DEFENSE ADVISORY COMMITTEE ON MILITARY PERSONNEL TESTING

September 17-18, 2020 Meeting

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

Minutes approved for public release.

Michael Rodriguez

December 08, 2020

Dr. Michael Rodriguez, Chair, DACMPT DATE

DEFENSE ADVISORY COMMITTEE
ON
MILITARY PERSONNEL TESTING

September 17-18, 2020

The Fall 2020 meeting of the Defense Advisory Committee on Military Personnel Testing (DACMPT) was held on September 17-18, 2020. The meeting was conducted virtually using the Microsoft® Teams online collaboration tool to accommodate travel restrictions resulting from the 2019 Coronavirus (COVID-19) pandemic. Dr. Sofiya Velgach (Assistant Director, Accession Policy Directorate [AP]) opened the meeting by stating that it was being held under the provisions of the Federal Advisory Committee Act (FACA) of 1972 (5 USC, Appendix, as amended), the government in the Sunshine Act of 1976 (5 USC, 552b, as amended), and 41 CFR 102-3.140 and 102-3.150 and open to the public. She said the meeting agenda was available on the DACMPT website[1]and that public comments would be received at the end of each day's scheduled sessions. Dr. Velgach then noted that the Spring 2020 meeting of the DACMPT had been cancelled due to the COVID-19 pandemic and described how the current presentations were selected to cover critical program updates throughout FY2020.

Dr. Velgach continued her introduction by announcing the recent passing of a DACMPT committee member, Dr. Kevin Sweeney. Dr. Sweeney, she explained, was an invaluable member of the committee, whose advice and insights had contributed greatly to the progress of military personnel testing over the years. She said his presence would be missed and expressed condolences to Dr. Sweeney's family.

Addressing the administrative components of the virtual meeting, Dr. Velgach described the method for taking attendance and informed participants that the meeting would be recorded via the Teams system. She instructed all participants to mute their devices and committee members to click the "raise hand" button when they wanted to speak. She asked all other participants to remain on mute until public comments were solicited at the end of each day.

Dr. Velgach concluded her introductory remarks by thanking the committee members for their participation and the presenters for their support of the committee's activities. She then directed introductions of all participants.

The attendee list is provided in **Tab A** and the agenda in **Tab B**. The chair of the committee has since provided a letter, written by the committee members, summarizing key committee findings. The letter is included in these minutes at **Tab C**.

1. <u>Accession Policy Update</u> (Tab D)

Ms. Stephanie Miller, Director, Accession Policy (AP) Directorate, presented the briefing.

[1] The DACMPT website Meetings page is located at https://dacmpt.com/meetings/.

Ms. Miller began by summarizing the mission of AP, which is to "develop, review, and analyze policies, resources, and plans for Services' enlisted recruiting and officer commissioning programs." She then presented an organizational chart detailing the structure and programs within AP. An additional chart summarized some of the critical items facing the Directorate in the areas of entrance processing, the GI Bills, accessions, testing, officer programs, personnel security, and transgender/gender dysphoria applicants. A table displayed the fiscal year (FY) 2020 recruiting mission by Service (Active Duty, Guard, and Reserve), followed by a table displaying results as of July 2020. For Active Duty recruiting, all Services met goal in terms of accession and quality. All Reserve Component goals have been met except for the Army National Guard and Reserves, which were at 90.66% and 84.28% of mission, respectively.

Ms. Miller then provided a list of Congressional and internal reports requested from her office. These include a report on the Armed Services Vocational Aptitude Battery (ASVAB), which indicated that eligibility based on the ASVAB is critical to readiness, and only 2% of applicants are disqualified solely due to aptitude scores. The military Services have met or exceeded annual recruiting mission while maintaining applicant quality and with limited reliance on individuals scoring in the 10-30th percentile range. A report on assessing English Learners is due in June of 2021 and will address (a) the impact of current testing and accession standards on English Learner applicants; (b) best practices from academia for assessing academic achievement; (c) best practices in teaching English comprehension, particularly to older students; and (d) the feasibility of enacting such practices within the Department of Defense (DoD). An internal report on diversity and inclusion will include a holistic assessment of all current DoD-issued aptitude tests to identify potential barriers for minority members, including test time limits, language, and modality.

When Ms. Miller said that relaxed testing protocols were implemented to accommodate for delays caused by the COVID-19 pandemic in taking the ASVAB Pending internet Computerized Adaptive Test (P*i*CAT) verification tests at Military Entrance Processing Station (MEPS) facilities, a committee member asked about the potential implications for applicants. Ms. Miller explained that applicants who take the P*i*CAT are reminded that their likelihood of performing well on the verification test decreases as more time passes between P*i*CAT completion and verification testing. She added, however, that it is still important to maintain some flexibility in scheduling to accommodate applicant circumstances and MEPS capabilities. Dr. Velgach explained that the verification testing window had been extended from 30 to 45 days. The committee member then asked if the Defense Personnel Assessment Center (DPAC) has been tracking trends in recent verification test scores, to which Dr. Segall replied they plan to do so. The committee member requested that DPAC provide an update on the effect of extending the verification testing window at the next DACMPT meeting.

2. <u>Milestones and Project Schedules</u> – (Tab E)

Dr. Mary Pommerich, Deputy Director, Defense Personnel Assessment Center (DPAC), presented the briefing.

Dr. Pommerich began the presentation with an overview of the projects to be covered in the briefing, including ASVAB development, the Career Exploration Program (CEP), ASVAB and Enlistment Testing Program (ETP) revision, and the Defense Language Aptitude Battery (DLAB).

- New Computer Adaptive Testing (CAT)-ASVAB Item Pools. The objective of this project is to develop CAT-ASVAB item pools 11 – 15 from new items. New form implementation is projected for October 2021.
- Developing New CAT Item Pool for the CEP. The objective of this project is to build a CAT pool from paper-and-pencil (P&P) Forms 20B, 21 A&B, and 22 A&B for implementation of the Internet CAT-ASVAB (*i*CAT) in the CEP. The new pools were implemented in January 2020.

- Automated Generation of Arithmetic Reasoning (AR) and Mathematics Knowledge (MK) items. The objective of this effort is to develop procedures for automating AR and MK item generation so that AR and MK pools can be replaced on a more frequent basis. Completion date was May 2020.
- Automated Generation of General Science (GS) items. The objective of this effort is to develop procedures for automating GS item generation so that GS item pools can be replaced on a frequent basis. The projected completion date is February 2021.
- ASVAB Technical Bulletins. The objective of this project is to develop a series of electronic ASVAB technical bulletins to meet American Psychological Association (APA) standards. The project is ongoing.
- CEP. The objective of this project is to revise/maintain all CEP materials, conduct program evaluation studies, and conduct research studies as needed. The project is ongoing.
- Evaluating New Cognitive Tests.
  - Mental Counters (MCt). The objective of this project is to conduct a validity study to evaluate the benefits of adding MCt to the ASVAB and provide data to establish operational composites that include MCt and operational cut scores for new composites. The Navy is taking the lead. Completion schedule is to be determined (TBD).
  - Cyber Test, formerly the Information/Communications Technology Literacy (ICTL) Test. The goal of this project is to develop and evaluate the Cyber Test. The Air Force is the lead, and the project is ongoing.
  - Nonverbal Reasoning Tests. The objective of this project is to address the ASVAB Expert Panel's recommendation to investigate the use of a test of fluid intelligence, such as nonverbal reasoning, and to plan and conduct construct validation studies. Project completion is TBD.
- Adding Non-Cognitive Measures to Selection and/or Classification. The objective of this project is to address the ASVAB Expert Panel's recommendation to evaluate the use of non-cognitive measures in the military selection and classification process. The measures being evaluated include the Tailored Adaptive Personality Assessment System (TAPAS), and Army, Air Force, Navy, and Marine Corps interest inventories. The project is ongoing.
- DLAB. The objective of this project is to transition to all computer-based testing and improve the predictive validity of the DLAB.
- Expanding Test Availability: Web/Cloud Delivery of Special Tests. The objective of this effort is to transition delivery of special tests from the Windows-based platform to a web-based and/or Cloud platform. The anticipated completion date is May 2021.

There were no questions or comments during the briefing.

3. <u>Device Evaluation Update and Future Use</u> (Tab F)

Drs. Tia Fechter and Dan Segall, Defense Personnel Assessment Center (DPAC), presented the briefing.

Dr. Fechter began by outlining the goals of the study, which include facilitating device expansion of the ASVAB internet Computer Adaptive Test (*i*CAT) and Pending internet CAT (P*i*CAT) by evaluating examinee performance differences among electronic devices such as tablets and smart phones. This will allow for more flexibility for ASVAB administration to reduce time spent in Military Entrance Processing Station (MEPS) facilities, increase the number of enlistees, and increase schools' participation in the Career Exploration Program (CEP). The data collected will allow DPAC to make a recommendation for which types of electronic devices should be approved or prohibited for ASVAB administration. It will also inform a Next Generation user interface that incorporates a Responsive Design approach. Dr. Fechter then showed tables that summarized the data collection plan and status. In all, 10,527 recruits and applicants took part in the study at 10 training locations and 17 MEPS. An additional table summarized the sampling plan and actual number of subjects per condition. Experimental group conditions varied by device, web browser, and subtests taken. Item special features were also taken into consideration, such as whether a graphic was included in the item, whether the item had been reconfigured, the complexity of the graphic, and the inclusion of mathematical symbols.

Dr. Fechter continued by discussing data cleaning procedures, such as removing cases with self-reported low motivation or inability to match records to original ASVAB scores. The final number of cases in the analysis database was 8,517, or 81% of all participants. To investigate whether the device used and familiarity with that device have an impact on ASVAB subtest scores, Analyses of Variance (ANOVAs) were conducted across all device conditions using two models. The dependent variables were subtest scores. The independent variables were device, device familiarity, sample (i.e., applicant, trainee), and administration order. The models were (a) mixed effect linear model with device treated as a random effect, and (b) mixed effects linear model with device treated as a fixed effect. A separate ANOVA was conducted for each subtest. Both models produced consistent results with each subtest, with the exception that the following interactions were significant for model 2, and therefore were used for interpreting additional results:

- Paragraph Comprehension (PC): Device x familiarity and sample x familiarity
- Mechanical Comprehension (MC): Device x familiarity

For Assembling Objects (AO), Arithmetic Reasoning (AR), General Science (GS), Math Knowledge (MK), and Word Knowledge (WK), there were no significant theta score differences between devices, and, therefore, Model 1 was used to interpret additional results. Dr. Fechter then showed tables summarizing the results. Both models produced consistent results within each subtest in regard to response times. For most subtests (except PC), there were significant response time differences between devices and, thus, Model 2 was used to interpret additional results for all subtests except PC. For PC, Model 1 was used to interpret the results. An increase in test-taking time would be offered if a difference of 30 seconds or more existed between device conditions. Each subtest's practical significance for response time is scaled to account for the fewer number of items administered during the evaluation. Dr. Fechter continued by showing a series of tables displaying the results for both main and interaction effects. These suggested that, in general, examinees take less time to respond to items on alternative devices in comparison to a standard PC.

Dr. Fechter then summarized the results as follows.

- The specific device an examinee uses to take the ASVAB does not significantly impact test scores.
- Examinees perform better on the ASVAB when they are familiar with the device they use.
- In general, examinees use less time responding to items on alternative devices in comparison to the XPS.
- Overall, based on these findings, ASVAB subtest scores among applicants should be comparable regardless of device used to take the tests so long as the examinee uses a device that is familiar to him/her *and* the test delivery application is designed to be responsive to a variety of device types.

These results led to the following recommendations:

- Design a test delivery application responsive to a variety of device types for ASVAB administration.
- Allow examinees to choose a device they are familiar with to take the ASVAB.
- Develop a test monitoring plan that tracks operational performance differences (scores and response time) between device types.
- Develop a data collection tool that reports device features (e.g., screen size, browser type and version, device type, etc.) for post-test monitoring and analysis.

- Develop and implement a post-test questionnaire intended to measure barriers to optimal performance.
- Operational implementation decisions must be made prior to moving forward with device expansion.

Dr. Fechter then reviewed the remaining analyses. These include conducting multiple differential item function (DIF) tests to determine if devices affect item difficulty. For items noted for DIF, explore whether item features, such as inclusion of a graphic, explain the differences detected. A comprehensive hierarchical Bayesian-based analysis must also be conducted to account for all variables, demographics, and item/score level differences.

Dr. Segall concluded the presentation by reviewing some of the operational considerations to be considered before implementation of the ASVAB on alternative devices. These include who should take the battery on mobile devices (e.g., applicants testing at home, students in schools), which mobile devices should be used (e.g., examinee-owned, DoD-provided, School-owned), and for what purpose (e.g., unproctored P*i*CAT, proctored test to obtain score of record). Decisions in this regard will depend largely on the complete outcomes of the device evaluation. Other considerations include the potential for test compromise from examinee-owned devices via screenshots, maintenance costs of DoD-owned devices, and score effects associated with testing on unfamiliar devices.

At the briefing's conclusion, Dr. Velgach remarked that the Accession Policy (AP) Directorate is excited about the research, because administering the internet-based Armed Forces Qualification Test (AFQT) Predictor Test (APT) and P*i*CAT on phones and tablets could open up more access to minority groups, who are more likely to have access to phones and tablets than laptops. She then asked for the committee's recommendations on how DPAC should proceed given the current results.

A committee member asked whether test time would be the same across devices or vary by device. S/he asked if the latter option might affect applicants' perceptions of fairness. Dr. Fechter said DPAC was leaning toward recommending a single, standard time because, based on their findings, the control condition produced the longest testing times, and the current time limits are sufficient to accommodate that outcome.

Another committee member noted that device familiarity had a greater impact on minority participants and asserted that DPAC must already have data on that question. Dr. Fechter said they do and one of their next steps is to introduce the demographic data into their Bayesian-based analyses. She said the current study lacked the statistical power and design (i.e., random assignment to familiarity) to make a determination at present. She said that they hoped the Bayesian-based framework will allow them to tease out some of that information. The committee member then asked if the number of significant results exceeded what would be expected due to chance, given the number of tests taken by each participant. Dr. Fechter said the analyses had controlled for the family-wise error rate.

A committee member then questioned whether the study suggests examinees should be allowed to choose devices based on familiarity, suggesting that their choices may not produce the desired familiarity effect. The committee member explained that, when students are given the choice of items to answer, their choice is often suboptimal, especially in the case of constructed response items. Dr. Fechter said DPAC would take a closer look at the recommendation. In response, another committee member suggested that examinees at home would almost necessarily choose a familiar device, but at school or a military site, underrepresented (i.e., minority) groups might not be familiar with the available devices. Dr. Velgach clarified that the question is different in proctored settings, where the option to bring a device is the issue. The committee member countered that allowing examinees to use their own devices during proctored testing might cause a problem with features being differentially available across examinees. S/he said unproctored settings, however, already present this problem, and it is important to impose some restrictions so that device choices do not cause inequities. Dr. Velgach said these are exactly the types of conversations they are having, but that the goal is to expand test accessibility to the widest audience possible, while at the same time accounting for the concerns that have been discussed. Citing test content security concerns, Dr. Segall stated that DPAC's intent during initial phases is to exclude the use of personal devices in proctored sessions. He also said there would not be an array of device choices. At home, however, he said the use of personal devices seems promising, and DPAC should be able to identify and control advantages gained by using unallowable features. All the same, he said unproctored examinees should be informed about disallowed features.

A committee member reemphasized the point that allowing device choice will not necessarily maximize the benefits of familiarity. S/he said evidence from item selection research indicates students may make a selection based on content familiarity, even though the item presents a cognitive task they cannot perform. S/he said one of the biggest issues with equity is the amount of information available when making a choice; that is, limited information in respect to what is optimal reduces the quality of the choice. The result would be test takers choose the device they are familiar with but not necessarily the device most appropriate for the item types. Another committee member then reiterated the need to provide guidance on device specifications and selection. Dr. Segall agreed, stating that DPAC must provide clear advice as to which devices and device specifications would be most appropriate, or optimal.

A committee member questioned the potential to optimize device selection, given that what is optimal may vary across tests and demographic characteristics. Dr. Segall explained that, to the extent that the data show some devices produce higher scores, DPAC might be able to tease that out, but said they will probably provide general advice, such as "do not use a device with which you are unfamiliar." The committee member then noted the small effect sizes for familiarity and commented that it might not make that much difference. Dr. Fechter said that is their major takeaway. She said most effects derived from device familiarity, but even those effect sizes were minimal. Dr. Segall then explained that the current study focused on a select number of devices and was not capable of addressing the issues associated with using an even wider range of device types. He said the plan was to keep monitoring the data and types of available devices to identify issues that may surface in the future.

A committee member stressed the importance of evaluating the impact of decisions, noting that some effects in the study were for non-AFQT tests and should have less of an impact on qualification. Dr. Fechter agreed and said they were unable to present results at the AFQT or Service composite level because the study was designed at the subtest level, where each participant did not respond to items from all subtests. A committee member concluded that the best course of action would probably be to recommend examinees use the most familiar device unless evidence shows a particular test is extremely difficult on a given device. The committee member agreed with the need to continue monitoring the situation for the introduction of new devices, especially as they affect test difficulty.

4. <u>ASVAB CEP Update</u> (Tab G)

Dr. Shannon Salyer, Manager, Career Exploration Center, and Mr. David Davis, U.S. Military Entrance Processing Command (USMEPCOM), presented the briefing.

Dr. Salyer began the briefing by reviewing ASVAB Career Exploration Program (CEP) numbers and metrics. These indicated drops in the number of schools and students participating in the program in the 2020 school year, which can largely be attributed to the fact that in-school testing was suspended on March 13, 2020 due to COVID-19. Although there was a drop in paper-and-pencil (P&P) administrations, the number of ASVAB internet Computer Adaptive Test (*i*CAT) examinees increased (72,299 in school year 18-19 compared to 76,232 in school year 19-20). The number of leads provided to the Services through the program decreased from 468,003 in 2018-2019 to 402,868 in 2019-2020, however the overall number of accessions based on CEP ASVAB scores increased over that same period from 28,614 to 30,755. Dr. Salyer then reviewed usage statistics for both the CEP and Careers in the Military (CITM) websites. There were increases in the number of users of both sites, as well as the number of returning users. The number of inquiries received through both sites increased as well. Dr. Salyer continued by displaying a timeline showing major activities undertaken each year to improve the CEP and the CITM websites and the number of users of the websites. This generally showed increases in usage as program and website initiatives were completed. A series of tables then displayed the past, current, and future status of the program in regard to various capabilities, methods, and metrics.

