometric checklist and validity framework; (b) ASVAB delivery and operations (elimination of paper-and-pencil [P&P] test delivery and the implementation of unproctored pre-screening along with a verification tool); (c) the ASVAB Career Exploration Program (CEP); and (d) enhancements to the testing program, such as the development of specific aptitude tests (e.g., Cyber Test and Assembling Objects [AO] test), the Tailored Adaptive Personality Assessment System (TAPAS), and automated item generation for some tests.

The Committee members who contributed to this report include Dr. Barbara Plake, Emeritus at the University of Nebraska-Lincoln; Dr. Michael Rodriguez (Chair), University of Minnesota; Dr. Neal Schmitt, Emeritus at Michigan State University, and Dr. Kevin Sweeney, College Board. The committee also experienced a shift in staff, as Dr. Jane Arabian retired in March 2018 following a long and successful career with the DoD. This report reflects two meetings

supported by Dr. Arabian and two meetings supported by Dr. Sofiya Velgach, Assistant Director, Accession Policy, as the Committee's Executive Secretary.

On that note, the DACMPT congratulates Dr. Arabian on her retirement and extends sincere appreciation for her professionalism, support, and commitment to the efforts of the DACMPT.

General Overview of the ASVAB Testing Program

The ASVAB Testing Program has been the Joint-Service enlistment test since 1976. The Program has two major components: the DoD Enlistment Testing Program (ETP) and the DoD Student Testing Program (STP), implemented in schools as the ASVAB CEP. Parallel forms of the ASVAB are used in both programs; new test forms have been introduced about every four years. In the ETP, the ASVAB is administered annually to approximately a half million military applicants, depending on recruiting mission requirements. In 2018, over 300,000 proctored enlistment tests were administered, as well as over 284,000 unproctored Pending Internet Computer-Adaptive Test (PiCAT) enlistment tests. Over 700,000 students were tested as part of the CEP in 2018 in over 12,000 schools. The P&P forms of the ASVAB have been phased out from Military Entrance Processing Stations (MEPS), though limited use has been retained in distant Military Entrance Test (MET) sites on an exception basis. Most applicants now take a CAT version of the test. Two P&P forms and one CAT form are available for CEP, and a new CAT version is being developed based on old P&P forms of the test.

The current ASVAB P&P forms (Forms 23 and 24) are used in the ASVAB CEP and one form (Form 27) is used in ETP at MET sites. The Defense Personnel Assessment Center (DPAC), formerly known as the Personnel Testing Division (PTD), is no longer creating new P&P forms. The elimination of P&P test forms is intended to make the testing program more efficient and secure. Beginning in September 1990, and continuing through the present, many applicants have been administered a CAT version of the ASVAB. Initially, there were two forms of this CAT-ASVAB and, over the 1990-1996 time period, they were used in five MEPS as an operational test.

This marked the first time DoD enlisted recruits based on scores from a CAT. In October 1996, DoD began the implementation of the CAT-ASVAB in additional MEPS across the country. CAT-ASVAB implementation was completed in all MEPS by August 1997. There are now five operational CAT-ASVAB item pools, with the current CAT-ASVAB forms in use being 05E through 09E. The same forms, now available in MET sites, are administered via the Internet and referred to as iCAT forms, with another form used for PiCAT, the unproctored iCAT.

Since 1968, the DoD also has sponsored the STP, now known as the CEP. The CEP uses the ASVAB in a P&P format to facilitate large test sessions. Student scores and support materials useful for educational, career guidance, and counseling purposes are provided free of charge to participating schools. Students may also use these ASVAB scores for enlistment purposes (with some exceptions). An iCAT version of this test is administered on a smaller scale.

ASVAB testing for enlistment purposes is a high stakes enterprise, much like major college entrance examinations. The ASVAB is a large-scale, nationally administered, internationally recognized test; the scores are used to determine the enlistment eligibility of individuals. Young adults are either granted or denied the opportunity to enlist in the military based upon their ASVAB scores, and opportunities to enter specific training programs and military occupations depend, in part, on ASVAB scores. Each individual decision is critical. One of the scores derived from the ASVAB, the Armed Forces Qualification Test (AFQT), is of particular significance. This score, reflecting verbal and mathematics skills and abilities, is used not only by each Service as a primary enlistment criterion, but also by DoD and the U.S. Congress as an index of overall military personnel quality.

Within DoD, every program must have an Executive Agent responsible for funding and operating the program. For the enlistment and student programs, the Executive Agent has been the Defense Manpower Data Center's Personnel Testing Division (DMDC-PTD) since 1989. For the CAT-ASVAB program, the Navy was the Executive Agent until October 1, 1994; on that date DMDC acquired responsibility for both programs. In 2017, the DMDC testing function, DPAC, was reorganized and is now part of the Office of People Analytics (OPA) within the Defense Human Resources Activity (DHRA).

