DEFENSE ADVISORY COMMITTEE ON MILITARY PERSONNEL TESTING

August 16-17, 2023 Meeting

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

Minutes approved for public release.

Nancy L. Tippins November 15, 2023

Dr. Nancy Tippins, Chair, DACMPT **DATE**

DEFENSE ADVISORY COMMITTEE
ON
MILITARY PERSONNEL TESTING

August 16-17, 2023

The Fiscal Year (FY) 2023 second session of the Defense Advisory Committee on Military Personnel Testing (DACMPT) was held at the Crowne Plaza Chicago O'Hare Hotel & Conference Center, Chicago, IL on August 16-17, 2023. The meeting was conducted in person; however, one DACMPT committee member and two presenters participated virtually using the Microsoft® Teams online collaboration tool. Dr. Sofiya Velgach (Assistant Director, Office of Accession Policy [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 all other governing Federal statutes and regulations, and open to the public. She said the meeting agenda was available on the DACMPT website[1]and public comments would be received at the end of each day's scheduled sessions.

Dr. Velgach thanked the committee members for their participation and the presenters for their support of the committee's activities. She then introduced the Director of AP, Dr. Katherine Helland. Addressing the administrative components of the virtual meeting, Dr. Velgach said she needed a complete record of attendance and distributed an attendance sheet. She also informed participants that the meeting was *not* being recorded on the Microsoft Teams® system. She instructed all Teams participants to mute their devices and to click the "raise hand" button when they wanted to speak. She then directed introductions of all participants.

The attendee list and agenda are provided in **Tab A** and **Tab B**, respectively. **Tab C** contains a list of acronyms. The Committee Chair has provided a letter, written by the committee members, summarizing key committee findings. The letter is included in these minutes at **Tab D**.

1. Accession Policy Brief (Tab E)

Dr. Katherine Helland, Director, AP, presented the briefing.

> Dr. Helland began by presenting an organizational chart for the Office of Accession Policy. She continued by citing a number of challenges in the current recruiting environment including (a) lingering effects of Coronavirus Disease 2019 (COVID-19), (b) minimal support from influencers to recommend military service, (c) low youth propensity to serve, (d) a limited pool of qualified youth, (e) the desire to maintain a highly qualified and diverse force and (f) maintaining adequate recruiting resources. Dr. Helland noted mitigating factors include: the existing professional and dedicated recruiting force, national support for a strong military, and robust virtual and social media engagement.
>
> Dr. Helland presented a chart showing results from surveys assessing the propensity to serve, which were administered from April of 2001 through Spring of 2022. These demonstrate that few youths are propensed to serve and propensity continues to decline, meaning the Services have to work harder to meet their mission. She then turned to the 2020 Qualified Military Available (QMA) Study. Key findings included (a)

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

the proportion of youth eligible for military service without a waiver was 23%, which is a decrease from previous estimates of 29%; (b) most ineligible youth are disqualified for multiple reasons (44%); (c) the largest increases in disqualification estimates observed between 2013 and 2020 were for mental health reasons and overweight conditions; (d) when considering youth disqualified for one reason alone, the most prevalent reasons were overweight (11%), drug use (8%), and medical/physical health (7%); and (e) the proportion of youth who are QMA, defined as both eligible and not currently enrolled in college, is 12%.

Dr. Helland then reviewed recruiting results for Fiscal Year (FY) 2023. These showed that for the Active Component only the Marine Corps and Space Force have met 100 percent of goal, while the Army, Navy, and Air Force are below 90% of goal. Furthermore, the Marine Corps, Air Force, and Space Force have met quality benchmarks through the end of May FY 2023, although these numbers fluctuate during the course of the year. Among Reserve Component forces only the Marine Corps Reserves have met goal. The Army National Guard has achieved 90-99 percent of goal, while the other components are all below 90 percent of goal. The Marine Corps Reserve and Air National Guard have met or exceeded DoD quality benchmarks through May 2023.