Dr. Salyer then turned to the CEP's response to COVID-19. Testing and in-person post-test interpretations (PTIs) were suspended temporarily in March 2020. Testing can now be conducted, with local Education Service Specialists (ESS) or test control officers (TCOs) coordinating with schools. ESS also can conduct virtual PTIs and hold virtual office hours. A national virtual PTI was conducted once a month in June, July, and August. An online form allows users to obtain their website access codes. Enhanced classroom activities have also been introduced; these activities include learning objectives and teacher guides and are in line with national standards. A one-page flier was created and distributed to ESS and recruiters outlining ways to market the program and leverage functionality in a virtual learning environment. The CEP is also represented at national conferences. A 4-day virtual conference of the American School Counselor Association yielded 221 leads. The Association for Career and Technical Education virtual conference will be held December 2-5, 2020.

Additional marketing activities include a parent campaign that included placement of ads in relevant publications, emails sent to parents in Indiana and Texas, publishing of a college preparation podcast, emails sent to core groups (e.g., counselors, department chairs), and a countdown calendar sent to Military Entrance Processing Station (MEPS) ESS for distribution.

Dr. Salyer continued by describing an effort to provide continuing education credits on the ASVAB CEP. MEPS ESS and other relevant personnel ability to present for CE will be based on their demonstrated proficiency to teach particular components of the CEP. An application to the National Board for Certified Counselors, Inc. and Affiliates has been submitted.

The next topic addressed by Dr. Salyer, with the assistance of Mr. Davis, was the provision of regional google analytic reports to MEPS ESS and TCOs. These are monthly reports that capture metrics and filter data by student access codes. They reflect data accumulated within the entire region, not just individual school boundaries. Among the performance metrics tracked are total users, new users, average sessions duration, and top jobs viewed.

Dr. Salyer next discussed ways that states can have an impact on the CEP. These include (a) legislative actions that either require the program be made available to students or authorize its use; (b) using program participation as an option to meet graduation requirements, a military and career readiness indicator, or a required or recommended tool for career exploration; and (c) mentioning the program on their state websites without requiring or endorsing its use. She then provided examples of relevant legislation that has been passed in Kentucky, Maryland, and Florida. Dr. Salyer then provided a snapshot of school district reopening plans, including remote learning only, hybrid options, and in-person options.

Dr. Salyer then provided an overview of business modernization activities that have been ongoing. A gap analysis was conducted regarding existing business practices and technology associated with the CEP, including testing inventory software, business administration software, session number assignment, marketing, and communication. Site visits were conducted at MEPS across the country, supplemented by phone calls with key personnel. A report was generated that identified functionalities that should be updated, modernized, or added, along with information on costs and infrastructure requirements. The findings were presented to stakeholders from Accession Policy Directorate (AP), USMEPCOM, and Office of People Analytics (OPA) in June. Next steps include planning for new communications, scheduling, and inventory solutions.

Dr. Salyer's final topic was a new program initiative called UNIFORM. The goal was to develop a web-based application to replace the Occupational Database (ODB). The ODB depends on manual entry of information, which is provided in various forms. As a consequence, there is a backlog of 3 years in terms of getting the information updated. A plan was developed to create a more modernized and robust application to house all Service-provided occupational information and streamline data collection, analysis, and distribution. The updated information would be available to various users, including the CEP and CITM websites. The effort was funded by AP using Forces of the Future Funds. AP played a significant role in bringing various stakeholders together. The application has been developed, and next steps include determining whether it will be an OPA application or enterprise solution, working with other DoD entities on data sharing, and working with AP to update relevant DoD Instructions.

At the end of the briefing, a committee member asked Dr. Salyer to describe legislation that would be pro-CEP. Dr. Salyer said Dr. Velgach could speak to this because of her work with the Defense State Liaison Office. She added, however, that pro-CEP legislation might be something like a requirement to show students all post-secondary options. She said they have learned, however, that if a state legislates something, then it has to be done, and they are still trying to obtain the resources needed to support the legislation that was passed in Texas and Indiana. Dr. Velgach then directed the committee to slide 25, which identified state legislation that specifically references the ASVAB-CEP. She noted that Texas' requirement that all high schools administer the test is challenging because the CEP program lacks the manpower to support that level of expansion, which occurred during a single year. Dr. Velgach mentioned additional state requirements, such as providing student personally identifiable information (PII) back to the state, which is problematic due to DoD privacy regulations. She then said some legislative requirements are more readily supported, such as the provision of information about the CEP and the military in general, which is something Accessions is already working with states through the State Defense Liaison Office. Dr. Velgach then stressed that the ASVAB CEP program is not initiating these efforts, but that they want to help states understand the implications of prospective legislation. The committee member said s/he recognized the challenges, and Dr. Velgach emphasized the importance of communicating the CEP program's capabilities with the intent of ensuring any legislation that passes can be supported.

5. <u>CAT-ASVAB New Forms Update</u> (Tab H)

Dr. Matthew Trippe, Human Resources Research Organization (HumRRO), presented the briefing.

> Dr. Trippe began the presentation by providing an overview of the project objectives. These include developing 5 new forms of the ASVAB to replace current operational forms and those in use for the Pending internet Computer Adaptive Test (P*i*CAT), as well as expanding the items available for form assembly. The new forms include unused or assigned items from forms 5-9 and additional items series. Forms 11-15 were assembled from ten experimental item series or sets of 100 experimental items per subtest. Each experimental form is reviewed for psychometric (e.g., model fit, information) and content quality. Items that survive the review process move on to form assembly, and enemy groups are identified to mitigate local dependence. Given that computer adaptive test (CAT) administration is based on forms from which a potentially unique set of items is administered to each examinee, the forms need to contain items from the full range of content and difficulty and sufficient information/score precision across the full range of ability. The goals of form assembly are (a) for each subtest, assign each item to 1 of 5 forms (11-15); (b) maximize conditional precision levels of each form; (c) constrain conditional precision levels to be comparable across all forms; (d) account for enemy items and distribute them evenly across pools; and (e) account for content taxonomies where applicable (General Science [GS], Assembling Objects [AO]).
>
> Dr. Trippe presented the form assembly simulation findings. For most ASVAB tests (i.e., Automotive Information (AI), Electronics Information (EI), GS, Mechanical Comprehension (MC), Paragraph Comprehension (PC), Shop Information (SI), and Word Knowledge [WK]), information alignment and information levels are comparable to existing forms. For Arithmetic Reasoning (AR), the information levels are generally lower than current operational CAT forms and lower than paper-and-pencil (P&P) information in the middle of the distribution. Average reliability is slightly lower than that of forms 5-9 (.02), but virtually identical to P&P forms. Information levels for MK are not well aligned with existing operational forms or observed applicant ability. This may be addressed in the future by providing targeted guidance to item developers. Average MK reliability is slightly lower than that of forms 5-9 (.01-.03), and slightly higher than that of P&P (.02-.03). Dr. Trippe then showed graphs demonstrating the results for each subtest.
>
> The tasks remaining include equating and equating analyses/evaluation. Equating on form 10 has been completed and the team is ahead of the learning curve for equating forms 11-15. Two P&P forms for use in the CEP will also be developed from form-assembly-eligible items not assigned to a CAT form. Seed items will be processed for the next set of CAT forms under projected yearly form-development rates of eight WK, four GS, AR, PC, and MK, and two EI, AI, SI, MC, and AO forms.

Regarding the quality of the information functions shown on slide 10, which Dr. Trippe said could only be better if the peaks were directly over the theta peak, a committee member suggested the peaks would be optimal if they were in the range where decisions are made. Dr. Trippe agreed. The committee member then asked if the target was optimal. Dr. Trippe said, if decisions were made based on individual tests, he could have drawn a reference line. He said, however, that the individual test scores are only used within composites. The committee member asked for clarification on the theta range and where decisions are critical. Dr. Trippe replied that he did not have a sense of this because decisions are not made based on individual tests. He suggested the possibility of creating a reference line where most of the action is but said he would have to give it more consideration. Dr. Segall said they could ask the Services where their decision cuts are and determine the associated range of GS scores. Dr. Trippe said that was a good suggestion.

Another committee member said most of the information is available at thetas that are quite high, around two standard deviations above the mean, and asked if that was the approximate location of the decision point. S/he noted that other tests have similar differences between the information curve and the distribution thetas, and so when combined in a composite, the tests may always work best for those who are well above the average theta. The first committee member replied that "best" is likely in, but also around, that range. S/he said widening the information curve is helpful. Dr. Trippe said they had tried a few different algorithms and could potentially chop off items at the high end to make the curve look more like the benchmark P&P curve. He said that would allow them to save items for future rounds, but they decided there was little harm in having the additional information, even if it was not used. The committee member concluded, first, that it would be worth mapping to the Service cuts and, second, that retaining the high-end items would cause some loss of efficiency. Another committee member agreed, proposing that most of the useful information is available around 1 standard deviation above the mean. Dr. Trippe said he could address the situation better in a future briefing.

As Dr. Trippe briefed slide 17 and identified an issue with the A parameter, or the new items, a committee member asked about the source of the new item pool. The pool of 1,000 new items, said Dr. Trippe, was written by teachers, or experts, and edited according to established procedures. The committee member characterized the new items as being of lower quality and less discriminating than the items in pools 5-9. S/he asked if the items in the pools were developed using the same procedures and if the difference in quality might be due to different item construction procedures. Dr. Trippe said that was a possibility because the items were developed over a long period of time and not always by the same group or vendor. He said they had conducted distributional analyses by item series, but a smoking gun explanation had not emerged. He reiterated that they have just observed a general depression in the A parameters.

The committee member then asked if the items had been generated through automation (i.e., automated item generation [AIG]). Dr. Trippe said they had not. The committee member said, as an aside, that it would be a mistake to use the new items as a model for AIG. Dr. Trippe agreed and said they would cherry pick any items used for future item developer training. He said they are trying to develop a formal pipeline to provide item statistics back to item developers to improve the quality of distractors, which would make the items more discriminating.

As Dr. Trippe briefed the Math Knowledge (MK) simulation results (slide 19), a committee member commented that there was no issue there.

6. <u>TAPAS Evaluation Project Overview</u> (Tab I)

Dr. Tim McGonigle, Human Resources Research Organization (HumRRO), presented the briefing.

> In 2018, the Defense Personnel Assessment Center (DPAC) contracted with HumRRO to independently review the body of Tailored Adaptive Personality Assessment System (TAPAS) research and make recommendations regarding the readiness of TAPAS for operational use in selection, classification, or other decision making. The evaluators had two objectives: (1) provide recommendations on readiness of TAPAS for operational use/implementation, including policy/issuance directing permanent implementation; and (2) make recommendations on future research and development, to include operational analyses/pilots. HumRRO's method for conducting the review involved assembling a team of nationally known experts in psychometrics, personality theory and measurement, and operational testing. The experts held four meetings between October of 2018 and July of 2019, which were attended by DPAC and Service representatives. Presentations were delivered by TAPAS developers, the RAND Corporation, Service representatives, and other key stakeholders. The resulting report was organized around the *Standards for Educational and Psychological Testing* put out by the American Educational Research Association, the

American Psychological Association, and the National Council on Measurement in Education. The factors that were reviewed included:

- Multi-Unidimensional Pairwise Preference (MUPP) Model
- Scores, Scales, Norms, Score Linking, and Cut Scores
- Reliability
- Validity
- Mitigating Social Desirability
- Fairness and Subgroup Differences
- Test Design and Development and Documentation
- General Recommendations

Each factor was given a rating of satisfactory, minimally sufficient, or insufficient.

Dr. McGonigle continued by explaining that, while the TAPAS Evaluation Project (TEP) team considered evidence from the entire TAPAS system and all listed uses (i.e., selection, classification, and special assignment), based on the documentation received the scope of this evaluation was limited to the operational use of TAPAS for selection decisions. TAPAS was considered as (a) a specific instantiation of a model developed by the Army with other users (Air Force, Navy, Marine Corps); (b) having a library of 27 facets with individual versions consisting of 13 to 15 of these facets; and (c) primarily evaluated for use in selection decisions to date, although other uses have received limited research. The TEP team identified 9 recommendations that should be addressed prior to operational use and 15 that can be addressed after operational use has begun. Assuming no adverse findings result from addressing the pre-operational recommendations, the TEP team believes TAPAS can be used operationally for selection purposes. Post-implementation recommendations will improve the quality of information provided by TAPAS. Accession Policy (AP), DPAC, and Services commented on the report and provided their recommendations/intentions for way forward. Concentration is on the pre-implementation recommendations.

The first consideration was the MUPP model, which was judged to have minimally sufficient evidence in its support. The MUPP model is based on the idea that choices are determined by the degree to which each statement in the pair best represents the respondent. TAPAS is designed to prevent faking by making the "correct" response not readily apparent because the statements are matched on strength of association with a single dimension and on social desirability. The TEP team found sufficient reasoning for the use of the MUPP model, but also identified three pre-implementation and seven post-implementation recommendations. The pre-implementation recommendations were:

- Test the proportionately redistributed probabilities assumptions when respondents agree or disagree with both statements in an item.
- Provide evidence that an unfolding model is an appropriate choice for disagree-agree responses to statements in the TAPAS item pool, including examination of principal components to single-stimulus responses along with generation of empirical item characteristic curves.
- Provide technical documentation about the computerized adaptive testing (CAT) algorithm.

The post-implementation recommendations were:

- Incorporate a population variance-covariance matrix estimate for TAPAS facets into the procedure used to estimate TAPAS theta values.
- Recalibrate the Generalized Graded Unfolding Model (GGUM) item parameter estimates for each facet and the corresponding social desirability estimates using a larger dataset consisting of more recent data and monitor parameters for drift from version to version.
- Investigate the possibility of estimating TAPAS item parameters from the forced-choice responses rather than relying on those developed from responses to single statements using the GGUM.
- Demonstrate algebraically, and document, any dependence of the TAPAS latent metric on unidimensional statement pairings.

- Reanalyze simulation data using a Root Mean Squared Deviation (RMSD) index between estimated and true parameters.
- Document item selection procedures to show that the precision of facet score estimation for core facets is psychometrically sufficient given that the number of statements per facet varies by TAPAS version.
- Calculate and document the item pool information function for each facet of the TAPAS using the information function.

ARI will lead the effort for addressing the pre-implementation recommendations using data from across Services.

The TEP team found insufficient evidence in support of the TAPAS scores, scales, norms, score linking, and cut scores. Versions 9, 10, and 11 are used interchangeably despite variation in facets between versions. Additional variations occur across Services, as does the use of composite scores versus facet scores. There is no single source of information about the scaling and comparability of TAPAS versions. The TEP team identified four pre-implementation recommendations:

- Demonstrate and document version equivalency, develop a comprehensive documentation on norming and equating, including a detailed description of the size and characteristics of the samples used to norm each version.
- Investigate the comparability of facet and composite estimates across versions to determine the impact of changes in construct, by version (movement in and out of facets/statements).
- Document the process by which the cut scores on the facets and composites were derived and the argument and measurement precision near any cut scores that are established to make decisions using TAPAS results.
- Provide clear guidance on score interpretation, including information about score meaning and score precision.

The Air Force Personnel Center (AFPC) will take the lead in developing/collecting required documentation on norming and equating, to include impact on facets and composite scores. If additional analyses are required, data from across the Services will be used. The Services will develop clearly articulated processes for cut score development, score interpretation, and process for score reports. Stakeholders (AP, DPAC, and Services) will develop a centralized standardization process/policy for identifying facets, norming, equating, and version control. The effort for developing a Theory of Action has been initiated.

The TEP team found minimally sufficient evidence to support the reliability of TAPAS. Because TAPAS uses an Item Response Theory (IRT) model for test design, it should be evaluated using indices of measurement precision (e.g., conditional standard errors and marginal reliability) rather than test-retest or alternate forms reliability alone. Non-operational analyses reported good marginal reliability and sufficient test-retest correlations. Operational analyses show lower test-retest correlations, but may confound several sources of error (e.g., test items vary within and across examinees, testing conditions/retest after failure versus not, and inconsistent retest intervals). Preliminary results provide minimally sufficient evidence of measurement precision, but also led to one pre-implementation and two post-implementation recommendations. Pre-implementation, the marginal reliability and conditional standard errors for new item pools should be calculated, along with information about the distribution of precision of estimates across a given sample (when developing new item pools). Post-implementation, an operational study should be conducted to estimate the test-retest correlation of TAPAS facet and composite scores in a context where test administrations can be considered replications and incorporate conditional standard error and marginal reliability indices. Additionally, calculate and document the conditional standard errors of test scores based on Fisher information in addition to those from the replication method. The AFPC will lead the effort in addressing the pre-implementation recommendations, using data from across the Services.

The TEP team found minimally sufficient evidence in support of the validity of TAPAS. TAPAS research has focused on adding prediction beyond ASVAB scores for a number of performance criteria (e.g., turnover, training performance, supervisor ratings) for selection. Evidence shows that TAPAS composite scores contribute small but consistent increases in prediction of attrition. The value of this increment must be determined by the Services themselves. Existing research evidence provides minimally sufficient evidence of incremental validity, but also led to three pre-implementation and four post-implementation recommendations. The pre-implementation recommendations were:

- Develop and document a validity argument for each operational version and use of TAPAS, specifying the outcomes the test is intended to predict, the intended population, and the evidence in support of intended interpretation.
- Conduct operational pilots evaluating impacts on critical Service performance outcomes.
- Provide documentation on construct validity at the composite level.

The post-implementation recommendations were:

- If retests are permitted by policy, evaluate and demonstrate the psychometric properties and validity of the testing system, including policies about retest intervals, number of retests, and length of time for which scores are valid.
- Calculate mean differences in scores by sex and ethnicity when sample sizes are sufficient for separate examination of validity by subgroup.
- Document the extent to which external judgment has been utilized to evaluate the TAPAS item pools for content, construct, and sensitivity reasons.
- Reduce the number of latent constructs to those that are absolutely essential (e.g., those weighted in composites).

The Services will continue to develop and propose additional analyses for validity assessments using the TEP team's recommended study designs. The pilot/analysis design documentation should include critical outcomes, the intended population, evidence in support of intended interpretation, and the impact of any confounding artifacts. Stakeholders will develop a comprehensive centralized technical infrastructure for summarizing all validation efforts to date. All applicable documentation must be included in the summary. The Services will provide existing validity documentation in support of current and proposed inferences to be included in the centralized infrastructure. Efforts will be undertaken to establish a nomological network of the composites to provide additional evidence of construct validity at the composite level. The Services had concerns on the appropriateness of this recommendation, since construct validity is established at the facet level. The Services will conduct validity analyses by sub-group and develop a centralized standardization process/policy for including constructs/facets in future TAPAS versions. A Theory of Action has been initiated, which will drive toward development of a centralized version. Stakeholders will discuss the feasibility of a centralized retest policy.