Accession Policy & Recruitment Update

Accession Policy and Recruitment Updates occur at each DACMPT meeting. Ms. Stephanie Miller, Director, Accession Policy, typically presents the briefings; in her absence, the Committee's Executive Secretary does so. Over the past two years, a number of significant challenges have been presented, many of which have also been addressed during the same period with promising outlook. The challenges to achieving the Department's goals in recruitment and accessions deal with a decrease in the qualified and propensed market, a decrease in the exposure of the youth population to the military, and national improvements in employment rates.

Budget resources have a direct impact on both accession and recruitment, to include the ASVAB Testing Program's security and sustainability. One indicator of the role and effectiveness of the testing program in place is the selection ratio and its success in identifying and placing successful personnel. This, in part, depends on the broad representativeness of the recruitment pool. The Force of the Future initiatives indicate additional promise to support high quality accession and recruitment outcomes as well as greater support and security for the future of the ASVAB and associated measurement endeavors. The DACMPT acknowledges the importance of recruitment, marketing, and advertising resources to support these efforts. The current efforts to increase these resources are encouraging.

Resource Update

During the period covered by this report, resource updates were brief and occasional. During the previous Biennial Report period, there were significant concerns expressed by the DACMPT

regarding the shortfall of resources, in part due to the federal budget sequestration. Since that period, major aspects of the military personnel testing program have been fully funded and resources have been restored. This has enabled progress on important security and future sustainability aspects of the testing program, including information technology (IT) test delivery support, in-house psychometric quality assurance, and important research studies to support the validity of ASVAB use across Services.

Comments

The DACMPT simply reiterates the importance of fully funding the ASVAB Testing Program. In the past, the DACMPT has encouraged alignment between the ASVAB funding structure and the validity claims that the program makes. In this way, the impact of funding decisions on the validity of the ASVAB will be transparent. Ongoing development and security of all military personnel testing programs will provide for the highest quality program to support selection and placement decisions.

ASVAB Project Milestones and Projects

At each of our meetings during this period, Dr. Mary Pommerich presented updates on major ASVAB research and development efforts, including milestones and project schedules. More detailed briefings on many of these projects were provided in one or more of the meetings as well. The DACMPT has looked forward to hearing about the progress toward moving to the Cloud for delivery of the testing programs and the IT infrastructure and security measures accompanying these advancements. In addition, the DACMPT maintains that item pool development on the ASVAB and AFQT are absolutely critical to maintaining an uncompromised test. The DACMPT is encouraged regarding continuous progress in these efforts.

PiCAT/VTest Updates

The PiCAT assessment, coupled with the verification test (VTest), is an alternative ASVAB assessment system that permits applicants to take the unproctored PiCAT with the opportunity to verify their PiCAT scores through the VTest. The VTest is proctored and allows PiCAT scores to serve as qualifying ASVAB scores. The goal of the PiCAT/VTest approach is to reduce the time needed for proctored ASVAB administration, thereby allowing more time for administering other/specialty tests.

The DACMPT received two presentations on the PiCAT/VTest for this biennial period. The first focused on the number of applicants taking the PiCAT/VTest, and the second addressed possibilities for reducing the number of candidates who fail the VTest.

Number of Applicants Taking the PiCAT/VTest

The first presentation provided an update of the Services' progress on increasing the rate of PiCAT utilization to 75% to reduce proctored testing time. Data from January through October 2017 show that, of 331,601 applicants tested, approximately 50% utilized the PiCAT for

qualification determination. Of that group, 33.5% had their PiCAT scores verified by the VTest and did not have to take the proctored CAT-ASVAB.

Reducing the Number of Candidates Who Fail the VTest

The second presentation explored ways to reduce the number of candidates who fail the VTest by examining: (a) failure rates across failure models and (b) failure rates by Service. Failure models signal if a PiCAT score is not trustworthy due to collaboration, confederate substitution, or pool breach. A follow-up analysis of non-passing VTest performance indicated that most of the failures (i.e., 76% of cases) were due to probable confederate substitution. The use of a confederate to substitute for the applicant could result in performance on the PiCAT that is higher or lower than would have occurred if the applicant had taken the test. The investigation of VTest failure rates across Services showed that the Marine Corps had the highest pass rate, as well as the shortest lag time between PiCAT and VTest administrations. Prior analyses have shown that VTest pass rates decrease as the interval between PiCAT and VTest administrations increases.

Recommendations/Comments

The DACMPT sees value in the PiCAT and agrees with the DPAC recommendation that the ASVAB Testing Program continue to use the current algorithm to determine pass/fail decisions, even for failures due to the confederate model. Services may consider ways of shortening lag time between PiCAT and VTest administrations as a way to increase VTest passing rates.

ASVAB Development

Several presentations on ASVAB development covered: (a) the seeding of new items into test delivery to obtain item pretest data; (b) the development of new item pools, including a plan to equate the new forms to the existing ASVAB scale; and (c) a study to examine whether test takers had sufficient time to complete the ASVAB tests.