The TEP team found satisfactory evidence regarding social desirability, fairness, and subgroup differences with respect to TAPAS.

- Social Desirability: Personality measures are notorious for response distortion because candidates attempt to determine what response is expected and present themselves in a way that they believe will make them a more attractive candidate. TAPAS was designed to reduce the effectiveness of this strategy, such that item pairs consist of equally desirable or undesirable traits from different facets. TAPAS uses a design that effectively mitigates social desirability and seems to be impermeable to coaching. The TEP team recommended that an evaluation be conducted to determine the extent to which the social desirability responding parameter is fixed over time.
- Fairness and Subgroup Differences: Personality measures tend to show small to non-existent differences between sex, race, and ethnic groups. TAPAS generally follows this pattern, but the effect size differences for the composites may be worthy of additional study depending on different uses among the Services. Additionally, examination of validity by subgroup will further enhance the fairness evidence.

The TEP team found insufficient evidence concerning the test design, development, and documentation of TAPAS. The standards for test design and development advise that the intended uses of a test are known

> and clearly articulated and guide the test development process. TAPAS was intentionally designed to be flexible to support each Service's needs. As a result, there are aspects of TAPAS test design and development that are better explained than others. TAPAS also suffers from a diffusion of responsibility for the evidentiary arguments for the multiple known, desired, and/or reasonably anticipated uses of the test within and across the Services. The post-implementation recommendation is that the publisher assemble a technical manual that specifies the approved test uses, the evidence for such uses, and the examinee population associated with each use/evidentiary argument. This technical manual should include information on psychometric test design, test development, scoring, reliability, validity, and fairness. The Services will continue assessing potential uses for TAPAS and clearly articulate and document the purpose and objective for each use, with supporting validity information, prior to implementation. The Services must use a MAPWG-established checklist/protocol for providing validity evidence. Accession Policy, in collaboration with DPAC, will initiate development of a centralized process for a documentation repository. The Services will provide all required documentation on psychometric test design, test development, scoring, reliability, validity, and fairness to be included in the repository. The Theory of Action has been initiated, which will drive development of the validity framework similar to the ASVAB validation framework.
>
> Dr. McGonigle continued by presenting some general recommendations that arose from the TEP team deliberations. These include:
>
> - Prior to implementation, develop an infrastructure for permanent operational testing (e.g., define and assign operational roles and responsibilities, create and document quality assurance procedures, develop standardized methods for continual development).
> - Post-implementation, investigate the comparability of samples and research findings for analysis within and across Services and at points in service (pre-accession and post-accession).
>
> In response, stakeholders will develop a centralized process/policy for future TAPAS research, development, and maintenance. The plan should include all applicable roles, responsibilities, and funding streams. Policies for identifying future constructs/facets, norming, equating, version control, and implementation must be covered in the plan. AP, in collaboration with DPAC, will initiate development of a centralized process for a documentation repository. The Services will provide all required documentation on psychometric test design, test development, scoring, reliability, validity, and fairness to be included in the repository.
>
> Dr. McGonigle ended by presenting some overall conclusions from the TEP team's work. The available evidence provides preliminary support for the operational use of TAPAS for selection. However, several items should be addressed, most notably scores/scales, norms, score linking, and cut scores, and test design and development and documentation. As noted, some of these items should be addressed prior to permanent operational implementation, while others could be addressed over time. Assuming no adverse findings result from addressing the pre-operational recommendations, the TEP team believes TAPAS can be used operationally for selection purposes. Post-implementation recommendations will improve the quality of information provided by TAPAS. Next steps include completion of additional analyses, collection of documentation, and completion of a Theory of Action to drive future development of a centralized version of TAPAS.

As Dr. McGonigle briefed the way forward on slide 10, Dr. Velgach asked the committee for its thoughts on developing composite-level construct validity evidence. She said the Services only assess construct validity at the facet level. A committee member asked if the TEP team had made any recommendations on that topic and said it would require determining what the composite actually measures. The committee member then expressed uncertainty about what the TAPAS composites were supposed to measure. Dr. McGonigle said the TEP team had the same uncertainty, and that was the genesis of Dr. Velgach's question. He explained that the situation was somewhat by design due to the empirical basis on which they were developed.

A committee member suggested correlating the facets with the composite and looking at the relative magnitudes of the correlations. Dr. McGonigle then noted that the composites have labels consistent with the criterion variables, but that there is no available evidence suggesting that is what they represent. Another committee member asked how the facets could be combined into composites. S/he said that is not usually dictated by the test but by the test's users, and that s/he believed the TAPAS developers did not recommend specific composite scoring methods, but left that up to the users. Dr. Velgach agreed that the Services use the TAPAS in different ways and for different purposes, and that the Services are responsible for demonstrating the utility of the composites for their intended purposes. Dr. Velgach also mentioned that the Services are concerned with predictive validity rather than construct validity. Dr. McGonigle concurred. The committee member remarked that the strict use of empiricism – saying it is valid because it works – is inappropriate and suggested there are legal requirements to demonstrate the construct relevance of assessments.

Another committee member remarked that the composites do not even work that well in respect to attrition. Dr. Velgach clarified that it is difficult to make gains above the Armed Forces Qualification Test (AFQT), but that the TAPAS's incremental validity over the AFQT is approximately .02. The committee member replied that s/he thought the AFQT predicts performance but not necessarily attrition, and so there should be room for incremental validity in attrition. Dr. Velgach referenced some recent analyses showing that the AFQT predicts attrition as well as performance and said that would be covered in more detail during the next day's AI briefing. The committee member said that research relates to utility vice validity. S/he then said the full TEP team report outlined how to use the Taylor Russell Tables to look at incremental utility and the RAND report contained a similar analysis examining the number of attrites that could be avoided by using the TAPAS plus AFQT scores as opposed to AFQT scores alone. Dr. McGonigle said that was true. He explained that the TEP team stopped with Taylor Russell analyses and that the next step would be to look at the economic impact of potential reductions in attrition. He said the TEP team thought, however, that was a task for the Services to perform. The committee member said s/he thought someone had already looked at the number of people who attrite and calculated the training cost, which would be one measure of utility.

Another committee member questioned the need for construct validity for composites, asking if DPAC performs construct validity work for ASVAB composites. Dr. Velgach said the composite validation for tests like the ASVAB is empirical, and the construct validation is performed at the facet level. The committee member said it seems unusual to do construct validation at the composite level, and Dr. Velgach explained that was the Services' concern as well. Another committee member said s/he thought they were recommending that construct validity be established at the composite level because that is where decisions are made. The first committee member added that, when predicting job performance with a test that measures multiple knowledge, skills, abilities, and other characteristics (KSAOs), construct validity evidence is only required at the KSAO level.

Dr. McGonigle clarified the rationale for the recommendation by highlighting the uncertainty regarding whether composites with the same name (e.g., Will Do) include the same facets across Services. A committee member suggested that one way to make that determination – perhaps not theoretically, but empirically – would be to look at the correlations of the facets against the Will

Do and Can Do composites. Dr. Velgach asked Dr. McGonigle if he thought better documentation of how the composites were developed would help with this question. Dr. McGonigle said he thought it would help. Dr. Velgach then suggested the most important concern is to ensure there is justification for the composites. Dr. McGonigle agreed, noting that a theme across the recommendations is a lack of understanding precisely what the TAPAS measures, and that more documentation would go a long way in that regard.

At the end of the briefing, Mr. Tom Carretta asked if the TEP team had discussed the need to work towards a common core of TAPAS facets that would be used across Services. He mentioned discussions in the Manpower Accession Policy Working Group (MAPWG) about how varying content across forms affects the pairing of facets, which affects scores. Dr. McGonigle said the comparability of scores across versions was of concern to the TEP team and that, outside of the operational context, the TEP team would love to have a single version of the TAPAS. He said, however, the TEP team recognized that a strength of the TAPAS is its flexibility, and so they did not recommend a single, standardized core set of facets. Dr. Velgach commented that the Theory of Action is intended to address the use of a common core of facets DoD-wide as well as the individual Services' needs in developing future versions.
A committee member said s/he understood that the pre-implementation issues would be addressed before the TAPAS is used operationally, but if the Services are using the TAPAS now, should they discontinue its use until those questions are answered. S/he then said a plan of action should indicate what results would be required before moving ahead; that is, is a .01 validity sufficient justification for moving ahead? The committee member said the report requires this question be answered. Dr. Velgach said that was a good point and explained that none of the Services are currently using the test operationally for selection purposes. She added, though, that the Army is using it in a pilot program within a "select-in" framework that allows for the selection of individuals in the 45-49 AFQT range, who are normally not selected at the same level as those with an AFQT score of 50 and above. She also said the Air Force uses the TAPAS for classification, but there was no centralized documentation available for evaluating that use. She added that a current goal is to ensure all required documentation is available to make sure the test is evaluated for these purposes.

Upon hearing that the TAPAS is being used for selection decisions, even if only to a limited extent, the committee member again questioned the appropriateness of using the TAPAS before more questions had been answered. Dr. Velgach responded that, without using the TAPAS in the pilot program, it would not be possible to show any additional incremental validity. The committee member replied that the best way to do that is to avoid using the TAPAS to make decisions, because using it restricts the range of outcomes. At this point, Dr. Tonia Heffner (U.S. Army Research Institute for the Behavioral and Social Sciences [ARI]) explained that the Army *is* running a pilot program, as Dr. Velgach explained earlier, but that no decisions are being made based on TAPAS data and "nobody is being impacted in any way, shape, or form." She said everyone takes the TAPAS, and then the Army determines if those who score in the 45-49 AFQT range perform as well as they would have if they had scored in a higher AFQT range. She said the Army is not making selection decisions but is tracking the data. She also mentioned that the RAND report focused on attrition, although the outcomes of interest in the ARI research are improved motivation and performance. She said the TAPAS does predict the motivational outcomes as well as subsequent performance. She said attrition is the weak spot, it just happens to get all the attention.

Dr. McGonigle concluded the discussion by explaining that the bulk of the available information on TAPAS validity is related to attrition, and that is what drove the TEP team's discussion, though he said they were interested in other outcomes as well. A committee member said that made sense and that s/he remembered data on reenlistment and organizational commitment measures, which were better than the attrition data. Dr. Heffner concurred, explaining that the former data were much better than attrition, with coefficients ranging from 0.2 to 0.4 or higher. She said that when used for job assignment within the Services (i.e., recruiters and drill sergeants), ARI is finding numbers as high as 0.4 and 0.5. She clarified that these results are obtained with the Non-commissioned Officer Special Assignment Battery (NSAB) rather than the TAPAS, and explained that the NSAB has similar dimensions but different questions, because it is against regulations to use the same items that are used in the Military Entrance Processing Stations.

7. Public Comments

After the briefing, Dr. Velgach opened the floor to public comments and asked participants to limit their comments to 5 minutes. There were no additional public comments.

8. Use of Calculators (Tab J)

Dr. Peter Ramsberger, Human Resources Research Organization (HumRRO), presented the briefing.

> Dr. Ramsberger began by noting that there is pressure from the Services and other stakeholders to allow calculator use on the ASVAB math tests, Arithmetic Reasoning (AR) and Mathematics Knowledge (MK). Recruiters think this will result in greater numbers of qualified applicants. Applicants question why calculators are not allowed when they have used them in high school math classes. However, the current policy remains that calculators are not allowed. The primary reasons for this were outlined in a March 2018 policy paper.
>
> - Allowing calculator use would not increase the number of eligible recruits due to the test scaling and equating requirements.
> - Although allowing calculator use would give the appearance of keeping up with the latest trends and better align with college entrance exams, the ASVAB and college entrance exams serve different purposes.
> - The change lacks the technical and psychometric merit necessary to enact it, and the reliability and validity of ASVAB scores would be at risk.
> - The cost of allowing calculators outweighs any possible benefits given the complexity of necessary re-standardization processes.
>
> Dr. Ramsberger continued by noting that HumRRO was asked to examine this issue by taking the following steps:
>
> - Review test anxiety literature for relevant information.
> - Obtain insight into how math is taught in high school and students' preparedness to perform hand calculations required on the ASVAB.

- Evaluate calculator use on other standardized tests (e.g., ACT, SAT, National Assessment of Educational Progress [NAEP]).
- Obtain insight from recruiters and applicants via focus groups.
- Assess the importance of math skills and hand calculations for success in military training and on the job.

The findings will be consolidated into a presentation for multiple audiences (e.g., recruiters, applicants, other stakeholders). Dr. Ramsberger then displayed a series of slides showing the test blueprints for AR and MK.

Regarding test anxiety, the main findings of the literature review were: (a) calculators are widely used in education; (b) racial and ethnic minorities tend to use them less frequently; (c) there is an inverted U-shaped function between calculator use and performance, with examinees at the upper and lower ends of the performance curve using them less; (d) research on the impact of calculator use has yielded mixed results, likely due to factors such as demographics, math ability, and item types; (e) examinees have more favorable attitudes towards math and math tests when calculators are allowed; (f) test anxiety can have a negative impact on performance (one study indicated that 85% of students experience at least some math anxiety); and (g) there is not much evidence to suggest that calculators mitigate the impact of test anxiety on performance, although studies have shown that, for adaptive tests, providing information on how the test functions can lessen anxiety.

In regard to how math is taught, the typical sequence is Algebra, Geometry, and Algebra II, followed by higher-level mathematics. Dr. Ramsberger showed a table displaying the percentage of students enrolled in various math classes in 2009, and the percentage of students whose most advanced high school mathematics credit was earned in each course as of 2013. According to a 2015 survey of 767 2- and 4-year college instructors, less than 50% of students are adequately prepared for college math. Approximately 39% of undergraduates reported taking a remedial math class after high school. Over 99% of schools require or allow calculators, and 90% of schools allow calculator use on tests. Teachers surveyed agreed that calculators should not be used until basic math skills are mastered. Standardized test consortia (e.g., Partnership for Assessment of Readiness for College and Careers [PARCC] and Smarter Balanced) policies dictate that if a student is expected to demonstrate proficiency without the use of a tool, it should not be allowed. The analysis of one of the ASVAB math test item editors suggests that non-calculator Smarter Balanced items align with ASVAB items, although they involve more diverse and language-intense contexts. Dr. Ramsberger then showed sample Algebraic Operations and Equations items from the ASVAB MK and Smarter Balanced tests to illustrate this point.

Dr. Ramsberger continued by explaining that four large-scale standardized tests were reviewed: the NAEP, ACT, SAT, and General Educational Development (GED). ACT allows calculator use on all math items, although they state that the items can be solved without a calculator, and many are best solved without a calculator. Dr. Ramsberger then showed a paragraph from ACT documentation explaining why they decided to allow calculators throughout the test, which highlighted the fact that calculator usage is prevalent in schools. NAEP classifies items as calculator inactive (not required), calculator neutral (not necessary, but allowed), and calculator active (difficult to solve without). NAEP, SAT, and GED include calculator-allowed and non-calculator items, with two-thirds of NAEP items non-calculator and two-thirds of SAT items calculator allowed. The GED includes 46 math items, with the first 5–6 being non-calculator. Dr. Ramsberger then presented a series of sample items from each of the tests, including calculator and non-calculator items where applicable.

The ASVAB math content editor reviewed 134 released items from each of the four tests and made determinations as to whether they mapped to AR/MK, their relative difficulty level compared to AR/MK, and whether a calculator would be needed to answer the question. She found that most items did map to ASVAB, with far more mapping to MK (n = 86) than AR (n = 28). Of the 18 items that were judged to not map to either test, 6 were calculator inactive, and 12 were calculator active. The content editor rated most items as being of medium difficulty by ASVAB standards. Only one NAEP and one GED item were deemed to require a calculator to obtain the correct answer. In all, 69 were judged to be calculator neutral (won't hurt, won't help), and 59 were seen as not requiring a calculator, although it could help.

A focus group protocol was developed to obtain the input of applicants and recruiters. Travel restrictions constrained this activity, however 30 recruiters attending the Career Exploration Post-Test Interpretation training in Knoxville, February 2020, were surveyed. Of these, 12 agreed to take part in a focus group on the topic. Survey results indicated that applicants occasionally (n = 10), or frequently (n = 15), ask if they can use a calculator when taking the ASVAB. The most common applicant reactions to not being able to use a calculator are to question why (n = 10), express concern about impact on performance (n = 9), or to just accept it (n = 7). When asked their opinion about whether calculators should be allowed, 14 recruiters had no strong opinion either way, 11 felt they should be allowed, 4 agreed with current policy, and 2 expressed mixed feelings. Focus group comments included a mention that many applicants are most concerned about the math portions of ASVAB. Concern was also expressed that not allowing calculators could have an impact on examinee confidence going into the test. As recruiters, they would like to see calculators allowed, but they also recognize that math skills are required for many jobs, and tools are not always available. Mention was also made that calculators are allowed in schools, so it does not seem logical that they are not on the ASVAB.

A survey was developed to obtain input from subject matter experts in selected career fields on the importance of math skills in training and on the job and the likelihood of needing to employ those skills without the aid of a calculator. Career fields were selected based on the number of incumbents, the importance of math skills based on existing job analysis data, and a mix of jobs with AR/MK in their selection composites. An online survey was generated that asked both general and occupation-specific questions. Thus far, 25 responses have been received. Dr. Ramsberger then showed a series of graphs summarizing the results. In regard to training, 24 respondents indicated that math was at least somewhat important, and 13 indicated that those skills need to be applied without the use of a tool. On the job, 19 respondents indicated that math skills are at least somewhat important, 13 responded that hand calculations are required.

Dr. Ramsberger then summarized the steps that would have to be taken to enact a change in the current calculator policy, including: (a) developing specifications for new item development, (b) reviewing existing items, (c) field testing new items, (d) conducting test scaling and equating, (d) developing norms (e) evaluating test fairness, (f) evaluating and establish new testing times, (g) developing and implementing applicable software updates, and (h) evaluating test reliability and validity. A recent analysis put the estimated time and cost at up to 10 years and 30 million dollars.

An additional consideration is the fact that other standardized tests provide specific guidelines for the types of calculators allowed: SAT lists 82 allowed calculators; ACT lists calculators that are not allowed and/or functions that must be disabled; NAEP provides calculators to 4th- and 8th-grade students and has restrictions on the types that 12th-grade students can use; GED allows one type of calculator and provides one embedded in the test software. This issue would need to be addressed by supplying all Military Entrance Processing Stations (MEPS) and Military Entrance Test (MET) sites with calculators, requiring test administrators to screen those brought by applicants, or embedding a calculator in the ASVAB Computer Adaptive Test (CAT) software. Questions also would arise regarding how the policy would be adapted to the Armed Forces Qualification Test (AFQT) Predictor Test (APT) and unproctored ASVAB. Additional challenges would be faced in the Career Exploration Program (CEP) where hundreds of high school students may test at the same time.

At the end of the briefing, a committee member asserted that calculator use should be decided based on the constructs measured by the AR and MK tests. S/he said that is true because the ASVAB measures aptitude rather than achievement. Dr. Ramsberger concurred, saying the purposes of the ASVAB are different than those of other large-scale testing programs. Dr Segall agreed and said it would be an expensive undertaking to modify the tests for calculator use. Dr. Velgach amplified that there is a perception among recruiters and Service leadership that the ASVAB should follow suit with other testing programs, and they want to know our rationale for not doing so. She said answering that question required examining the overlap in content between the ASVAB and what is taught in high school, how ASVAB content compares to that of other standardized tests, and, most importantly, how ASVAB content aligns with the knowledge that is required on the job. She said the challenge is to put all that information together to provide a comprehensive picture that answers the question of why the "no-calculator" policy is appropriate for the ASVAB.

Dr. Ramsberger then described a meeting with Service representatives, who suggested that test anxiety might be reducing acceptance rates. He said Dr. Pommerich tried to explain what it would take to integrate the use of calculators, and why it would not change acceptance rates, but that he thought the audience probably did not grasp the explanations due to their technical nature. He said he thinks that is part of the issue: it is a foreign concept to the stakeholders and not easily explained. He said, when the ACT moved to calculator use, they conducted a study where they gathered data from a nationally representative sample, but that their analyses were so complex that it would be very difficult to describe to non-psychometricians. Ms. Miller said she has had the same experience when talking with senior leaders and, when she tried to explain that it could be a 10-year, 30 million dollar effort, they did not believe it and suggested that it should not take longer than 6 months. She said the leaders think it should take no more than programing a calculator feature into the ASVAB software. Ms. Miller said that their challenge is explaining, from a technical standpoint, that the change is not necessary. She added that the work just shown provides evidence that hand calculations are a reality of a lot of job functions in the military, and that those calculations need to be performed efficiently. She concluded by saying, absent some compelling argument, the move to calculator use will probably be required by senior leadership because it is a perception issue from their standpoint. She said it would be helpful if the DACMPT would provide their frank assessment of the matter to counteract the senior leaders' perception that the ASVAB is antiquated.

A committee member responded by saying that the answer should be the content on slide 29, which showed that hand calculations are important in most military jobs, as indicated by the subject matter experts (SMEs). The committee member said s/he did not know what better information to provide, and that the information on the slide is easy to understand. Dr. Bobbie Dirr (Air Force Personnel Center) asked if it would be possible, in the case that integration of calculators is required, to allow calculators for tests used in selection but not for tests used in classification. The committee member replied that it would probably be just as costly to implement either course of action.

Another committee member mentioned research by the Iowa Testing Program, which was considering allowing calculator use on all their subtests. S/he said when they provided the calculator option on a subtest that was much like the MK, the students became confused and took longer to complete the test because they tried to figure out how to use a calculator to solve problems that did not really require a calculator. S/he said allowing calculators on that test wound up having a negative effect on performance and anxiety. Dr. Salyer said she was proud that the CEP did not discriminate based on capabilities available across schools, but that it would be a strain if they had to allow calculator use on the paper-and-pencil (P&P) test, which she said accounts for almost 80% of their administrations. She said incorporating the use of calculators may seem like an easy fix to recruiters who want to increase their numbers, but that they should be aware that it might eliminate one of their testing options.

A committee member reinforced the point made in relation to the Iowa Testing Program research and added that calculator use also has implications for how items are written. S/he said, for instance, if item writers know calculators are going to be used, it impacts item development and construction. S/he followed by describing an unexpected finding in other research that the effect of calculator use is often differential based on overall performance level. S/he said that in some studies, the very high performing students were far less likely than the lower performing students to use calculators and, many times, the lower performing students wasted time trying to figure out how to use them. The committee member stressed the point that the impact is differential and often most damaging to the lower-performing students.

Ms. Miller followed by asking why other large-scale testing programs made the decision to incorporate calculators; were the test designers not familiar with the research? A committee member explained that, unfortunately, many large-scale assessments, particularly at the state level, are required to reflect content standards and instruction, which are often written to include, or even require, the use of calculators.

Dr. Pommerich reiterated that the Defense Personnel Assessment Center (DPAC) still must develop user-friendly documentation to summarize the outcomes and transmit their perspective on the matter. She asked if the committee had any thoughts or best practices on how to communicate some of the ideas mentioned at the meeting. Before the committee members responded, Mr. Davis suggested including the operational, as well as the technical, costs of making the change. Dr. Velgach said that was correct and cited costs such as those related to implementing U.S. Military Entrance Processing Command (USMEPCOM), as well as ASVAB-CEP requirements.

Regarding how to communicate the argument against calculator use, a committee member recommended focusing on what is actually required of recruits during training and on the job: If they cannot use calculators in those conditions, then recommend against changing the policy; but if they can, then perhaps consider changing the policy. S/he said it might be useful to collect data from a wider, more representative sample of military personnel. The committee member also recommended providing details about the work that is required as part of the 10-year, 30 million effort. Referring to slide 30, s/he mentioned specifically the need for norming, equating, and all the other changes that go into an effort like this.

Dr. Velgach thanked the committee members for their input, saying that it always carries a lot of weight with the leadership. She also mentioned that they had planned to collect more data on calculator use from Service-members, but the COVID-19 pandemic prevented that from happening. She said they will continue to work on that task.

Dr. Carretta asked if anyone knew how much the commercial testing services must “invest” to determine that they can use calculators in at least some instances. Dr. Ramsberger said ACT described what they did, but not the cost, and said he would have to look to see how long it took them to do it. Dr. Carretta said that type of information might be useful when making the argument in respect to the ASVAB. Dr. Cyrus Fouroughi (Naval Research Laboratory) then stated that the Navy takes an anti-calculator position, because they have a lot of ratings (i.e., jobs) that required hand-calculations. He cited their “Nukes” rating as an example and said that they would want an opt-out option if the policy is changed, because it would impact them operationally. He said they could put that in writing if it would help. Dr. Velgach thanked Dr. Fouroughi for the information, noting that opt-out options would be difficult to integrate into the highly standardized ASVAB program. She also asked him to provide a written statement, because other Services had provided statements to the contrary.

9. **Complex Reasoning** (Tab K)

Dr. Scott Oppler, Human Resources Research Organization (HumRRO), presented the briefing.

> Dr. Oppler began by explaining that fluid intelligence has been found to be a strong predictor of training and job success. The 2006 ASVAB Review Panel suggested that DoD consider adding a test of fluid intelligence to better balance the ASVAB's composition (between fluid and crystalized intelligence). The potential benefits of adding a test of fluid intelligence to the ASVAB include higher prediction of training and job success, lower susceptibility to compromise, and increased qualification rates for non-native and non-heritage English speakers. A previous attempt to create a test of fluid intelligence took the form of the Abstract Reasoning Test (ART). It was developed by Susan Embretson of Georgia Tech, and had a format similar to Raven's Progressive Matrices items (multiple choice, 6 or 8 response options per item). The Defense Personnel Assessment Center (DPAC) commissioned the development of one form (30 items). It was administered (for research purposes) to language training applicants in 2017. The items were found to be relatively easy and time-consuming. Dr. Oppler then presented two sample ART items.
>
> Dr. Oppler continued by stating that DPAC would like to develop a non-proprietary automated item generation (AIG) algorithm and difficulty model for a non-verbal test of fluid intelligence (aka, "Complex Reasoning Test"). The goal is to improve item development efficiency and reduce or eliminate field-testing requirements. Items should be similar to Raven's Progressive Matrices items and at a difficulty level appropriate for qualifying military applicants into jobs of varying complexity. A first step is to evaluate the viability of developing a complex reasoning test using an existing, non-proprietary item generation tool created by the Sandia National Laboratories, known as the Sandia Generated Matrix Tool (SGMT). Initially, work will be done to (a) identify the features of the items that impact their difficulty parameters and (b) develop a difficulty model to improve the efficiency associaed with the creation and calibration of the items.
>
> Dr. Oppler then summarized the research plan, which includes conducting a literature review to examine prior attempts to model difficulty of matrix reasoning items, and the Abstract Reasoning Test (ART) evaluation study conducted in 2017. The next step will be to assess item generation capabilities of SGMT and identify relevant research questions. During stage 2, a pilot study will be designed and conducted to address research questions identified in Stage 1. Concurrently, a preliminary set of complex reasoning items will be created using SGMT (k = approx. 60). The final step will be to collect and analyze the data. Stage 3 will involve developing a research plan for a larger study, creating a larger set of items (k = 200, including items surviving the pilot study), and collecting and analyzing the data to develop and evaluate a difficulty model.
>
> Dr. Oppler next summarized progress to date. The literature review has been completed. Findings include that the variance accounted for in item difficulties is appreciable, but has varied widely ($R^2 = .40$ to .80). Substantially less success has been achieved in modeling discrimination or guessing parameters. The review of the ART evaluation study has also been completed. This work was briefed to the Manpower Accession Policy Working Group (MAPWG) and DACMPT in 2019. The form was administered to 2,162 highly qualified examinees (military personnel applying for language training). Findings were that items tended to be relatively easy (60% had p-values > .80), the 25-minute time limit may not have been sufficient, and only six (of 30) items had more than three distractors selected by 3% or more of the examinees.

Work has also been completed on assessing the item generation capabilities of the SGMT. There are two item types: transformation iItems and logic items. Item generation is not fully automatic in that it requires users to specify structural elements, but the tool can then be used to develop what are essentially item clones. Several features no longer function as intended, given the outdated software (programmed in an early version of Java). Many distractors, as generated, appear to be less than plausible, requiring human intervention to make them more attractive. Items are not generated in a high-resolution format, necessitating additional human processing. However, the tool still allows for the manipulation of a variety of features that have been shown by the developers to influence item difficulty and, therefore, may be adequate for purposes of the pilot research (but probably not for larger-scale needs). It has subsequently been learned that the Army Research Institute for the Behavioral and Social Sciences (ARI) has been granted access to an updated version of the tool (developed for another Government agency) that addresses many of the problems currently associated with the original version. Likewise, ARI has also indicated interest in the development of a more thoroughly modernized version of the tool. DPAC and ARI have recently begun discussing the possibility of collaborating to more efficiently and effectively develop these item types (i.e., transformation and logic items) and to jointly pursue mutually beneficial research to determine the potential usefulness of such items.

The relevant research questions to address in Stage 2 of the plan have been identified. They include:

- What is the impact of the two different item types (transformation and logic) on the dimensionality of the test?
- How does the number of response options influence the psychometric properties/timing of the items, and what are the pros/cons of reducing the number of response options to 4?
- To what extent is performance on these items influenced by practice?
- Can the difficulty of the items be adequately controlled/modeled by systematically varying certain features of the items in order to make the test sufficiently difficult for qualifying military applicants into jobs of varying complexity and avoid (or reduce) the need for calibration with live examinees for scoring and adaptive item selection?

Ongoing work includes designing the pilot study to evaluate: (a) the possible multidimensionality associated with the two item types; (b) the impact of number of response options on the psychometric properties of items and test scores; and (c) practice effects. The study will also evaluate the extent to which items can be made sufficiently difficult for military selection and classification purposes. The current plan is to collect data from non-military examinees via the internet (i.e., using Mechanical Turk [Mturk]). Potential examinees will be screened to reflect characteristics of ASVAB examinees. A question remains regarding which version of an item generation tool to use to develop the necessary items, given that DPAC has not yet been granted access to the updated version of the tool.

At the end of the briefing, a committee member asked whether the military had found Mturk to be viable for use in military research. Dr. Oppler said he did not know if DPAC or the military had used it in the past. Dr. Segall reported that DPAC has not used it, but that HumRRO has a lot of experience with it. He then asked Dr. Oppler if there was a way to restrict the types of respondents. Dr. Oppler replied that Mturk has features that allow samples to be specified on multiple participant characteristics, but that it costs more to use those features. He said it is also possible to obtain samples of highly rated participants for a fee, and that other researchers have had success with those types of samples. He mentioned recent Society for Industrial Organizational Psychology (SIOP) presentations that provided tips for how to make the most of those types of data collections. Dr. Oppler then described some recent discussions with DPAC about options for this particular data collection and that, given it will be a pilot to collect basic information (e.g., data on different numbers of response options and item dimensionality), he thinks they would not have to rely on a specifically military sample. He said, however, when the research starts to deal with more realistic calibrations, they would need military samples. Another committee member reinforced the conclusion that Mturk allows samples to be defined in any number of ways, and that it is widely used in research. S/he said probably a third of the cohort whose professional journal articles s/he reviews are now using samples from Mturk or similar systems, which research indicates provide samples that are at least as good as college or single organization samples.

The committee member said s/he thinks the research on Raven, though a little dated, indicates very large subgroup differences, and that if it is important to look at demographic differences, there might be a low-cost method to examine the issue. Dr. Oppler said he had not explicitly built that into the design, but it is certainly something they would consider. He said they are planning to collect large sample sizes per cell because they want to look at how some of those factors impact the item calibration results. He said the numbers should be sufficient to look at the subgroup question. He then remarked that research suggests a male-female effect size, for example. That is, one meta-analysis conducted about 10 years ago showed about a third of a standard deviation effect size. He said discussions related to that type of effect have centered around the visualization aspects of some of the items. He said that, given they are planning to vary visual/spatial requirements, they are hoping to see something in the evidence about differential item functioning for particular types of features that might be varied in some item types.

Dr. Velgach said she is hoping that complex reasoning can assist in reducing the performance differences between demographic groups, especially in the English-learner population. A committee member said that is certainly possible but, the “rather lengthy” instructions might need to be reduced or presented orally. Dr. Oppler noted that most of the verbal component will reside in the instructions rather than the items themselves. The committee member concurred but reiterated that examinees will still need to understand the directions. Another committee member said, in industry, someone s/he knows has been working on a test with no verbal instructions and that examinees learn how to work the items through practice. The committee member said it might not be appropriate for this test, but it seems to work with simple item types, and that s/he would look for a point of contact who could provide more information.

The committee member then asked a question not related to English learners: assuming both types of items reflect fluid intelligence, is the plan to determine the predictive capability of both types of items. S/he said transformational items might be easier than logic items, but one or the other might be a better predictor of training performance. Dr. Oppler said the literature suggests that the logic items are more difficult than the transformational items, but that the latter can become very complicated. He said the more complicated transformation items are more comparable, in terms of difficulty, to the logic items. He added that the problem with the findings related to the logic items is that they are administered after the transformation items, and so the examinee does not arrive at the logic items with the kind of mindset it takes to answer those items. He said it is not clear whether the difficulties result from the change of item set. He said that, after looking at a few of the logic items, a pattern emerges, and the logic items are not as difficult as the transformation items. He said they can make the logic items “more crowded,” but once a person is aware of the pattern, they may not be as difficult as the transformation items.

Regarding the determination of concurrent or predictive validity, Dr. Oppler said they are still not close to being able to collect data from people that will have criterion data available. Dr.

Segall replied that the issue could be explored in the training schools, but they have not thought it through at that level of detail. He said the first step is to assess dimensionality – if the item types load on the same dimension – and then determine if difficulty is an artifact of the way the items are presented (e.g., the sequence). He said if Study 1 shows the item types are measuring the same dimension, they might be able to develop transformational items that span the full range of difficulty and eliminate the logic items. He said they would want to get validity information afterwards. A committee member then mentioned that s/he had seen Mturk studies that collected criterion data and suggested even self-report criterion data might be useful. Dr. Oppler said that is an option to consider.

10. Adverse Impact (Tab L)

Dr. Greg Manley, DPAC, presented the briefing.

> Dr. Manley began the briefing by defining adverse impact (AI) as the unintended discrimination of a protected class due to the result of a selection procedure. AI is not a property of a test per se. However, AI may occur when a test's scores are used as the basis for selection. A selection test may contribute potential for AI when it shows sizable mean test score differences between a majority group and a protected class (minority). Effect sizes of the standardized mean difference gives us an index to examine a test's potential for AI. The four-fifths rule is often used to determine the occurrence of adverse impact: "A selection rate for any race, sex, or ethnic group, which is less than four-fifths (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact." (Section 60-3, Uniform Guidelines on Employee Selection Procedures [1978]; 43 FR 38295 [August 25, 1978].) The ratio comparing the selection rates is called the impact ratio. Dr. Manley then presented a formula for calculating the impact ratio and determining the significance of that value and the confidence intervals around it.
>
> The four-fifths rule and accompanying statistics are applied to the ASVAB by comparing qualification rates across the focal and reference groups of interest with regard to: examinees who qualify for entry into the military (i.e., those scoring in Armed Forced Qualification Test [AFQT] category IIIB or higher, AFQT ≥ 31); examinees who qualify for enlistment incentives (i.e., those scoring in AFQT category IIIA or higher, AFQT ≥ 50); and adverse impact assessed using initial test scores only (i.e., scores from retests or confirmation tests are excluded from the analyses). Dr. Manley noted that significance testing is not necessarily useful for analyses with very large numbers of applicants (i.e., > 2000). Effect sizes (i.e., standardized mean differences) provide a method of evaluating potential for adverse impact across individual ASVAB and special tests, where no direct selection occurs. Dr. Manley then presented a formula for computing effect sizes and their confidence intervals. He explained that effect sizes can be plotted and classified with respect to Cohen's standards of evaluation, with small effect sizes starting at .20, moderate at .50 and large at .80. He explained that the ASVAB testing program evaluates AI by comparing males and females, non-Hispanic Whites and Hispanic Whites, Non-Hispanic Whites and Non-Hispanic Blacks, and non-Hispanic Whites with non-Hispanic Asians. He noted that non-Hispanic Asians now represent more than 2% of the applicant population.
>
> Ideally, AI is assessed on a regular basis. The data presented here measured AI for applicants testing in fiscal year (FY) 2019 (October 1, 2018 through September 30, 2019). In the past, AI was evaluated in FY05, FY09, FY11, FY13, FY15, and FY17. Dr. Manley then displayed a chart showing the FY19 sample sizes for each of the groups of interest. He continued by presenting charts showing the following results:
>
> - Impact ratio and 95% confidence interval for AFQT cut scores FY 2019 IIIB and IIIA, all education levels
> - Comparison on impact ratios for FY09, FY11, FY13, FY15, FY17, and FY19
> - Effect sizes and confidence intervals for ASVAB scores for FY19 on males versus females, non-Hispanic Whites versus Hispanic Whites, non-Hispanic Whites versus Hispanics, non-Hispanic Whites versus non-Hispanic Blacks, non-Hispanic Whites versus non-Hispanic Asians

- Comparison on impact ratios for FY09, FY11, FY13, FY15, FY17, and FY19 on AFQT test score and non-AFQT test scores for: males versus females, non-Hispanic Whites versus Hispanic Whites, non-Hispanic Whites versus Blacks, and non-Hispanic Whites versus non-Hispanic Asians.