Seeding of New Items into Test Delivery

The first presentation addressed a July 2017 DACMPT recommendation to explore the possibility of gathering evidence that the new seeding strategy would not adversely affect item parameter estimates. In the new strategy, randomly selected test takers would receive 15-item sets of seed items in one or more ASVAB subtests. In the prior strategy, all test takers received one seed item in each of the ASVAB subtests. The briefing focused on the random assignment and sampling design but did not acknowledge the lack of direct comparison to the prior seeding strategy.

Development of New Item Pools

Construction of new item pools. The first presentation on item pool development summarized the steps that the Human Resources Research Organization (HumRRO) is taking to implement procedures for constructing new forms/pools for ASVAB. HumRRO provided a synopsis of the

new item seeding strategy in contrast to how item seeding had been implemented previously. Previously all test takers received one pretest/seed item across all of the ASVAB assessments. In the new seeding approach, randomly selected test takers receive multiple items in a subset of ASVAB assessments (for example, 29% of the test takers receive 15 General Science [GS] and 15 Arithmetic Reasoning [AR] items, but no seed items from other assessments). Once these seed/pretest items have been administered and calibrated, their item parameters need to be rescaled to be put onto the common scale. The proposed approach uses the mean and standard deviations from the latent distributions for the group of operational items and for the group of seeded items. The seeded items are then put onto the operational parameter distribution using a linear transformation.

Development of ASVAB pools 11-15. This presentation summarized efforts to identify enemy items (items that cue the answer to other test questions) and, once identified, to distribute them across the ASVAB test pools, thereby ensuring that they would not appear in a test administered to an applicant. The presentation also described the CAT simulations that support test form assembly. The presentation concluded by identifying tasks to be completed in preparation for final pool construction to support form administration. Because of the large number of items in the development process, due in part to the aggressive item seeding/tryout schedule, a methodology is needed to identify items that cue others and, therefore, should not be delivered to a test taker in a single administration. Two strategies are being deployed to address this concern: (a) a more automated method for identifying potential enemy items and (b) distribution of these enemy items across item pools so they are not available for item selection during the use of a single pool. Two strategies were tried out for automatic item enemy flagging (content-based judgments and a "hotspot" tool). Simulations were used to explore the information functions derived from applying the item selection algorithms across the newly developed pools using differing numbers of items per CAT administration and across two or three pools. The results suggest that for the simulated forms, some subtests (notably Mathematical Knowledge [MK]) include more difficult items than were previously administered using the P&P tests. Mitigation of this result is expected from infusing middle range difficulty items and additional newly developed items. Equating efforts are underway for pool 10 and are in preparation for new pools 11-15.

Equating new forms onto the ASVAB scale. The third presentation summarized the steps planned to equate the new forms to the existing ASVAB scale. This equating plan is congruent with previous equating processes and, therefore, should provide acceptable results.

Time Sufficiency for ASVAB Test Completion

Another presentation summarized the results of a study that investigated whether there is sufficient time for examinees to complete the tests that are on the ASVAB platform by looking at actual examinee response time distributions. In this study, data from 2015 – 2018 across WinCAT (a Windows-based CAT ASVAB) and *i*CAT administrations were used. It should be noted that, starting in 2015, an aggressive item seeding/pretest strategy was implemented on the WinCAT platform, which resulted in additional items (and additional testing time) for test takers. The study also anticipated the implications of using alternate devices in future ASVAB administrations on test administration time. After inspecting the actual test score distributions for

candidates who completed 95% and 99% percent of the items, two subtests appeared to demonstrate speededness with the 95% completion criterion (MK and AO).

Recommendations/Comments

Regarding seeding strategies, the DACMPT acknowledged that, at this point, it is not practical to implement a sample based on the previous seeding strategy for comparison purposes. A second July 2017 DACMPT recommendation addressed the possibility that the new seeding strategy might alter the test specifications in a way that affects item parameter estimation. This reflects a general concern about potential context effects introduced by seeding 15-item sets in selected subtests per person. The response offered an experimental study employing items from Form 4 (used for linking and equating), in addition to new seeded items with parameter estimates but not used operationally, to evaluate the possible parameter drift. However, the DACMPT noted that this still does not estimate the potential context effect regarding coverage of subtest content and the shift in test specifications.

The DACMPT acknowledges that the development of new item pools is progressing well and encourages the DPAC to continue monitoring completion time sufficiency. The DACMPT also concurs with the proposed time adjustments to ensure that at least 99% of examinees are able to complete the test in the testing time allotment.

Word Knowledge Automated Item Generation

Two briefings on the ASVAB Word Knowledge (WK) Automated Item Generation (AIG) Project were presented to the DACMPT by Dr. Isaac Bejar (Educational Testing Service). The first briefing was presented at the July 2017 meeting and the second briefing was provided at the September 2018 meeting.

The September 2018 meeting was the final briefing on this project. The project yielded over 3000 WK items. The WK generator has been fully developed and is now ready for operational testing in the DPAC system.

Recommendations/Comments

The DACMPT is encouraged with the pursuit of AIG methodologies to address ongoing item development. The success in developing a WK generator is promising, and similar methods should be explored for other content areas. As the WK generator is implemented in operational systems, DPAC staff will provide continued important review for its functioning and ongoing success.

APT (AFQT Predictor Test)

The APT is intended to predict performance on the AFQT, providing for a brief screening of candidates early in the recruitment process. This test is a short adaptive assessment that consists of 4 AR, 8 WK, 3 Paragraph Comprehension (PC), and 5 Math Knowledge (MK) items, which is

expected to take 10-20 minutes to complete. The initial estimated AFQT performance was based on a simulation approach. Data for this study were based on 1750 applicants, who took both the APT and the CAT-ASVAB in May 2017. In addition, there were 939 applicants who took both the APT and PiCAT. The purpose of this study was to update the prediction equations using these operational data sets. The results indicated that the prediction equations based on the operational data better fit the data than did the simulated results, especially for the proctored CAT-ASVAB with the APT. The fit was not as good for the application of the results for PiCAT – APT.

Recommendations/Comments

The DACMPT supports the continued use and exploration of the APT as a brief predictive screener. The DACMPT also provided a number of specific recommendations regarding further exploration of APT effectiveness, including a cross-validation of the current prediction weights.

ASVAB Psychometric Checklist

Over the past several years, the various parties involved in developing and using new tests have developed a set of criteria that they hope to use in evaluating new instruments. A briefing on the Psychometric Checklist was provided by Dr. Greg Manley from DHRA/DPAC in July 2017. Dr. Manley explained that the purpose of the psychometric checklist was to identify a fixed set of criteria for evaluating the merits of tests under consideration for addition to the ASVAB platform. In practice the checklist is used to assist in evaluation, rather than to make definitive go/no-go decisions. The checklist also seeks to establish ASVAB standards for each of the major sections and specialty tests currently associated with entrance testing. Dr. Manley identified the major sections of the checklist as being Theoretical Development, Measurement Precision, Validity, Equating, Fairness & Sensitivity, and Operational Considerations. A copy of the checklist was provided to the DACMPT as part of the briefing materials.

Recommendations/Comments

The DACMPT was pleased to receive this briefing and believes that the usage of the checklist represents exemplary practice and urges the continuation of its use. The checklist and its purpose were both clear and comprehensive. However, the DACMPT did note that the checklist does not address the issue of technical documentation associated with the test being proposed. The checklist could be improved by adding a major section addressing documentation or adding a documentation subsection to each major section.

Tailored Adaptive Personality Assessment System - TAPAS

The TAPAS is a forced choice personality inventory developed by the Drasgow Consulting Group. The DACMPT has received three briefings on the TAPAS during the past two years. The first briefing consisted of reports from the Army, Air Force, and Marine Corps on their research and experience with the TAPAS. Dr. Chris Nye from the Drasgow Consulting Group and Dr. Heather Wolters from the Army Research Institute presented the update on TAPAS research in the Army. This report was extensive and presented the results of several research efforts

investigating the reliability and validity of TAPAS as it is used by the Army. The presenters also suggested some future research initiatives. Specifically, the results of three studies examining TAPAS reliability were presented; three research efforts to improve the TAPAS (i.e., recalibrating the TAPAS statement pool, developing Smart-CAT, and exploring triplet forced-choice items) were discussed, and five sets of analyses examining different aspects of TAPAS predictive validity were presented. Developing Smart-CAT and exploring triplet forced-choice items were identified as work to be conducted in the future.

Regarding the three reliability studies, we noted that the test-retest reliability is best thought of as parallel forms test-retest, because examinees receive different questions at each testing time. Additionally, we noted that each study had different TAPAS dimensions that contributed to the "Can Do," "Will Do," and "Adaptation" composites. Because these composites were not constructed the same way, the ability to compare results across studies is limited. The DACMPT was told that security concerns prohibited the sharing of which dimensions contributed to each composite at the meeting, as it was a public meeting. In comparing Study 1 to Study 2 in their presentation, briefers asserted that Study 2 produced higher reliabilities than Study 1 for the composites. The DACMPT pointed out that, although on average this was true, it was actually only the case that 2 of the 3 composites had higher reliabilities. Because 33% of the composites had lower reliability in Study 2, this assertion overstated the results. For Study 3, the DACMPT commented that the population obtained from Amazons' Mechanical Turk was significantly different than the populations in Studies 1 and 2 and was also different from the population to which the results would be generalized. These differences are important and should be acknowledged as study limitations.