Dr. Manley continued by noting that the magnitude of impact on the ASVAB has remained generally consistent across fiscal years, but still varies in size from negligible to large across tests and groups. A comparison of impact across different testing programs gives some indication of whether the observed FY19 magnitudes are reasonable. Sufficient information for estimating effect sizes is available online for two other large-scale testing programs: SAT – 2016 College Bound Seniors (Math and Reading) and National Assessment of Educational Progress (NAEP) – 2015 Grade 12 (Reading, Math, and Science). Dr. Manley then presented a series of charts showing the comparisons of effect sizes across testing programs (i.e., math, reading/verbal, and science content areas) for males versus females, non-Hispanic Whites versus Hispanics, non-Hispanic Whites versus non-Hispanic Blacks, and non-Hispanic Whites versus Asians.

Dr. Manley continued by stating that, for the AFQT tests (and General Science), the direction and magnitude of overall impact is generally consistent with that observed on comparable SAT and NAEP tests, which suggests that impact on ASVAB tests may reflect legitimate differences in the studied groups. Comparisons across programs may be somewhat restricted due to differences in group definitions, testing populations, test content, etc. "To the extent that members of one group do more poorly on a subset of items that are a legitimate part of the content domain, we would be reluctant to call the discrepancy evidence of bias" (Shepard, 1987). Adverse impact does not reflect bias if validity research shows that the test is equally valid for relevant groups. Historically, a regression-based approach has been advocated to evaluate the existence of bias. Lack of bias is indicated when the regression line relating the test score [X] and a criterion [Y] is the same for each group. Previous research on the ASVAB technical tests showed similar prediction lines across (1) males and females and (2) Blacks and Whites (Wise et al., 1992), suggesting no bias for the tests and groups studied. The Defense Manpower Data Center (DMDC) recommended in 2010 that an updated validity study be conducted for relevant tests and groups. However, a lack of access to criterion data across Services (except Air Force) presents an impediment to updating the study. More recent thinking in the realm of bias detection is that regression-based approaches may not accurately reflect bias. The recent acquisition of training outcome data from the Services may make it possible to examine AFQT for test bias. Moderated Multiple Regression (for Final School Grade) and Logistic Regression (for Pass/Fail) may be used to evaluate intercept and slop bias for AFQT scores and group membership when predicting training outcomes. The lack of variance on the Pass/Fail criterion presents a challenge (most individuals pass). Reducing potential impact will be a high priority when considering revisions to the ASVAB and AFQT contents.

Dr. Manley next addressed special tests that are administered on the ASVAB platform: Mental Counters (MCt) is a counting test of working memory currently being administered by the Navy. The Cyber Test (Cyber) is a test of basic computer and information systems knowledge used by all Services. Coding Speed (CS) is a speeded test of assigning code numbers to words and is also used only by the Navy. Dr. Manley then displayed a series of charts showing FY19 effect sizes and confidence intervals for:

- Special test scores: males versus females, Non-Hispanic Whites versus Hispanic Whites, Non-Hispanic Whites versus Hispanic, Non-Hispanic Whites versus non-Hispanic Blacks, and Non-Hispanic Whites versus non-Hispanic Asians
- ASVAB test scores: males versus females, Non-Hispanic Whites versus Hispanic Whites, Non-Hispanic Whites versus Hispanics, Non-Hispanic Whites versus non-Hispanic Blacks, and Non-Hispanic Whites versus non-Hispanic Asians in MCt Cyber, and CA samples.

Dr. Manley continued by noting that MCt, Cyber, and CS generally exhibited small to moderate effects, usually as low as or lower than most ASVAB tests. White-Black comparisons were generally larger for MCt than for the other group comparisons. CS usually had very small effects (near 0), *however*, this test suffers from other issues. For example, it can be affected by lag time in internet delivery, given that it is a speeded test. It is also known to be affected by the test delivery device and suffers from coachability and susceptibility to invalid strategies that result in high scores. Dr. Manley concluded by noting that the potential for AI is not the only consideration for making changes to the ASVAB.

After noting the consistency in the AI findings over time, as well as with other national measures of similar constructs, a committee member said the magnitude of the disparities remains troubling and probably reflects many factors. S/he said NAEP and other large organizations have attributed the disparities to segregation and differential opportunities to learn, which s/he characterized as powerful explanations that speak to the importance of preparation for opportunities like military careers. The committee member then suggested that the difference between effect sizes of, for example, .04 and .08, though described as small, are quite different, in that .08 is twice the size of .04. S/he said such differences affect a lot of people in large populations of examinees, like military applicant samples. The committee member followed, however, by saying s/he appreciated the emphasis on AI versus effect size, because while effect size denotes differences between scores, AI measures the impact on selection, which is where the impact of segregation and opportunities to learn affect careers and continued learning. S/he said the main message over the years is that the ASVAB creates no more disparity than do the other national standardized assessments, and that is a strong statement, in fact, stronger than the implications of effect size differences.

Dr. Manley agreed that AI is the bottom line and said he wished it could be determined for the other tests as well. He said the best they can do for the other tests is to measure effect sizes. He also recognized the striking nature of the differences in effect sizes between the military tests and other national standardized programs and said the Asian versus non-Hispanic Whites (shown on slide 37) is a good example. Referring to the Asian versus non-Hispanic White differences, Dr. Velgach suggested that the larger effects found in military testing might be due to a greater language barrier in that population. She said they were looking for alternate approaches to reduce the impact of that barrier.

Another committee member agreed that AI is more consequential because it relates to decisions affecting opportunity. S/he said the Black-White ratios suggest that less than half the proportion of Blacks, as compared to Whites, are being selected, which restricts opportunities for African Americans. S/he said this may be justified by evidenced outcomes, but it remains a potential concern. The committee member also mentioned being struck by the results concerning training outcomes in that, if nobody fails, it is an ineffective criterion measure. S/he suggested that future predictive bias studies may require identifying long-term criteria, such as promotion rates, reenlistment, and other more distal outcomes that demonstrate greater variability. Dr. Velgach explained that the next step is to look at prediction by subgroup using other outcome data that are available from the Services, with one potential outcome being attrition. Dr. Segall said DPAC is looking into alternative criterion data, and he is still hoping to use final school grades from one or two of the Services, which he said should have more variability. He also said it would be a good idea to look at promotion rates.

Another committee member echoed the significance of the criterion problem and asked if it would make sense to collapse across years in some of the high-density occupations. S/he said collecting data at the same point in time over several years for specific occupations or groups of occupations is something that other organizations do. S/he said every employer is obligated to look at AI and assemble the data that attest to job relevance and business necessity in support of bias studies. S/he said that approach might make sense for the military, given that trends reveal only small differences across years, but it would require job requirements to remain generally stable over time. Dr. Manley asked if the committee member was speaking about performance or attrition measurement. The committee member said there might be a performance measure common to a fairly large group of people, such as a high-density occupation. S/he said looking at 3 years (e.g., 2017-2019) might provide a sample sufficient to demonstrate that the tests are not biased, even though AI might remain troubling.

Dr. Manley said he concurred and explained that they already have data from the Services that go back a number of years. He then suggested they should use fewer than 10 years of data, because the situations change over time to a degree that might affect inferences drawn about the current situation. Regarding training outcomes, he said Air Force data would be the most useful because those data include course grades. He then concurred that promotion rates, number of promotions, reenlistment, and attrition were all solid measures of performance that, unlike regular (i.e., semi-annual) evaluations, demonstrate notable variance.

A committee member asked if variability in pay raises is locked into tenure and if that might be another criterion possibility. Dr. Manley said pay raises are tied to promotions as well as tenure. The committee member then asked if pay raises were at all tied to merit, to which Dr. Manley responded that bonuses might be, but probably not pay raises.

A committee member asked if DPAC, despite the difficulty posed by computer-adaptive item administration, has been able to explore differential item functioning (DIF) as another source of potential bias. Dr. Segall said they routinely examine DIF during item pool development, and because items are piloted with applicants, they have sizeable samples for many demographic groups. He explained this allows them to look at subgroup differences, such as those between Blacks and Whites. The committee member then noted two items of importance: First, when disparities exist, the differences do not necessarily indicate bias. Second, the AI is not negligible, and it might be more worthwhile to try to understand the AI rather than the mean differences. The committee member again asserted that mean differences are not that interesting, especially when they are so similar to those found in other large-scale assessments. S/he said the real issue is the consequences of those differences in decision making.

Dr. Segall remarked that the racial and ethnic composition of the U.S. military mirrors, to a large degree, the American population, especially in the Black-White ratio. He said that seems curious, given that Blacks qualify at a much lower rate than Whites, but the answer lies in the fact that the Black applicant group is proportionally larger than expected, when compared to the proportion in the American youth. He said the higher Black application rate cancels out any AI caused by testing. He added, if there was no AI, the proportion of Blacks in the military would be much higher than the proportion of Blacks in the population. He then commented that the military is judged, in part, by the composition of the enlisted corps. Against that standard, he said, the military looks to be in good shape despite the AI numbers.

A committee member responded that s/he would hesitate to suggest that a disadvantaged group ought to be overrepresented in the military, because the military was criticized heavily for that during the Vietnam War. Dr. Segall said he would avoid wading too deep into that discussion, because it is a policy matter. Another committee member suggested, however, that in peace time, it might be more important to ensure accessibility to military opportunities than to reflect population ratios. The first committee member expressed interest in seeing racial composition differences by occupation, noting that the military has also been criticized for overly representing African Americans in branches that are perceived as being more dangerous, such as the Infantry. S/he added that there are likely no easy answers to these questions.

Dr. Velgach then remarked on the need to ensure the standard is valid across groups, even if it results in differences in opportunity. This prompted agreement from a committee member, who reiterated the importance of understanding the differences right around the selection cut point.

In reference to the comments on occupation-specific compositions, another committee member noted that all military jobs are low-paying jobs, but that it would be good to see disadvantaged groups represented sufficiently in technical occupations that provide greater opportunity in post-military careers. Dr. Velgach explained that the current AI analyses are constrained by the fact that much of the requisite data are held by the Services. She said the plan is to ask the Services to do some of the analyses on occupations and the related qualification composites, which are developed and maintained by the Services.

A committee member noted that, in private industry, AI is so undesirable that employers often look at ways to isolate the AI to a certain job, locale, or timeframe. S/he said no organization does what DPAC is doing, which is to look at everyone who applies. The committee member then remarked that it may not be appropriate for the military to conduct AI analyses by job, because applicants are not applying for jobs, so much as they are applying to a Service. S/he then asked if it might be appropriate to examine AI by locale, such as by recruiting stations or geographic areas where AI is significantly worse, or better, to understand the phenomenon. S/he said it might be helpful to look at the characteristics of applicants in those areas to determine who is contributing to the subpar AI results. Dr. Segall said he liked the idea, and that they have access to applicant geographic data.

11. Testing – Next Generation (Multi-Dimensional Evaluations and Future Vision) (Tab M)

Drs. Mary Pommerich and Tia Fechter, Defense Personnel Assessment Center (DPAC), and Dr. Scott Oppler, Human Resources Research Organization (HumRRO), presented the briefing.

> After providing an overview of the briefing, Dr. Pommerich noted that the ASVAB underwent a systematic review in 2005–2006, with testing experts making recommendations for improvements and enhancements to the military's Enlistment Testing Program (ETP). The panel was motivated by a difficult recruiting environment and the belief held by some that the ASVAB was outdated and in need of an overhaul. The Manpower Accession Policy Working Group (MAPWG) condensed and prioritized the panel's recommendations. A modified Delphi approach was used to prioritize the condensed recommendations. Dr. Pommerich then showed a list of 17 recommendations. Those receiving a priority rating of 1 included implementing ASVAB Computer Adaptive Test (CAT) at Military Entrance Testing (MET) sites, considering classification accuracy when evaluating content changes, and reevaluating the contents of the ASVAB.
>
> Many changes have been introduced in the Enlistment Testing Program (ETP) because of the 2005–2006 review, however the contents of the ASVAB itself have not yet changed. Prior discussions of possible changes have floundered due to (a) a lack of consensus on the philosophy of the ASVAB; (b) logistical difficulties associated with making changes (such as dropping subtests) that would impact existing

composites and systems set up to operate on those composites; and (c) concerns about insufficient resources to accommodate a revised ASVAB that would take more time than the current battery (if new tests of interest were added to the current battery). Given the complexities associated with making changes to the ASVAB, DPAC now believes it is best to consider all new and existing tests at once, rather than on a case-by-case basis. DPAC hopes to resolve the ASVAB impasse via Next Generation Testing efforts. Key steps include (a) studying new tests of interest, (b) evaluating the tests in the ASVAB, (c) consolidating information gathered to aid decision making about the status of individual tests, and (d) conducting focus groups with stakeholders to develop a shared vision for Next Generation Testing. Key questions to be answered through this process are what tests should be included in the ASVAB, or ASVAB platform, and what other changes are needed to modernize the ASVAB/ETP.

Next Generation Testing efforts will focus on the ASVAB, as well as the special tests that are administered alongside the ASVAB in the ETP. ASVAB and special tests are administered jointly on the ASVAB platform and share a common look. Special test scores are used in addition to ASVAB scores for classification purposes. A key distinction is who is responsible for development and maintenance of the tests. DPAC has responsibility for the ASVAB; Service proponents have responsibility for the special tests. Due to the limited time for total testing, it is necessary to consider all tests to be administered on the ASVAB platform in conjunction.

Dr. Pommerich then began a progress report on ASVAB modernization efforts. Efforts to develop or refine new tests of interest, such as the Tailored Adaptive Personality Assessment System (TAPAS), the Cyber test, Mental Counters, and Complex Reasoning are ongoing, as is an evaluation of the tests currently in the ASVAB. An argument-based approach to validation of the ASVAB is also ongoing, while this has been completed for the Armed Forced Qualification (AFQT) tests. The psychometric checklist for evaluating new tests also continues to be reviewed and updated. The Services/proponents complete the updated psychometric checklist for new tests of interest, documenting all new information since a checklist was last completed. This will allow stakeholders to develop a shared vision that defines the purpose and general makeup of the ASVAB and ETP for Next Generation Testing. A MAPWG focus group was held to further this process; plans for future focus groups will be summarized later. Next it will be necessary to establish a systematic process for evaluating potential changes and making decisions regarding tests in the ASVAB and the ETP. This has been initiated and will be discussed later in this presentation.

After the ASVAB evaluation and focus group, logistical questions with stakeholders, including the feasibility of lengthening the ASVAB and the feasibility of dropping existing tests, will be revisited. Stakeholders will summarize the impact of potential modifications to the battery and identify resources to support a revised battery. The information will then be compiled to allow for identification and discussion of potential changes to the contents of the ASVAB and the special tests administered in the ETP.

Dr. Pommerich continued by recapping the new tests currently of interest. These include the Cyber test. New forms are being developed to address issues of compromise vulnerability and obsolescence. A CAT version is also in process to better target item difficulty to applicant ability. The TAPAS is being discussed in light of recommendations from the TAPAS Evaluation project. Improvements to the instructions for the Mental Counters test are being studied as a way of eliminating the persistent floor effect in the applicant population. Finally, a non-verbal test of fluid intelligence (Complex Reasoning) is under development with items modeled after Raven's Progressive Matrices items. Dr. Pommerich noted that, due to resource constraints, total testing time across the ASVAB and special tests (as well as potentially outdated content) will be a key consideration for Next Generation Testing. A chart displayed the testing times associated with each of the ASVAB and special tests. Dr. Pommerich stated that there is a strong interest in assessing how the ASVAB might be modified to accommodate the new tests.

Dr. Pommerich then turned to the ASVAB evaluation, which was motivated by ongoing research to thoroughly evaluate new tests of interest, which has not occurred for the ASVAB itself. A comprehensive assessment of the current battery will give insight into its utility, quality, and potential for modification. Potential changes to the ASVAB could include dropping, shortening, or combining existing tests and/or merging new tests with existing tests. Dr. Pommerich then presented a list of 11 efforts being conducted as part of the ASVAB evaluation effort and a separate list of the individuals involved. She then reviewed the progress made on each of the steps.