Dr. Eric Charles summarized a Marine Corps study designed to predict boot camp attrition for enlisted personnel based on performance on the TAPAS (there are also plans to do a similar study with officer candidates). The analysis is based on Loess Lines (which are useful depictions of curvilinear relationships). Visual inspection of these Loess Lines was the basis for constructing a "handmade" bootstrap predictor. This approach appears to be a promising effort to identify those Marines who are likely to leave the Service early. This approach to examining the TAPAS's ability to identify attrition is worthy of continuing research.

Dr. Mark Rose of the Air Force presented the results of research on the TAPAS that has been conducted by Air Force personnel. This research addressed issues related to the fakability of the TAPAS, its test-retest reliability, and its validity. Using several different research designs, there was little evidence that faking was possible, or that it occurred, though substantial faking was evident in response to a social desirability measure that employed the usual rating-scale format. Test-retest reliabilities were low in comparison to meta-analytic reports of test-retest reliability for similar measures, though it was noted again that one reason for this difference might be that most of the literature on test-retest reliability likely included the administration of the same items over time. In the Air Force data, items from the same item pool (but not the exact same items) were administered on the two occasions. Validities computed against a self-report situational judgment test of ethical decision making were of modest magnitude and most were statistically significant. Notably the validity of the composites (Can Do, Will Do, and Persistence) were lower than the Big Five composites computed for the purpose of this study.

Recommendations/Comments

The DACMPT thought the information presented about the TAPAS was useful and, in many ways, comprehensive. However, due to study limitations the DACMPT urged the Army not to over-interpret the results presented.

Regarding the research efforts to improve TAPAS, the DACMPT had few comments and thought all three areas discussed were worthy of future or continued work. The areas discussed were:

1. Recalibrating the TAPAS statement pool to better control the range of parameter estimates.
2. Developing a Smart-CAT algorithm to improve marginal reliabilities by dynamically adding items to dimensions with the largest standard errors.
3. Exploring three-statement forced choice item formats (as opposed to the usual pairs-based format), to increase the efficiency of the assessment.

The DACMPT did indicate that, because changes to the TAPAS instrument based on these efforts may impact the technical characteristics of the assessment, the current TAPAS technical characteristics should be documented as well as possible, ideally through a technical manual. If changes are made, the documentation should be updated to reflect them.

Regarding the predictive validity analyses discussed, the DACMPT agrees that the TAPAS, on the whole, holds promise to provide predictive value in the areas presented, particularly given its reliabilities. It was noted however, that although it provides incremental predictive power beyond the AFQT, the overall predictive validity of the TAPAS is modest, rarely accounting for more than 2% of the variance even when scales are used in combination. Consequently, the DACMPT is concerned about overstating the predictive power of the TAPAS, particularly when describing the assessment to non-technical audiences.

Discussion of the results presented by Dr. Mark Rose of the Air Force led to several DACMPT recommendations.

1. Provide more technical information about the construction of the composites to aid the DACMPT in assessing the technical merits of TAPAS.
2. When presenting research results, the limitations of the studies should be specifically acknowledged.
3. The current technical documentation should be assembled into a comprehensive technical manual that is updated as the TAPAS assessment evolves.
4. The Army should use caution in describing the predictive power of the TAPAS, particularly to lay audiences.
5. If possible, studies from the literature in which test-retest reliability is estimated as a function of administration of alternate forms across time (or two different administrations of a CAT) should be examined. Data from those studies would be a more appropriate comparison to the test-retest reliabilities of the TAPAS computed in this study.
6. The validities of the TAPAS should be supplemented by validities computed against appropriate actual performance measures in the Air Force studies.

A report by the RAND Corporation, critical of the research and potential contribution of the TAPAS measures developed by Dr. Drasgow and his colleagues that have been evaluated and used in some form by all Services, stimulated the external review request by DPAC. A panel of five distinguished and knowledgeable researchers are evaluating research conducted on the TAPAS, making recommendations for future research, and commenting on the readiness of the TAPAS for operational use. The DACMPT was briefed on the progress of this committee at its last two meetings. The DACMPT agrees that the panel, with research support from HumRRO, is competent to provide this evaluation and the DACMPT looks forward to the report. Examination of the TAPAS's incremental validity, test-retest reliability, and its susceptibility to faking as well as other possible problems should inform future use of the TAPAS and the required research support.

ASVAB Validity and Criterion Measures

Over the last two years, the DACMPT has received several briefings regarding validity issues related to the ASVAB and other tests being developed. They have included briefings on a conceptual framework for discussions of validity, discussion of a checklist of items that should be considered when evaluating a test, and the appropriateness and feasibility of criterion measures used in gathering validity evidence.

ASVAB Validity Framework

Dr. Art Thacker of HumRRO updated the DACMPT on progress HumRRO has made in the articulation of a validity argument framework for the ASVAB for both subtests that comprise the AFQT and the other tests that comprise the ASVAB platform. Dr. Thacker started with identification of the main uses of ASVAB scores: (a) admission into military Services and (b) placement into training programs or advanced educational opportunities. Dr. Thacker then clarified how a "theory of action" is related to the interpretative argument, which then leads to gathering evidence to support this interpretative argument. Through several examples of draft theories of action, more clarity was provided to the steps needed to complete a validity argument. Dr. Thacker then provided several claims that could be investigated using a validity argument framework, including that the AFQT measures "G" or general cognitive ability. "G" is broadly predictive of performance. Candidates categorized based on the AFQT are sorted according to their likelihood of success in military occupations. Claims about the utility of AFQT scores for various occupations are evaluated based on performance data and associated outcomes. Each of these claims are related to sources of evidence that support the validity of these claims.

Criterion Measures

Dr. Cristina Kirkendall of the Army Research Institute provided a briefing on plans to examine the criteria used by the Services and how they may be relevant or appropriate measures by which to judge soldier outcomes. The goals of the project are to gather information on the outcomes measured by the Services and the manner in which they are measured. The ultimate goal would be to attempt some standardization of outcome measures across Services to allow for joint validity studies. A second goal would be to identify relevant outcomes that might not be assessed now, as well as to assess the psychometric properties of the outcomes now assessed. The

proposed approach to this issue was to identify two proximal and distal outcomes for each cognitive, personality, and interest predictor, and move forward with evaluation of the quality and functioning of these criterion measures.

Recommendations/Comments

In the discussion around the draft theories of action, the DACMPT suggested that the final outcome in the theory should be "readiness for service," not just success in jobs/training. This work represents a promising effort to clarifying the philosophy underlying the ASVAB and establishing a consistent and coherent validity framework.

Regarding criterion measures, the DACMPT strongly supports the attempt to document what outcomes are being measured and how and whether there are opportunities to develop common criteria across Services. However, the DACMPT does not believe one should begin with predictors and try to develop criteria around them. Logically this approach appears to be backward. One should begin with an assessment of what the relevant outcomes are, if they are now assessed and how. The next step should be to determine commonalities across Services and identify any gaps in the criterion space. Once relevant criteria (e.g., training and job performance, retention, reenlistment, satisfaction) are identified, then one would identify predictors that would allow inferences about soldier outcomes. If this logic is followed, the validity framework discussed in the earlier briefing described above would be more informative.

Norming Needs

The DACMPT received one briefing where the topic of updating the ASVAB norming was considered in January 2018. Dr. Richard Riemer (DPAC) presented the briefing regarding the need for an update to the Profile of American Youth (PAY) 97 study that provided normative data on the qualifications and nature of the youth population from whom the military recruits personnel. It has been 20 years since the last effort to collect such normative data. Normative data is important to support current decision making relative to the population of eligible candidates for military service.

The briefing provided convincing evidence regarding the minimal changes in the population of eligible youth to support the current norms. Population data were reviewed including demographic profiles across age groups, gender, racial/ethnic categories, as well as academic achievement performance based on the National Assessment of Educational Progress (NAEP) by youth group in Mathematics and Reading. The DACMPT acknowledged the appropriateness of these comparisons over time. The briefing and subsequent discussions focused on the extent to which changes in the population characteristics could have a meaningful effect on AFQT estimates and decision-making rules.

Recommendations/Comments

The DACMPT supports the recommendations made by DPAC staff. First, DPAC should repeat NAEP ability and demographic trend analysis every four years to continue to monitor shifts in

national academic achievement performance, particularly in the area of mathematics. Second, DPAC should continue to monitor validity data generated by the Services over time. In addition, the DACMPT recommends ongoing monitoring trends in AFQT scores and key outcomes such as attrition over time, given ASVAB performance and educational attainment. At this time, there is no need to pursue renorming.

Cyber Tests (Information and Communication Technology Literacy)

The DACMPT has received periodic updates regarding the progress of Cyber Test development and use. Dr. Gao provided a briefing in March 2019. Some Services have expressed interest in the Cyber Test, as it provides focused information about knowledge, abilities, and skills in areas of growing personnel needs. The current effort is focused on creating a computerized adaptive version of the Cyber Test. In order to achieve this goal, analyses were needed to (a) identify whether the assessment is sufficiently unidimensional to support the item response theory model underlying the CAT and (b) evaluate how well the CAT performs under simulated conditions.

The unidimensionality analysis indicated that the assessment is sufficiently unidimensional to support its use for a CAT. An analysis of the item pool indicated there was an appropriate distribution of items in the pool in terms of content distribution and item parameter similarity to previously administered 29-item fixed forms. Initial results are promising.

Recommendations/Comments

The DACMPT supported the recommendation of the Manpower Accession Policy Working Group (MAPWG) to transition at a future date to a Cyber Test delivered as a 15-item CAT. Sufficient evidence supports this move with advice regarding the CAT model and the number of pools that results in a good balance of score precision and the availability of a reserve pool, if needed. Bringing additional items – that are either targeted at the cut score or more discriminating at low-to-moderate difficulty – into the pools will further improve score precision.