- Step 1: Trace History of the Current Tests. The goal is to document where the ASVAB tests came from and why they were included in the battery. It has been completed. Dr. Pommerich showed a table listing which of the current tests were in the battery at various points in time since 1968 and provided an overview of the origins of each of the tests.
- Step 2: Complete Psychometric Checklists. The goal is to complete the psychometric checklist for the current ASVAB test and Coding Speed and evaluate the psychometric value and limitations of each test. Final checklists have been completed for Assembling Objects (AO), Arithmetic Reasoning (AR), Math Knowledge (MK), and Paragraph Comprehension (PC). Draft checklists have been completed for General Science (GS), Work Knowledge (WK), Mechanical Comprehension (MC), Electronics Information (EI), and Coding Speed (CS), and a draft is in process for Auto Information/Shop Information (AI/SI). Initial pros and cons have been identified for GS, AR, WK, PC, MK, EI, AI/SI, MC, and AO. Dr. Pommerich then presented an example of this work, the pros and cons for MK. Two tables showed the pros and cons for each of the tests.
- Step 3. Evaluate Usefulness and Appropriateness. The goal is to evaluate the usefulness and appropriateness of existing tests for the current youth population. One aspect of this is to track trends in test scores over time, which has been completed. Dr. Pommerich then listed the steps taken to accomplish this and presented a series of charts showing score trends from 1985 to 2018 for each subtest overall, by gender, and by race. A table provided a stability rating on a 1-10 scale for each test. Another step taken was to evaluate the overlap between latent ability and score information for the current testing population. A series of charts displayed these outcomes. Another step that has been completed is to evaluate what fraction of the population possesses the knowledge/skill assessed by each test by examining the overlap between latent ability and score information. Dr. Pommerich showed a table providing a finiteness rating for each of the subtests on a scale of 1 to 10. A final step in the process will be to use job/task analysis ratings to evaluate the relevance of content contained in the science and technical tests to success in technical training. This is in progress with a plan being formulated to collect subject matter expert (SME) judgments.
- Step 4: Evaluate Item Development Costs. The goal is to identify yearly costs for item development for each test, as well as the desired form replacement schedule, the number of items needed per year, and the total yearly cost. Dr. Pommerich then showed a table listing each subtest and a total yearly cost rating on a 1-10 scale.
- Step 5: Evaluate the Ease of Developing Good Items. This includes assessing the finiteness of domains, given that tests with limited content domains will result in more difficulty developing good items. The potential for using automated item generation (AIG) programs was also assessed, as were the item retention rates. Dr. Pommerich showed a table listing the subtests and their finiteness ratings on a scale of 1-10. Another table provided feasibility of using AIG to develop items for each test, which was also rated on a 10-point scale. Dr. Pommerich then showed charts that displayed retention rates for items by subtest.
- Step 6: Evaluate the Durability of Test Content. This includes assessing the extent to which test content is less relevant to today’s testing population, and the extent to which content is likely to require changes or updates in the near or long term. A table showed relevancy ratings for each subtest on a 1-10 scale. Another table showed obsolescence ratings for each subtest.
- Step 7: Evaluate the Efficiency of Content Coverage. This includes considering prior research examining the redundancies in content coverage across tests, gaps in content coverage, and potentially unnecessary content coverage. A table showed content efficiency ratings and recommendations from prior research.
- Step 8: Evaluate Vulnerability to Compromise. This involves considering features of tests and item pools that could make them vulnerable to compromise, as well as examining previous incidences of compromise. A table provided vulnerability ratings for each of the subtests.
- Step 9: Evaluate Other Vulnerabilities. These include coaching, practice effects, hardware effects, mode effects, local dependency, and device familiarity. One table provided ratings for each of the subtests regarding coachability. Another table provided practice effect ratings. Another table

provided ratings on a 10-point scale indicating the likelihood of score differences across paper-and-pencil (P&P) and CAT modes. Another table provided local dependence ratings (based on susceptibility), effort exerted to identify enemy items, and item selection controls. Device familiarity ratings assessed the degree to which familiarity with the device on which the test is being given could impact performance and/or response times.

- Step 10: Evaluate the Efficiency of Each Test. This involved examining testing time allotted and testing time used. Dr. Pommerich presented a table showing the number of scored questions for each subtest, their reliabilities, time allocated, and time observed. Efficiency ratings were calculated by dividing reliability by mean time spent.
- Step 11: Synthesize Findings. The goal is to synthesize findings across all evaluation criteria and tests and summarize the desirability and expendability of each test. A table displayed the ratings for each subtest on each of the criteria, as well as an average for each test across ratings.
- Step 12: Evaluate Psychometric Impact of Shortening Tests. One step in this process is to review DAC briefings and feedback from prior discussions related to shortening the ASVAB in the Career Exploration Program (CEP). This has been completed. Another step is to evaluate the potential impact on CAT-ASVAB test precision. This was done using real data to compute the latent ability means and standard deviations for the total group of all subtests. The next step is to simulate CAT-ASVAB data and compute measures of precision for selected shortened lengths. The impact of shortening tests on test validity was completed by summarizing content coverage under current test lengths as a baseline for evaluating the impact of shortening them. Next steps include conducting simulations and estimating the change in validity coefficients for the shortened tests. The impact of shortening tests on qualification rates will be accomplished by simulating CAT-ASVAB data and comparing qualification rates for current versus shortened test lengths for both AFQT and Service composites. The potential impact of shortened tests on adverse impact ratios and effect sizes for demographic groups of interest will be accomplished using real data to compute the latent ability mean and standard deviation for demographic groups, and simulating CAT-ASVAB data using N (mean, SD) distributions for demographic groups and comparing qualification rates, impact ratios, and effect sizes for current and shortened test lengths. Simulations of CAT-ASVAB data using observed average item latencies will be used to project item and total response times for current and shortened tests to evaluate the potential impact on testing time.
- Step 13: Impact of Using a Math Composite. The goal is to evaluate the psychometric impact of shortening AR and/or MK and computing a composite score (labeled Mathematics Expressions - ME) to use in place of AR and MK scores. This will involve (a) identifying options for shortening AR and MK and computing a composite score; (b) reviewing CAT-ASVAB history for computing an AS composite from AI and SI scores; (c) simulating CAT-ASVAB data and comparing reliability, qualification rates, and impact ratios for AFQT scores created from Verbal (VE), AR, and MK versus VE and ME; and (d) evaluating the impact on Service composites. Dr. Pommerich provided a summary of the status of each of these steps, which are or will be in progress.
- Step 14: Feasibility of Combining AR and MK into One Test. The steps involved in this assessment include (a) reviewing AS to AI and SI history to identify potential issues in combining two tests into one, (b) evaluating the dimensionality of AR and MK, (c) identifying feasible options for combining into a single test, and (d) evaluating the feasibility/desirability of using multidimensional CAT. Regarding AI/SI, dimensionality studies of the P&P AS data showed statistically significant two factors with low correlation (~0.60). The AS pool was split into separate pools for AI and SI items, and the AS (composite) score was derived from unidimensional scoring of AI and SI items separately. An examination of the dimensionality of AR and MK showed a high correlation between AR and MK scores (average correlation value from P&P test scores was ~0.72; average correlation value from CAT test scores was ~0.74). Bilog MG calibrations were conducted on the combined AR/MK data and separate AR/MK data on several P&P forms (25F/G; 26F/G). A unidimensional model fit the combined AR and MK data reasonably well. Correlations of the discrimination parameter estimates from the separate and combined calibrations were all greater than 0.9 except for MK in one P&P form (26G; 0.78). Correlations of the difficulty parameter estimates from the separate and combined calibrations were all greater than 0.99. Correlations of the guessing parameter estimates were all greater than 0.9. Confirmatory analyses using iFACT and a bi-factor model also indicated that a

unidimensional model would fit the examined data adequately well. Correlations from the G-factors of the one-factor and bi-factor models are high. Explained common variances (ECV) by the G-factor in the bi-factor model are high (> 0.8). Next steps include conducting further investigation of content coverage, examining the differential validity of AR and MK, determining the item selection algorithm (content balance or split pool), and determining an appropriate MA score definition.

- Step 15: Psychometric Impact of Combining AR and MK. The steps involved here include (a) evaluating the potential impact on test validity, (b) evaluating the potential impact on CAT-ASVAB test precision, (c) evaluating the potential impact on qualification rates for the total group (d) evaluating the potential impact on adverse impact (impact ratios and effect sizes) for demographic groups of interest (as defined in adverse impact analyses), and (e) evaluating the potential impact on CAT-ASVAB testing time. Dr. Pommerich then provided greater detail on how each of these steps will be accomplished.
- Step 16: Feasibility of Combining EI and Cyber into One Test. The steps involved here include (a) evaluating the content overlap between EI and Cyber, (b) evaluating the dimensionality of EI and Cyber, (c) identifying feasible options for combining EI and Cyber into a single test, and (d) evaluating the feasibility/desirability of using multidimensional CAT. Dr. Pommerich then provided an overview of the steps to be followed. She also provided an overview of a related step (16.5), which is evaluating the feasibility of combining GS, EI, and MC, which involves the same processes as described in Step 16.
- Step 17: Psychometric Impact of Combining EI and Cyber (CE). This involves creating CAT pools for CE and repeating the steps outline earlier for evaluating the psychometric impact of combining AR and MK into a single test. Similar processes will also be used to evaluate the psychometric impact of combining GS, EI, and MC into a single test.
- Steps 18-24: Evaluate the Psychometric Impact of Dropping Existing Tests, including AI, SI, AO, EI, MC, GS, and WK. The Air Force is conducting a related effort that could meet this goal. DPAC will determine the need for additional work upon completion of the Air Force effort.
- Step 25: Synthesize Findings. The goal is to synthesize and condense the findings of all steps into one rating per test.

Dr. Fechter presented a more detailed discussion of the synthesis process. She began by noting that most criteria that subtests are evaluated on have relative degrees of importance compared to one another. The number of criteria is too vast to readily make judgments on the relative importance of each ASVAB subtest. The synthesis process provides for a more rigorous, research-backed approach for rating the quality (and potential expendability) of the ASVAB subtests. Three methods for achieving this goal are under consideration. The first is the Delphi Method, which is a decision-making approach that engages a panel of experts in providing their opinions on matters of importance for use in determining what could or should be with respect to policy, goal setting, forecasting future outcomes, and so on. It is designed to reduce characteristics of group/panel discussions that can hinder concrete decision making: It reduces influence of a dominant voice within a group discussion by eliciting independent input through a questionnaire format. It inhibits irrelevant or redundant material by employing an iterative process and providing summarized feedback for discussion and consideration before the next round of judgments are made. Finally, it mitigates group pressure by making use of a median statistical index to avoid the need for conformity/consensus. Limitations of the Delphi method include that it (a) is time consuming, (b) can be negatively impacted by low response rates, (c) may be subject to bias if the summary process allows investigators to inadvertently impose their own views, (d) can be affected by differential knowledge of participants, and (e) does not account for possible interactions of future events. Dr. Fechter then provided an example of the Delphi method being used to establish ASVAB priorities. MAPWG members served as the experts and rank-ordered 17 recommendations using web-based surveys and a 5-point Likert scale. Ratings were elicited on the recommendations and a set of criteria (e.g., timing, cost, benefits). The result was a prioritized list of recommendations. For the ASVAB evaluation, a possible approach would involve having experts rate the importance of each scale developed for evaluation of ASVAB subtests on various criteria and establishing weights for each evaluation criterion based on the Delphi process. The importance ratings and the scale values assigned for each subtest could be used to determine a single “importance” rating that can be used to rank order the subtests.

The second synthesis method discussed was cross-impact analysis. This is a tool used to evaluate the probability of future events or states and emphasizes the interactions between possible future states, changes, trends, or decisions. It involves four phases.

- Exploration: State possible interaction among events.
- Probabilistic: Determine how probabilities are elicited. Judge events as stand-alone and adjust for cross-impacts post hoc. Include the possibility of the cross-impacts.
- Synthesis: Determine how to collect and summarize judgments.
- Application: Collect judgments and evaluate whether adjustments are needed due to non-coherent input (lack of convergence). Implement multiple rounds using either a game-based format or Monte Carlo simulations.

Limitations to this method include that it (a) relies on the adequacy of pre-determined probabilities and specification of cross-impacts, (b) can result in tedium and fatigue as the number of conditional probability judgments to be made increases, and (c) accounts for interactions only among pairs of events and not higher order effects. However, there is some flexibility in how it is carried out. For instance, symbolic emphasis or impact may be judged in a matrix using a coding scheme (e.g., ---, --, -, 0, +, ++, +++), and verbal descriptions may be used (e.g., How does event A impact event B?). Dr. Fechter concluded by suggesting that, for the ASVAB Evaluation, a synthesis of Delphi and Cross-Impact could be employed, by making determinations about possible interactions (e.g., if Cyber is added, what is the probability of Cyber and EI being combined?).

The third approach is called Utility Analysis. This is a decision-making tool that assigns importance or monetary values to various criteria to evaluate the institutional gain or loss anticipated from various possible courses of action. Decisions are made to maximize benefits, while reducing associated costs for those benefits. Dr. Fechter then displayed a formula for calculating a quality indicator, or the dollar payoff to the organization for use of a particular decision tool or event. Some considerations in using this approach include (a) the value of various outcomes needs to be expressed in "equal units of satisfaction," which are additive over many decisions *or* must be treated as ordinal; (b) it is often costly to engage in the required accounting process for the algorithm inputs; (c) some believe the choice of a dollar metric leads to a false sense of precision and may be misleading to policy makers; and (d) the mathematics of Decision Theory are involved and laborious. For the ASVAB evaluation, one approach would be to convert all evaluation plan metrics to dollar scales, where each indicator of quality is monetized. For instance, subtest vulnerability, like susceptibility, can be converted to anticipated costs if piracy were to take place. The evaluation would involve comparing the added utility of each ASVAB configuration in the plan.

Dr. Fechter concluded this portion of the presentation by outlining next steps, which include seeking reactions from the MAPWG and DACMPT regarding the approaches, conducting a trial approach with the goal of arriving at a single importance rating for each ASVAB subtest and special test identified for inclusion in the ASVAB, and reporting on the findings.

Dr. Oppler then presented plans to hold focus groups with military testing stakeholders and users to gather information to develop a shared vision of the next generation of testing. The central questions are, what tests should be administered as part of the ASVAB or on the platform in the future, and what other changes are needed to modernize the ASVAB and ETP? The hope is to synthesize the findings and develop a pathway forward that will converge on a possible solution acceptable to all. An initial focus group was held with the MAPWG over a half-day at February's in-person meeting. Sofiya Velgach gave a brief review of the recruiting process and how the ASVAB and special tests are used by the military for selection and classification. Dr. Oppler then led a guided group discussion and information gathering. MAPWG members were asked to speak to their perspectives as MAPWG members, not to other stakeholders' perspectives. Participants were told that the objective was to collect information, not to reach consensus. There was some redundancy across questions, which could be helpful to consider things from slightly different angles. Participants expressed consent to have the session recorded. Key discussion points were:

- Do we need to change the ASVAB/ETP? This included identifying likes/dislikes about the status quo, primary reasons for change, and specific goals to be accomplished through change.

- What would we build if there was not already an ASVAB; that is, if we were starting from scratch?
- What should the ASVAB/ETP predict, and what outcomes should a revision focus on?
- What stakeholders should be involved in the process?

The perceived psychometric positives about the ASVAB included that the AFQT is predictive of job performance, all scales/tests have good measurement precision in addition to predictive validity, and it has no more adverse impact that other measures of similar constructs. In regard to content, Math/Verbal combination plus technical tests gives the ability to do classification (cannot just be a cognitive ability test; needs to keep the technical side) and it measures both crystalized and fluid intelligence. The perceived positives in terms of administration are that it is common across Services, can be administered in both P&P and CAT formats, and the pending internet CAT-ASVAB allows testing prior to shipping to Military Entrance Processing Station (MEPS) facilities. Finally, it was agreed that the ASVAB has a long history and reputation.

The perceived psychometric negatives include the presence of adverse impact (although still less than other tests). There is not a great balance between population aptitude and what the Services need for some military occupational specialties. Finally, the perception is that score reports are not easy to understand. In regard to content, the opinion was offered that it is not broad enough and should be expanded to add fluid intelligence and non-cognitive measures. Non-cognitive assessment is part of the ETP, but not of ASVAB. There should be a DoD standard, joint-Service assessment for this purpose. Finally, current Cyber items quickly become obsolete. In terms of administration, the perception is that administrative time is too long, with too many tests required in one session. There is a lack of parallelism in modes of administration for ETP and CEP (CEP requires P&P). Finally, P*i*CAT needs to be proctored to get good results. Other concerns raised were the general perception that the ASVAB does not change with time (e.g., same ten subtests; no calculator; time since last renorming). The name of the battery is outdated ("vocational"). There is an inability for the testing/psychometric voices to be heard by higher-ups, who often do not understand technical issues (e.g., changes to the ASVAB would require renorming). Finally, there is suboptimal communication between research entities.

A variety of issues emerged regarding what is perceived as missing from the current battery. These include non-cognitive assessments that measure/evaluate/predict (e.g., propensity to engage in negative behaviors, teamwork, trainability). Also perceived missing are measures of factors such as fluid intelligence, written communication, situational judgment, and problem solving. Suggestions regarding what could be eliminated from the current ASVAB included GS, EI, and AO (they are of less use for some Services) and AI (because of adverse impact and content). Other suggested improvements included expanding unproctored testing, using AIG to reduce item development time, and combining subtests to reduce administration time.

The primary psychometric reasons offered for changing the ASVAB included (a) increasing incremental validity over AFQT, (b) increasing classification efficiency, (c) increasing differential validity across job types, (d) increasing diversity, and (e) making the battery more resistant to compromise. In regard to content, it was noted that the training needs and nature of the work may have changed, and, over time, there is a need to assess the "whole person," as well as to better identify people who fit the military culture. Regarding administration, it was suggested it would be valuable to reduce testing time and take advantage of technological advances in both information technology and measurement. Changes could also help in addressing the perception that the battery does not change with time and is not measuring the right things. Changes are occurring elsewhere and the lack of change vis a vis the ASVAB is viewed as hurting the program from a leadership perspective. Modernization efforts could also increase the face validity of the test in the eyes of decision makers. Finally, changing the ASVAB could increase the number of eligible applicants and result in changes to policies that meet the requirements for the total workforce.

Reasons offered for not changing the battery include the research and non-research costs involved (e.g., renorming, information technology changes), the fact that the current test still meets the need at hand, a lack of consensus about what changes to make, and the need to maintain comparability with the CEP. Possible mitigations to these barriers include utility analyses to demonstrate the return on investment,

> frequent communication with stakeholders, obtaining stakeholder buy in, and eliminating existing misrepresentations regarding the battery.
>
> Stated goals to be achieved with a revised ASVAB/ETP include psychometric (e.g., increase prediction, differential prediction, classification efficiency), increasing the breadth of content coverage, reducing testing time and increasing flexibility to administer at home, improving face validity and perceptions of the test, and improving cost efficiency.
>
> In considering what a new ASVAB would look like, it was suggested the term “vocational” is outdated and should be eliminated. Additional suggestions were to include *g* in core tests used to determine enlistment eligibility, strike a better balance between crystallized and fluid intelligence, and include a measure of spatial and/or psychomotor ability. It was also suggested that non-cognitive measures should be included, some potentially as core tests, as well as interest measures; that the Cyber test should be included as a technical test; and that there should be physical/occupational assessment. Further, it was suggested that testing time should be about 90 minutes on average, the number of mandatory tests reduced, that new tests should be adaptive and administered on computers, and that tests should employ other item types rather than just multiple choice. Finally, it was suggested that ancillary information, such as biodata and parental educational attainment, should be collected.
>
> A variety of suggestions were offered for what the ASVAB should predict, including training success, on-the-job performance, attrition, first-term enlistment, person/job fit, physical fitness, leadership potential, and commitment. There was a three-way tie in regard to what outcomes the focus group thought the revision should focus on, with the highest priority being completion of training, job performance, and attrition. The next highest priorities were increasing classification efficiency and reducing adverse impact.
>
> Dr. Oppler then displayed a list of stakeholders from various realms (e.g., military/DoD, education, examinee) who could be targeted to take part in future focus groups. The next steps involve identifying SMEs to participate in focus groups, developing protocols, scheduling and conducting the groups, compiling the results, and forming and convening an ASVAB Stakeholder Advisory Committee to help guide decision making about the next generation of testing. DPAC will continue efforts to complete the ASVAB evaluation steps outlined earlier, as well as efforts to evaluate the special tests in a similar fashion. Work will also continue on establishing a methodological approach to rating the quality and expendability of ASVAB tests, and to develop a shared vision for next generation testing.