Assembling Objects

Work is underway to separate Connections (AC) and Puzzles (AP) as two subscales of the AO test. Previous briefings on AO suggested that this scale was multidimensional and that the AC and AP items needed to be scaled separately. In order to accomplish this, separate calibrations were conducted using current item pools. Based on preliminary analyses, it appears that AC can be successfully scaled using the current item pool, but there were problems with convergence of the solution for the AP items. The problem appears to be due to lack of higher difficulty AP items. To solve this problem, special seeding of new items will be conducted by using the 5000 new AC and AP items generated by an external contractor. These pools are more difficult than the current item pools.

Recommendations/Comments

The DACMPT supports the efforts to improve the qualities of the AP measure and pursue the separate scoring and use of AC and AP. If item tryouts reveal AC and AP items have a sufficiently broad range of difficulty, then it will be possible to establish new scales for AC and AP.

DPAC Device Evaluation for ASVAB

This research effort was designed to examine whether it would be acceptable from a measurement and logistic perspective to expand the types of devices on which the ASVAB can be delivered. Currently the CAT ASVAB is delivered using a laptop computer. Other testing programs have considered and sometimes expanded the devices for their test delivery. DPAC plans to engage in a staged device delivery study, considering first those subtests that either are highly unlikely to be impacted by the use of a variety of devices (WK) and those that may possibly have some interaction with performance, such as AO and PC. Devices to be considered in the experimental design include tablets, notebooks, and phones. Test-taker familiarity with these devices will also be gathered in the study. Following a review of the literature on the impact of test scores when administration is completed on mobile devices, a decision was made to limit the study to the consideration of specific notebook, tablet, and smart phone devices. Because these devices are products of different vendors and use multiple operating systems, there are limitations to the generalizability of the results beyond these specific devices. Administration is currently on-going and will employ both recruits and applicants over a 6-month period involving several ASBAB subtests (GS, AR, WK, PC, MK, Mental Counters [MCt], and AO). Data analysis plans were presented with the intent to use multivariate analysis of variance (MANOVA) with equated subtest score and response time as the two dependent variables, as well as a separate analysis to address the impact of device familiarity on examinee performance. Analyses will also consider if there are item features (such as the inclusion of a graphic) that interact with examinee performance for the different devices.

Recommendations/Comments

The DACMPT is eager to see the results of the study. Consideration should be given to doing the full model MANOVA (including interaction terms) instead of looking first at the main effect of the model. Although others have found device effects, they appear to be test and context specific. The potential of using a number of device options will provide greater flexibility to the ASVAB Testing Program.

CAT Equating Procedures

Briefings on equating plans were provided to the DACMPT in January and September 2018. Dr. Matt Trippe, HumRRO, gave updates on the ASVAB psychometric support activities contracted with HumRRO. In addition to a short recap of previous briefings, Dr. Trippe summarized the steps planned and executed to equate the new forms to the existing ASVAB scale. The equating approach is consistent with previous equating processes.

In September 2018, Dr. Trippe summarized the results from the CAT-ASVAB Form 10 equating study, a form that was based on P&P forms intended to be used in the CEP iCAT. He provided evidence of a quality equating to the current score scale. The DACMPT particularly noted the high level in equating quality for the AFQT – a core component of the testing program.

Recommendations/Comments

The DACMPT acknowledges the importance of equating – it is a core task of all continuous testing programs with multiple forms or item pools. The approach taken for the ASVAB forms is consistent with industry practice and should bring forms onto the ASVAB base scale.

ASVAB Career Exploration Program

Dr. Shannon Salyer (DPAC) provided briefings regarding the ASVAB CEP at each meeting. She has also shared information, brochures, and promotional materials developed to increase awareness and adoption of the ASVAB CEP. The DACMPT continues to be impressed with the thoughtfulness and comprehensiveness of the strategies employed to build awareness and utilization of the ASVAB CEP. The numbers of schools, educators, and students engaged in the ASVAB CEP continues to grow.

A number of efforts are noteworthy. The extensive work involved in building and deploying the online tools at the ASVAB CEP website are impressive. The DACMPT received a briefing on the results of an Expert Panel Review of the program. The collaboration with CAVEON (test security consulting firm) is an important one to maintain security and monitor online presence of potentially compromised materials. Finally, the iCAT administration of the ASVAB CEP test is a promising advancement, reducing the need to maintain P&P form testing sessions that are difficult to staff.

Recommendations/Comments

The DACMPT appreciates the updates on the ASVAB CEP, as it plays an important role in early identification of military enlistees and a much broader role in supporting the career exploration needs of youth across the country. The ASVAB CEP staff are encouraged to continue monitoring how states and other organizations use the CEP components to continue to learn how to maximize awareness and use. As iCAT administration expands, the ASVAB CEP program is encouraged to continue to seek ways to increase the availability of qualified personnel for test administration/monitoring; one possibility is the group of individuals trained in each school to administer statewide testing programs (typically called district assessment coordinators). The ASVAB philosophy and validity argument, currently being developed and reviewed, should include the ASVAB CEP. Finally, it is important for the program to continue to articulate inappropriate uses of the ASVAB CEP components, tools, and resulting scores – for example, to avoid setting cut-scores for high school graduation requirements (as some states try to establish career planning and goals to meet federal school accountability requirements).

Adverse Impact

Over the past two years, the DACMPT received two briefings on the adverse impact of various tests. The first briefing contained analyses of the ASVAB tests, and the second briefing provided a report on three new special tests.

Dr. Greg Manley (DHRA/DPAC) provided a report on the potential adverse impact of the use of the AFQT and the remaining ASVAB subtests in composite form as well as an analysis of the subtests in the battery. Dr. Manley provided a definition of adverse impact and the manners in which it is determined, as well as a definition of differential prediction. He also provided breakdowns of scores achieved by males and females as well as various racial groups (i.e., non-Hispanic blacks, non-Hispanic whites, non-Hispanic Asians, and Hispanic whites). For AFQT composites, there was some evidence of adverse impact against women and non-Hispanic blacks for the 50-score cutoff, but little evidence of adverse impact at the 31-score level used for Service entry. Asians score lower on highly verbal components of the AFQT and ASVAB suggesting that English may not be the first language of some of these test takers. For the ASVAB subtests, there was adverse impact against female examinees for some subtests, but this was not true for composite scores. Impact has changed minimally in the years that analyses have been conducted, and mean score differences between groups mirror those obtained in studies of major national tests such as the NAEP and SAT. There was no evidence of differential prediction that would suggest under-prediction of minority or female subgroup scores.

In a second briefing, Dr. Manley provided a report on the level of potential adverse impact of three new tests: MCt, Cyber Test, and Coding Speed. As usual, adverse impact was defined as different rates of selection of members of underrepresented groups relative to a majority group; in this case, non-Hispanic white males. Adverse impact is a function of differences in scores across groups as well as the selection ratio (i.e., the proportion of the total group selected). Because these tests, as well as others, are used with different cut scores by the Services, the standardized group differences provided in the report by Dr. Manley represent the potential for adverse impact, particularly when tests are used to select a relatively small proportion of the total group of recruits, as is the case with the Cyber Test. Standardized group differences summarized by Dr. Manley indicated the greatest group differences occurred on the Cyber Test and for African Americans in comparison to non-Hispanic white males. There also was a moderate difference between gender groups favoring men for the Cyber Test. All standardized differences for the three tests were similar to or lower than those exhibited by other ASVAB tests.

Comments

The DACMPT acknowledges the importance of monitoring adverse impact as it provides important information for test development and test score interpretation and use. It is a component of the validity argument regarding fairness. The DACMPT will continue to request occasional updates to provide additional reflection on the potential impact of adverse impact on force readiness. At this point, the DACMPT has not seen evidence to suggest the need for significant shifts in test design or test score interpretation and use.

Abbreviations and Acronyms

AC	Assembling Objects - Connectors
AFQT	Armed Forces Qualification Test
AO	Assembling Objects, an ASVAB Test
AP	Assembling Objects - Puzzles
APT	AFQT Predictor Test
AR	Arithmetic Reasoning, an ASVAB Test
ARI	Army Research Institute for the Behavioral and Social Sciences
ASVAB	Armed Forces Vocational Aptitude Battery
CAT	Computerized-Adaptive Test
CAT-ASVAB	Computerized-Adaptive Testing Version of the ASVAB
CEP	Career Exploration Program
DACMPT	Defense Advisory Committee on Military Personnel Testing
DHRA	Defense Human Resources Activity
DMDC	Defense Manpower Data Center
DPAC	Defense Personnel Assessment Center
DoD	Department of Defense
ETP	Enlistment Testing Program
GS	General Science, an ASVAB Test
HumRRO	Human Resources Research Organization
iCAT	CAT-ASVAB on the Internet
IT	Information Technology
MANOVA	Multivariate Analysis of Variance
MAPWG	Manpower Accession Policy Working Group
MCt	Mental Counters, a Cognitive Test
MEPS	Military Entrance Processing Station
MET	Military Entrance Test (Site)
MK	Math Knowledge, an ASVAB Test
NAEP	National Assessment of Educational Progress
OPA	Office of People Analytics
P&P	Paper-and-Pencil
PAY	Profile of American Youth
PC	Paragraph Comprehension, an ASVAB Test
PiCAT	Pending iCAT, an Unproctored CAT-ASVAB Pre-screening Test
PTD	Personal Testing Division
STP	Student Testing Program
TAPAS	Tailored Adaptive Personality Assessment System
VTest	CAT-ASVAB Verification Test
WinCAT	Windows-based CAT ASVAB Test
WK	Word Knowledge, an ASVAB Test