When Dr. Pommerich described the Step 11 synthesis findings (slide 70), a committee member commented on the amount and promising nature of the information provided. The committee member asked if Dr. Pommerich could help the committee determine what it should attend to in its feedback. Dr. Pommerich noted that Step 11 is still in progress, so she hopes to brief that in a future meeting. Then she said she would appreciate feedback on how to consolidate the results using a more rigorous approach, as well as on Dr. Oppler’s work with the focus groups. She said Drs. Oppler and Fechter would be briefing the focus group work momentarily.

Dr. Fechter asked for committee feedback on the next steps in the effort to consolidate ASVAB evaluation findings (slide 119). A committee member said collecting information on utility that will gain buy-in from stakeholders will depend largely on the composition of the panel. S/he asked what DPAC was thinking, as well as whether there might be multiple panels or a super-panel. Dr. Fechter said they were thinking of establishing a panel like the MAPWG, with representatives from each Service. The committee member replied that it might be helpful to use a layered strategy, starting with a MAPWG-type group, and then getting feedback on that group’s prioritizations. S/he said the process may blow-up if the final users of the results sense that they have been disenfranchised by the decision-making process. Dr. Fechter said that was a good point and that Dr. Oppler's presentation on focus groups would address that concern, because it will engage so many stakeholders.

Another committee member asked if the 5-point rating scales mentioned in relation to the Delphi method have been used previously in other contexts. Dr. Fechter said the Delphi method had been used to prioritize recommendations made by the ASVAB expert panel in 2005. She said that effort used a 5-point criterion weighting scale. The committee member asked if the scale included a midpoint representing a neutral category, because research, which is fairly consistent, suggests using a midpoint is not beneficial in decision making.

As Dr. Oppler briefed slide 136, a committee member asked if there was a specific question the committee should address. Dr. Oppler then proceeded to slide 141, which listed the stakeholders for the targeted focus group. He said, as the slide shows, there are multiple stakeholders from the military realm, as well as from the education realm. He explained that, not only did they have to worry about who was being brought into the military from the military's perspective, but how are they are being prepared on the educational side. He concluded by saying there are a lot of stakeholders, including applicants and Congress that must be considered. Referring to slide 142, he then described the plan for focus groups and asked for the committee's thoughts. A committee member said s/he was interested in hearing how the process works going forward, because it includes such broad coverage of different realms. The committee member also said s/he also was struck by the inclusion of state boards of education and asked if Dr. Oppler meant Federal or state boards, noting that Minnesota, and possibility other states, lack state boards, although Minnesota does have chief school officers and assistant officers, or commissioners, who should play an important role from the state perspective. Dr. Oppler said he appreciated the committee member bringing a greater degree of granularity to the table, which he said will be helpful in how they move forward in identifying participants. The committee member encouraged the inclusion of state leaders, whether it be state boards, commissioners, or school officers, because they represent such an important realm. The committee member also said including Congress would be very powerful and great for the program.

Commenting again on slide 142, a committee member made the point that aggregating information across indicators should consider that the outcomes of various focus groups may not be equally valuable. S/he said each indicator is valuable, but they are probably not all equally valuable. The committee member explained that, even after manual weighting, the effective weights are still partly determined by the amount of variability in each indicator. Accordingly, s/he said, it is important to pay attention to which indicators are affecting the composite, or the aggregated information. S/he said the indictors that possess more variance will affect the composite more than those with less variance, even though the latter indicators may be more valuable. The committee member stressed that some factors are "just really important," and they should not just be thrown in the mix. Dr. Fechter replied, with respect to the rating scales for the indicators for different subtests, they recognized that not all of them should be treated as equally important. She said, however, that the Delphi method should help determine the significance of each. She said she appreciated the committee member making the point about the effect of variability.

Another committee mentioned that DPAC might want to consider asking the question, if there was no ASVAB, what would we develop? S/he said the literature suggests three important factors:

ability, personality, and interest, and that it might be possible to cut the ability portion significantly and include a personality measure and an interest measure. S/he said that would be a clear break with the past. Dr. Oppler said the MAPWG focus group had discussed that idea. He said they asked the MAPWG what they would do if building a program from scratch, and what would it look like. Dr. Oppler then referred the committee to slides 136 and 137, explaining that there was both agreement and disagreement about issues such as, what should be core and what should be Service-specific. He said there was talk about personality and interest inventories in addition to cognitive ability. The committee member suggested that 90 minutes of testing time might be possible, as was shown on slide 137. Dr. Oppler then said the MAPWG focus group discussed a range of content, not just different flavors of cognitive ability. He also mentioned that the "what is missing" list included an even wider range of suggestions. Regarding cognitive ability, the committee member said that if DPAC really believes in *g*, it would certainly be possible to have a test much shorter than the AFQT.

A committee member closed the discussion by saying s/he could not believe the scope of the project and that it was excellent work. Dr. Velgach characterized the team's approach as thoughtful and deep and said the next step is to determine how and whether to proceed.

12. <u>Future Topics</u> (Tab N)

Dr. Dan Segall, Defense Personnel Assessment Center (DPAC), directed the discussion.

Dr. Segall presented a list of potential topics for future DAC meetings, as follows:

- ASVAB Resources
- ASVAB Development (pool development, evaluating/refining item and test development procedures, item writing guidelines and tools)
- Adverse Impact
- P*i*CAT/VTEST (Verification Test) Updates
- AFQT Prediction Test
- TAPAS Evaluation
- Test Security Compromise
- ASVAB Validity (improving the validation process and a review of Service validity studies, ASVAB validity framework, and criterion domain/performance metrics)
- Career Exploration Updates (web site, *i*CAT expansion)
- Adding New Cognitive Tests (Cyber Test, Working Memory, Complex Reasoning including Adverse Impact)
- Adding New Non-Cognitive Measures (personality and interest measures)
- Automatic Item Generation
- Web and Cloud efforts
- Device Evaluation and Expansion
- ASVAB Evaluation

Dr. Segall mentioned that most topics on the slide are recurring topics. He said cyber gaming had been slated for the present meeting but had been pushed back due to time restrictions. He also mentioned that a committee member had asked if there was anything new in the non-cognitive testing area. Dr. Velgach remarked that a committee member had asked for more information on the ASVAB Pending internet Computer Adaptive Test (P*i*CAT)-verification test interval. A committee member then expressed interest in an update on the Arithmetic Reasoning (AR) and

Math Knowledge (MK) automatic item generation, as well as on the next steps in the Device Evaluation Study (i.e., evaluating item features and their impact on item difficulty). Dr. Segall added DPAC's vision for implementation. A committee member asked what the committee could do to assist with the Next Generation ASVAB effort and said it might be helpful to focus on a few select areas. Dr. Segall said they can target some subtopics. A committee member said there should be an update on new forms development and equating, and Dr. Segall said he will see where they are with the results. Dr. Velgach relayed Ms. Miller's hope that they will be able to start talking about the Space Force, which will be new for the DACMPT. Dr. Segall then told the committee that, if they thought of anything else, they could include it in an email to Dr. Velgach.

13. Public Comments

Public comments opened with Dr. Carretta saying the Air Force has used Mturk in several studies related to their testing modernization effort and observed, in some of the more recent studies, a much higher level of bad data. He also mentioned that the Drasgow Consulting Group (DCG), in a study looking at item development for dark tetrads, reported a roughly 80% bad data rate. Dr. Carretta also said he saw some correspondence between Dr. Fritz Drasgow (DCG) and Dr. Paul Sackett (University of Minnesota), in which Dr. Sackett reported having much higher rates of bad data than in the past. He said Dr. Drasgow even found an article suggesting that this was a trend in Mturk, so he is looking at other data collection services, such as one called Prolific, which may do a better job of screening subjects. Dr. Caretta then recalled that someone had mentioned, in the AI discussion, moving beyond training performance, and said he has a project near completion that is looking at something called, "months of mission-ready service." He said this is a criterion that accounts for performance during the first term of enlistment and considers factors such as attrition and promotion rates. He said they would be presenting that work to the MAPWG.

There were no additional comments from the public.

In closing the meeting, Dr. Velgach said she would welcome feedback on this virtual instance of the DACMPT meeting and thanked everyone for their participation. The committee chair said he appreciated all the information, and that the presentations were nicely done and provided the committee a lot to think about.

Tab A

LIST OF ATTENDEES

Defense Advisory Committee on Military Personnel Testing (DACMPT)
September 17-18, 2020

Name	**Position**	**Organization**
Dr. Michael Rodriguez,	Chair Interim Dean, College of Education & Human Development	DACMPT, University of Minnesota
Dr. Neal Schmitt	Professor Emeritus	DACMPT, Michigan State University
Dr. Barbara S. Plake	Professor Emerita	DACMPT, University of Nebraska-Lincoln
Dr. Nancy Tippins	Owner and Manager	DACMPT, Nancy Tippins Group, LLC
Dr. Sofiya Velgach	Designated Federal Officer (attendance req'd by FACA)	Accession Policy Directorate
Ms. Stephanie Miller	Director	Accession Policy Directorate
Mr. Christopher Graves	Senior Scientist	Human Resources Research Organization (HumRRO)
Ms. Sachi Phillips	Project Manager	HumRRO
Dr. Daniel Segall	Director	Defense Personnel Assessment Center (DPAC)
Dr. Mary Pommerich	Deputy Director	DPAC
Dr. Shannon Salyer	Manager, Career Exploration Center	DPAC
Dr. Tia Fechter	Personnel Research Psychologist	DPAC
Dr. Greg Manley	Personnel Research Psychologist	DPAC
Ms. Olga Fridman	Analyst	DPAC
Dr. Lihua Yao	Principal Mathematical Statistician	DPAC
MAJ Kevin Doherty	Test Control Officer in Charge	US Marine Corps
Dr. Donna Duellberg	Voluntary Education Program Manager	US Coast Guard

Mr. Charles Lamer	Accessions and Recruiting	US Space Force
Dr. Tonia Heffner	Chief, Selection and Assignment Research Unit	US Army Research Institute (ARI)
Dr. Cristina Kirkendall	Research Psychologist	ARI
Dr. Erin O'Brien	Research Psychologist	ARI
Dr. Alisha Ness	Research Psychologist	ARI
Dr. Colin Omori	Postdoctoral Research Fellow	ARI
Dr. Michelle Kim	Postdoctoral Research Fellow	ARI
Dr. Jackie Torres	Postdoctoral Research Fellow	Consortium Research Fellows Program
Dr. Mark Rose	Chief, Market Research & Analysis	US Air Force Recruiting Service
Dr. Tom Carretta	Senior Research Psychologist	US Air Force
Dr. Sophie Romay	Senior Personnel Research Psychologist	Air Force Personnel Center
Dr. Bobbie Dirr	Personnel Research Psychologist	Air Force Personnel Center
Mr. Ken Schwartz	Chief, Testing and Survey Policy	Air Force Personnel Policy
Mr. Robert Tiegs	Testing Director	US Military Entrance Processing Command (USMEPCOM)
Mr. Billy Crook	Command TCO/Management Analyst	USMEPCOM
Mr. David Davis	Chief, Testing Division	USMEPCOM
Mr. Jaime Clayton	Enlistment Testing Program Manager	USMEPCOM
Mr. Stephen Richardt		USMEPCOM
Dr. Cyrus Foroughi	Engineering Research Psychologist	US Naval Research Laboratory
Ms. Odeyra Curcic	Senior Operations Policy Analyst	Defense Counterintelligence and Security Agency
Mr. Tom Blanco	Vice President	S&T Consulting

Dr. Tim McGonigle	Program Manager	Human Resources Research Organization (HumRRO)
Dr. Peter Ramsberger	Program Manager	HumRRO
Dr. Matthew Trippe	Senior Staff Scientist	HumRRO
Dr. Scott Oppler	Principal Scientist	HumRRO
Dr. Laura Ford	Program Manager	HumRRO
Dr. Deirdre Knapp	Principal Scientist	HumRRO
Dr. Ping Yin	Senior Staff Scientist	HumRRO
Dr. Furong Gao	Senior Staff Scientist	HumRRO
Ms. Deanna Hudella	Executive Director, Federal Programs	Pearson Vue
Ms. Kristy Park		

Tab B

DEFENSE ADVISORY COMMITTEE ON MILITARY PERSONNEL TESTING
AGENDA

September 17-18, 2020
Virtual – Microsoft Teams
Link – Click Here

Dial-In: 1 571-429-6145, Conference-ID: 160 597 151#

Meeting details will be posted on: https://dacmpt.com

September 17, 2020 (Eastern Time)

1200-1215	Welcome and Opening Remarks	Dr. Sofiya Velgach OASD (M&RA)/AP*
1215-1230	Accession Policy Update	Ms. Stephanie Miller OASD(M&RA)/Director, AP*
1230-1300	Milestones and Project Schedules	Dr. Mary Pommerich DPAC/OPA*
1300-1315	*Break*	
1315-1415	Device Evaluation Update and Future Use	Dr. Tia Fechter / Dr. Dan Segall OAP/DPAC*
1415-1445	ASVAB CEP* Update	Dr. Shannon Salyer OAP/DPAC*
1445-1500	*Break*	
1500-1545	CAT – ASVAB* New Forms Update	Dr. Matt Trippe HumRRO*
1545-1645	TAPAS* Evaluation Project Overview	Dr. Tim McGonigle HumRRO*
1645-1700	*Public Comments*	

<u>September 18, 2020 (Eastern Time)</u>

1200-1245	Use of Calculators	Dr. Peter Ramsberger HumRRO*
1245-1315	Complex Reasoning	Dr. Scott Oppler HumRRO*
1315-1330	*Break*	
1330-1430	Adverse Impact	Dr. Greg Manley OPA/DPAC*
1430-1600	Testing – Next Generation (Multi-Dimensional	Dr. Mary Pommerich/Dr. Tia Fechter OPA/DPAC* Dr. Scott Oppler HumRRO*
1600-1615	*Break*	
		Navy Selection and Classification
1615-1630	Future Topics	Dr. Dan Segall DPAC/OPA*
1630-1645	*Public Comments*	
1645-1700	Closing Comments	Dr. Michael Rodriguez, Chair

*** KEY:**

ASVAB = Armed Services Vocational Aptitude Battery
ASVAB CEP = ASVAB Career Exploration Program, provided free to high schools nation-wide to help students develop career exploration skills and used by recruiters to identify potential applicants for enlistment
CAT = Computer Adaptive Testing
HumRRO = Human Resources Research Organization
OASD(M&RA)/AP = Office of the Assistant Secretary of Defense (Manpower & Reserve Affairs)/Accession Policy
OPA/DPAC = Office of People Analytics/Defense Personnel Assessment Center
TAPAS = Tailored Adaptive Personality Assessment System

Tab C

UNIVERSITY OF MINNESOTA

Twin Cities Campus

Office of the Dean
College of Education and Human Development

104 Burton Hall
178 Pillsbury Drive S.E.
Minneapolis, MN 55455

Phone: 612-626-9252
Fax: 612-626-7496
http://cehd.umn.edu

October 22, 2020

Ms. Stephanie Miller
Director, Accession Policy
Pentagon, Washington DC, 20301

Dear Ms. Miller:

The Defense Advisory Committee on Personnel Testing (DACMPT) is pleased to provide this committee report of our meeting of September 17-18, 2020, conducted virtually through Microsoft Teams. Below, we provide summaries and recommendations from the DACMPT emanating from that meeting. The virtual meeting was productive and nicely facilitated. Connectivity was consistent and supported each presentation and DACMPT interaction with presenters and among committee members. We are also happy to see the development of the DACMPT website, https://dacmpt.com, as it is an important resource for DACMPT members current and future, as well as those affiliated with the defense offices interested in military personnel testing.

The meeting began with opening remarks from Dr. Sofiya Velgach and Dr. Rodriguez (chair). We acknowledged the passing of Dr. Keven Sweeney on September 3, a member of the DACMPT. Also, Drs. Barbara Plake, Neal Schmitt, and Nancy Tippins were in attendance. In addition, staff and representatives from Defense Personnel Assessment Center (DPAC) and various military units were present.

The DACMPT report and recommendations follow, in the order of the meeting agenda.

Accession Policy Update

Ms. Stephanie Miller, OASD(M&RA)/AP, provided a brief introduction to the accession policy updates. She began by restating the mission of the Manpower and Reserve Affairs (M&RA) unit and presenting organization charts that encompassed the unit. Ms. Miller noted that there had been wholesale turnover in the unit over the summer, and depending on the outcome of the November elections, additional changes may occur. Although the mission of the M&RA has remained constant, the means of achieving recruiting goals has changed due to current unemployment rates and economic changes. She reviewed initiatives in eight areas (USMEPCOM, GI Bills, Accessions, Testing, Officer Programs, Personnel Security, Transgender/Gender Dysphoria, and Miscellaneous), noting that some of the plans would be discussed in more detail later in the meeting. Ms. Miller closed by updating the DACMPT on the recruiting goals of each of the Services to date and the Congressional/Internal reports (i.e., Report on the ASVAB, Report on Assessing the English Learner, Report on Diversity and Inclusion) that have been submitted or will be completed soon.

Recommendations
In future meetings, the DACMPT would like to hear more about assessments in the form of gaming and measures of interest.

Project Milestones

Dr. Mary Pommerich, OPA/DPAC, reviewed the major ASVAB development projects, including the development of new item pools for the ASVAB, new CAT item pools for the CEP, the automated generation of arithmetic reasoning (AR), mechanical knowledge (MK), and general science (GS) items, the Career Exploration Program, the ASVAB and ETP revision, and the Defense Language Aptitude Battery. The develop of new CAT-ASVAB item pools and the CEP were discussed in more detail later in the agenda. A number of the project deadlines have slipped because of the freeze associated with moving many operations to the cloud. The transfer to the cloud has had more consequences on the ability to stay on track.

Device Evaluation

Dr. Tia Fechter provided an update on the research effort to examine whether allowing examinees to use mobile devices would increase access to the ASVAB and reduce the time needed for administration. Data from over 8500 examinees (recruits in the USMECOM settings and in training locations) were used to examine if test performance and time for test completion were affected by taking the tests on mobile devices. The seven devices included in the study involved both tablets and smart phones using different operating systems. The control device was the Dell XPS as that is the device used for the operational assessment. Using a counterbalanced design, the researchers asked the examinees to take the assessments either first on the control device and then again on their assigned device or vice versa. Test takers were asked whether they were familiar with the assigned mobile device. Analyses were conducted at the subtest level as each examinee took only a subset of the ASVAB tests. Results indicated no performance difference for taking the test on the mobile devices as long as the examinee was familiar with the device he/she was assigned. Less time was needed when the assessments were delivered on the mobile devices compared to the control device, but the time differences between the mobile devices and between the mobile devices and the control device did not reach practical significance (as not more than 30 seconds difference was seen in the time for completion across the control and mobile device conditions). The conclusion drawn was that if the test was designed with mobile device delivery in mind, and if the examinee is familiar with the mobile device, comparable performance would result from test delivery on the standard (control) device and appropriate mobile devices.

Next Dr. Dan Segall provided a presentation related to policy decisions regarding alternatives and for impacts of implementing an option for mobile device delivery for ASVAB tests. Among the operational implementation decisions to be made are (a) to whom would such a policy apply (e.g., applicants in MEPS, examinees in unproctored settings; students in schools taking the CEP); (b) which devices would be allowed (e.g., examinees' own device, devices owned and maintained by DoD, devices provided in school settings); and (c) for what purposes (e.g., proctored or unproctored applications). Concerns were raised about implications for test security (due to capacity for screen shots on mobile devices), equipment maintenance if owned by DoD, examinee choice across available devices which might not be well-informed with respect to test performance, and impact on test performance if examinees used an unfamiliar device.

The DACMPT concurred with the concerns articulated by Dr. Segall regarding operational implementation of a mobile device testing policy. Additional concerns were raised about possible differences in device familiarity across economically disadvantaged groups. If device choice was a factor in the application of the policy, then again concerns were raised about whether examinees would make optimal choices. The DACMPT was more comfortable with implementing this policy in unproctored settings with examinee-owned devices. However, there would still be concerns about screen shots impacting test security. In settings for high stakes decisions such as qualification for service, an additional concern for examinee-owned devices is whether the device had software that would compromise the integrity of test performance. If a decision is made to allow for mobile device use, monitoring should be done to examine the consequences, determine which devices are being used, and identify any perceived barriers to examinee performance. If the decision is to permit examinee-owned mobile devices in unproctored applications, the goal of reducing time for administration of the ASVAB in MEPS would not be realized.

Recommendations

The DACMPT recommends that if a decision is made to permit test administrations on mobile devices, examinee-owned mobile devices be used only in low stake testing or unproctored administrations. Further, mobile device use should be monitored, documenting what devices are being used and identifying any issues or concerns.

ASVAB Career Education Program

Dr. Shannon Salyer provided the briefing on the ASVAB Career Education Program update. In the presentation, the DACMPT year-to-date reviewed usage metrics on a number of CEP components. Participation, use of scores, requests for score reports, and interpretation support via websites were all increasing. This increasing usage is promising not only for users who explore military or military-related careers, but for the potential identification of future enlistees. Most notably, access and exploration of Careersinthemilitary.com was also increasing.

A long-standing goal has been to encourage additional engagement of school counselors, both in K-12 and higher education advisors. The provision of CEUs for counselors is making progress and seen as a key incentive to promote greater use of ASVAB CEP tools and resources.

The response to COVID-19 constraints was reviewed. Some actions appear promising, including the implementation of virtual post-test interpretations and combining schools in these efforts.

Recommendations

The DACMPT endorses and encourages the CEP to explore other information obtained from usage statistics to share with the military branches, beyond scores. This may provide additional leads as well as indications of interests among high school students. The DACMPT also encourages further examination of how COVID-19 modifications may provide greater access to score interpretations to larger audiences and across schools.

Another area to monitor is the occasional state legislation that introduces opportunities for the military to engage public school students regarding career options, including the adoption of ASVAB CEP for career planning for students, as states are requiring college and career planning for all high school students, and some for students in middle schools.

CAT-ASVAB New Forms

Dr. Matt Trippe updated the DACMPT on HumRRO's efforts to generate sets of item pools that would be used for Forms 11-15 of the ASVAB assessments. Starting with 1000 items per assessment, the research team reviewed items for psychometric quality, coded items as enemies when conflicting items existed in the set, and subjected items within a pool to a forms assembly algorithm to create five pools of items that had the capacity to maximize conditional precision across the examinee ability distribution while maintaining comparability in precision across the pools (among other test assembly criteria). The goals of the pool assembly process were to be responsive to the unique item selection algorithm from the CAT, produce comparable scores across pools, assess the full range of ability and subtest content, and provide score precision across the ability range. The results indicated that these goals were achieved for most of the assessments, but some problems were indicated in the results for Arithmetic Reasoning (AR) and Mechanical Reasoning (MR). For AR, the information function for the new pools should have somewhat lower precision than the previous set of pools 5-9; however, the impact was only a small decrement in reliability (.02). The reason for this loss of precision appears to be because the items developed for pools 11-15 had, in general, lower a-parameters (i.e., lower discrimination) than the items used for pools 5-9. For MK, the problems in lower precision appears to be a result of the new items being noticeably harder (higher b-parameters) than the items developed for pools 5-9. Again, the impact of lower precision for pools 11-15 was lower reliability values (.01-.03). Plans are in place to instruct item developers for MK to generate additional items that are less difficult.

Comments
The DACMPT expressed concerns that the higher difficulty of both the AR and MK items may affect the cutoff scores for AFQT; however, it is too early to know the overall impact as these analyses were done only at the individual assessment level and not for composites such as the AFQT. Further, these pools will need to be equated to be on the same scale as pools 5-9.

Recommendations
As new items are being developed for ASVAB assessments, attention should be paid to the item characteristics, in particular discrimination and difficulty, to help maintain psychometric equivalence across years and forms. When the equating of pools 11-15 to previous pools 5-9 has been achieved, consideration of the impact on composites, especially for AFTQ, should be examined.

TAPAS Evaluation Project

Dr. Tim McGonigle of HumRRO presented a summary of the report by a special team of external scholars in the personality and selection area of the TAPAS and research involving the TAPAS. The review was organized around eight areas addressed in the *Standards for Educational and Psychological Testing (AERA, APA, NCME,* 2014). In each area, the authors identified evidence on the TAPAS that was satisfactory, minimally sufficient, and insufficient to justify test use for selection. The DACMPT was very impressed with the report as well as plans for addressing the areas that were judged to be minimally sufficient or insufficient. The expert team made four pre-implementation recommendations regarding TAPAS scales, scores, and norms. The DACMPT believes all four are important, but we also think there must be some effort to make TAPAS scales and scores comparable across Services as is stated elsewhere in the report and summary. In one version of the TAPAS, there are 27 facets. Although no one Service appears to be using all of these facets in their composites, there does not seem to be much congruence in their use and interpretation across Services.

Recommendations
The Services should attempt to come up with a relatively small number of facets (perhaps five to ten, a subset of the current facets) that would be administered to all military recruits. Reliability and norms should be estimated for the Services as a whole. Each Service would then develop support for their unique use (validity and cut scores) in their context or ideally across Services. In discussing reliability, the expert team noted that test-retest and alternate-forms reliability were presented for various versions of the TAPAS, but these classical forms of reliability are inappropriate with tests such as the TAPAS that are developed using an IRT model. Although the DACMPT supports the recommendation that more appropriate indices be developed, it is likely that data regarding marginal reliability and conditional standard errors will support test use given favorable classical indices.

The expert team also recommended that validity arguments be developed for each use and outcome of the TAPAS. The DACMPT would underscore the need for a comprehensive centralized technical infrastructure for summarizing all validation efforts as is currently reflected in the *way forward* section on validity. The expert team also recommended that such arguments be developed for the composites used by the different Services. It was noted that developing construct validity arguments for TAPAS composites will be especially difficult.

The DACMPT recommends that in assessing construct validity arguments the following be considered: (a) conceptual arguments for each facet included in a composite (as well as its weighting in the composite) and a rationale as to why that facet be included, (b) previous research on similar facets in the military or elsewhere, and (c) intercorrelations between the facets and with the composite itself.

The DACMPT recommends that plans of action or way-forward tasks should include efforts to assess the impact of test use (utility) on military recruitment and selection efforts. For example, the .01 or .02 incremental validity for the use of TAPAS in the prediction of attrition can be translated into the number of persons whose attrition is avoided, and the potential cost of recruiting and training soldiers who leave can be calculated. If the Tapas is to be used to assess attitudes, it may be possible to relate scores to some distal behavioral outcome such as reenlistment.

Finally, an estimate of the level of impact that is deemed to be important could be considered as part of the Theory of Action. Such specification should lead to better and more rapid decision-making regarding the use of the test. In connection with various plans for additional pre or post implementation research efforts, the results that would be considered sufficient to proceed with use of the test should be specified (e.g., incremental validity of .01 or level of subgroup difference that is intolerable, magnitude of conditional standard errors).

Overall, the DACMPT was favorably impressed with the expert team's report and the plans developed to address the pre- and post-implementation recommendations contained in that report. If completed and the results are satisfactory, the TAPAS should be a valuable tool in military selection.

Use of Calculators

Dr. Peter Ramsberger reported on HumRRO's project to examine the impact of the use of calculators when taking the ASVAB assessments, especially Arithmetic Reasoning and Mathematical Knowledge. In this project, the researchers reviewed the literature on the effects of test anxiety on test performance (with the expectation that allowing calculators would reduce test taker anxiety), evaluated the use of calculators for other standardized assessments, obtained insight regarding the use of calculators from recruiters and applicants via focus groups, and assessed the importance of math skills and hand calculations for success in military training and on the job. They found mixed results on the impact of test anxiety on math performance with limited evidence that allowing calculator use had an overall effect of lowering test anxiety. They found that most large-scale standardized tests, such as the ACT, SAT, NAEP, and GED allow calculators at least for some math subtests that are deemed "calculator active/needed/useful." Importantly, based on an online survey of subject-matter experts across a variety of career fields in the military, hand-calculations (without calculators) were reported as important both in training and in military jobs. Further, adopting a calculator use policy on the ASVAB would have significant operational and psychometric consequences. Such a decision would require many time consuming and costly changes, including the need to:

- Develop specifications for new item development
- Review existing items
- Field test new items
- Conduct test scaling and equating
- Develop norms
- Evaluate test fairness
- Evaluate and establish new testing times
- Develop and implement applicable software updates
- Evaluate test reliability and validity

Resource projections for making this policy change are in the range of 10 years and 30 million dollars.

Comments

The DACMPT agrees that the most compelling reason for not allowing calculators when taking the ASVAB is the need to align test conditions with job requirements. Because the survey found that being able to do hand-calculations was important both for training and job performance, it should be part of the construct that is measured on the ASVAB in order for the ASVAB is to predict performance in training and military jobs. This a fundamental difference between the ASVAB and the other standardized tests that were found to allow the use of calculators. In the latter instance, tests such as NAEP, ACT, SAT, and GED are designed to either measure achievement in mathematics (NAEP) or to predict success in higher education (ACT, SAT, GED) in programs that do allow the use of calculators for some mathematical tasks.

The DACMPT raised other questions regarding the policy for allowing calculators, including what type of calculator would be allowed, who would provide the calculator, whether it would be desirable to have a calculator embedded in the test delivery software (which would create non-comparable testing environments for CEP with ASVAB P&P administration). However, these considerations pale when placed aside the fact that in order for the ASVAB to result in valid interpretation for its intended purposes, it should not incorporate the use of calculators. Based on current information about the constructs measured by the test and the job and training requirements, the DACMPT believes that calculators should not be permitted. If there comes a time when doing hand-calculations is not an important skill for training and job performance, reconsideration of this policy might then be appropriate.

Recommendations

Calculators should not be allowed with ASVAB assessments because the scores and composites based on ASVAB assessments are used to predict success in training and job performance that may involve the use of hand computations. More detailed surveys regarding the role of hand computations in training and job performance could further clarify to what extent hand calculations are a component of training and jobs in the military.

Complex Reasoning

Dr. Scott Oppler, HumRRO, presented his firm’s work on a measure of Complex Reasoning. The intent of the test is to create a measure of fluid intelligence that will better balance ASVAB measures of crystalized intelligence, increase the prediction of training and job success, minimize susceptibility to compromise due to the non-verbal items, and increase the qualification rates of non-native and non-heritage English speakers. The research plan for the measure of complex reasoning is intended to identify ways to improve item development efficiency while reducing or eliminating field-testing requirements through an item generation tool and determine the features of items that affect the difficulty levels. Stage 1 is complete, in which prior attempts to model parameters of matrix reasoning items, the Abstract Reasoning Test evaluation study, the item generation capabilities of the SGMT are examined and the relevant research questions are identified. Stage 2 is in progress, including the design and implementation of the pilot study which uses an MTurk sample. Stage 3 will follow the completion of Stage 2, including the design and implementation of a full operational study.

Recommendations
The DAC has few recommendations at this point in the research plan; however, there was discussion about reducing the verbal load of the instructions to enable those whose first language is not English to understand the test requirements easily. The DAC is enthusiastic about this study and is hopeful that the results contribute to a useful measure of fluid intelligence.

Adverse Impact

Dr. Gregory Manley of OPA/DPAC presented the results of adverse impact and effect size analyses for racial/ethnic and gender groups defined by the US Census categories for these variables. These data included differences in selection rates at two levels of AFQT cut scores (≥31 and ≥50) and standardized mean differences on the various ASVAB tests. Also included were adverse impact ratios for different subgroups over biennial analyses since 2019 and comparisons of impact ratios for new special tests (Cyber, Mental Counters, and Coding Speed) and with two other major tests (NAEP and SAT) administered nationally and across different subject matter areas. These data are largely as one would expect given the extant data in other nonmilitary arenas. Except for non-Hispanic Blacks, there is little evidence of adverse impact when the lower IIIB cutoff score is used. However, at the higher cutoff score used for IIIA, there is evidence of impact against non-Hispanic Blacks and Hispanic Whites with little adverse impact for Females and Non-Hispanic Asians. Impact ratios and effect sizes are consistent over time. There are some subject matter differences: Asians seem to perform less well on verbal tests, and impact ratios and effect sizes for females are largest in the Math and technical areas. Because of the relatively poor performance of the Black recruits, it might be helpful to unpack these differences further.

Recommendation
The DACMPT suggests examining potential regional differences in impact ratios that might reveal larger or smaller differences or if other factors indicate where larger or smaller adverse impact ratios are found. It might also be useful to see what the impact on Black candidates means in terms of the MOS assignments they receive upon entry into the military. The presence of subgroup differences in scores does not mean the tests are biased. To investigate bias, an attempt is usually made to assess whether the performance of different groups is predicted by the same regression equation. These examinations are usually conducted using a moderated regression analysis. To conduct these tests, one must have an outcome measure that is itself unbiased and on which there is variability. With the exception of the Air Force where training grades are available, performance in training is simply a pass-fail decision, and very few recruits fail their training program. Good performance post-training data, again with the seeming exception of the Air Force, are not typically available.

The DACMPT recommends that the possibility of biased prediction be investigated with Air Force training and performance data. In the absence of similar data from the other Services, we recommend that other potential outcome data be considered for these analyses where appropriate (e.g., promotions, bonuses, reenlistment). As a general recommendation, we encourage future reports of adverse impact to draw stronger attention to effect size indices rather than statistical significance – for example, some findings in the briefing were highlighted due to statistically significant z-tests, although the effects were quite small.

Testing—Next Generation

Drs. Mary Pommerich, Tia Fechter, and Scott Oppler presented a set of briefings on the broad topic of the next generation of tests in the ASVAB and beyond. Over the years, numerous changes have been made to the ASVAB, updating many features; however, no changes to content have been made in recent times. The Next Generation Testing efforts are intended to promote opportunities for changing content. This includes consideration of the ASVAB and special tests – all of which are administered on the ASVAB platform. An important aspect of this effort is the development of a validity framework for the ASVAB and AFQT in particular.

A useful component of this work is the comprehensive nature of the ASVAB evaluation plan, including possible synthesis approaches using DELPHI methods regarding adding, dropping, combining or shortening subtests. This addresses cross-impact analysis, utility analysis, and cost analysis.

The makeup of the ratings panel will be important, and stakeholder representation will provide a strong basis for defending results. MAPWG is one vehicle to ensure representation.

Recommendations
The DACMPT is highly supportive of the Next Generation efforts, particularly as it grounds and supports the validity framework. It will be important to find ways to broaden input at various stages to ensure broader buy-in and engagement. Focus groups will help broaden participation in the process overall.

Focus groups produced interesting results regarding likes/dislikes and possible candidate characteristics and abilities missing. Regarding focus group arenas of stakeholders, consider state school officers (commissioners/assistant commissioners). Not all states of state boards of education – that particular corner of the education arena might be better conceptualized as state school leaders.

The DACMPT encourages looking for ways to attend to indicators that have value in themselves – as you aggregate information across indicators and sources and focus groups

As a final thought, the DACMP encourages the Next Generation team to consider erasing all history of the ASVAB. What would it look like if you start with a clean slate? Maybe this is a thought experiment, but it is worth considering the ideal and what would be core considerations for inclusion in the ASVAB. We encourage the ASVAB team to consider finding time to seek additional input and support from the DACMPT in this effort in future meetings.

Future Topics

Dr. Dan Segall, OPA Director, DPAC, closed the meeting with a quick summary of future topics. Many of the future topics, such as ASVAB resources, development, evaluation, adverse impact, and validity; CEP; TAPAS; web and cloud efforts were recurring areas of interest for the DAC. In future meetings, the DAC expressed interest in hearing more about automatic item generation for arithmetic reasoning (AR) and mechanical knowledge (MK), plans for development of new forms and equating, cyber gaming applications, the measurement of interests, and assessment for the Space Force.

It is clear that full funding of the ASVAB and associated programs has helped secure greater progress on a number of key activities. The DACMPT encourages the federal government to maintain full funding for the ASVAB program and associated projects, to maintain a high level of security and quality, which meets the DoD goals for accession, training, and force readiness. Overall, the meeting was informative and useful. The DACMPT appreciates the high quality efforts of Accession Policy and DPAC staff, and the research staff of each of the Services. Their frank interactions with the committee continue to be helpful and appreciated. As always, the DACMPT is interested in supporting these efforts, as it provides a strong basis for the defense of the intended interpretations and uses of the ASVAB as well as its future. We look forward to our next meeting.

Sincerely,

Michael C. Rodriguez, Ph.D.
Interim Dean, College of Education & Human Development
Chair, Defense Advisory Committee on Military Personnel Testing