Dr. Helland then turned to actions taken to address these issues. These include actions related to growing propensity: (a) increasing OSD and Service marketing and advertising to grow propensity; (b) senior leaders developing relationships with civic, ethnic, and business organizations; (c) collaborating with other Federal agencies, universities, and other influential community organizations; (d) developing a recruiting toolkit that highlights the benefits of military service, and (e) rebuilding relationships with high schools and key influencers. Another focus is on expanding eligibility by (a) reducing time limitations on disqualifying medical conditions, (b) establishing a conditional Delayed Entry Program (DEP) for active duty applicants with specific disqualifying medical conditions awaiting adjudication waiver decisions, (c) reexamining OSD and Service policies to increase eligibility and reduce barriers to enlistment; (d) expanding the use of applicants scoring in the Category IV range on the Armed Forces Qualifying Test (AFQT), and (e) creating preparatory training programs to help individuals with low AFQT and/or physical fitness scores to qualify. Additional efforts focused on the process improvements include expanding processing opportunities (e.g., extended hours), establishing a Prescreening Coordination Cell to reduce the wait time for medical prescreens, and collaborating with the Services to identify and eliminate processing barriers.

Dr. Helland continued by providing more details on the Future Soldier/Sailor preparatory training programs cited earlier. The Academic Skills Development Track includes a combination of instructor-led and self-paced study to improve word knowledge, reading comprehension, arithmetic reasoning, and test taking skills. The operational Army program is mandatory for applicants scoring in the AFQT 21-30 range and voluntary for those in the 32-49 range. The Navy plans to make the program mandatory for applicants in the AFQT 21-30 range. The Physical Fitness Track is aimed at improving the overall health of participants and prepare them physically and mentally for basic training. This will ultimately improve their health in the long term so they can successfully serve their country. The Army operational program is targeted towards applicants who exceed the accession body fat composition standard by greater than 2 percent but less than 6 percent. Applicants who exceed by up to 2 percent (based on gender, age, and height/weight) are able to ship directly to basic training assuming they meet all other applicable standards. The Navy program is directed at accessions who exceed the body fat standard by no more than 6 percent.

Dr. Helland then discussed testing efforts that are of interest to senior leaders and Congress. These include expanding the delivery of the Armed Services Vocational Aptitude Battery (ASVAB) to alternative devices, developing a Tailored Adaptive Personality Assessment System (TAPAS)-based joint enlistment composite, developing a TAPAS-based compatibility composite, and developing a new special purpose tests (i.e., Mental Counters [MCt] and Complex Reasoning [CR]). Congress required completion of a Pre-Enlistment Assistance Policy and Program Review with a specific concentration on programs designed to improve aptitude and physical fitness. This has been completed. Congress also mandated the creation of a Computational Thinking (CompT) test as an adjunct to the ASVAB, which is in progress.

At the end of the briefing, Dr. Helland noted the committee's expertise, especially in respect to information relevant to policy development. She then stressed the importance of protecting the

integrity of the ASVAB but noted that, if the Services have units or jobs that are not fully manned, then the readiness issue is cause for asking questions such as, why are applicants prohibited from using calculators when using them might increase accession rates. She said explaining validity is difficult in the face of readiness challenges. Dr. Helland then asked the committee in what areas could or should AP be willing to take risks? What should policy be in regard to how long test scores are valid? Should we maintain the 2-year window or extend it? She thanked Drs. Velgach and Pommerich for their papers explaining why they can and cannot do certain things in the testing realm. She said they are providing briefings on various subjects, including recommendations on types of assistance or preparatory programs that Services can use. They are anticipating Congress will provide their perspective on testing and quality related issues in the Fiscal Year (FY) 2024 National Defense Authorization Act (NDAA).

A committee member asked if AP had prep course outcome data. Dr. Velgach said the Army has had a program in place since July 2022, and it is seeing an increase in test scores, but data on performance is still limited at this time. There are plans for tracking outcomes long-term. Dr. Velgach also said the Navy will initiate an academic preparation course soon. The committee member asked about the length of course, and Dr. Velgach said it would be at least 3 weeks but could range between that and 90 days. She said the plan is that participants complete the 3-week course and, if they score high enough, they are sent to basic training. The committee member commented on the use of a mix of self-paced and instructor-led instruction and said it will be interesting to see how that plays out. Another committee member asked about the extent to which the prep courses are making a difference in whether Services are making mission. Dr. Velgach said, based on Army feedback, prep courses *are* making a difference to making recruiting mission. In respect to long-term objectives and goals, they see up to 20 points increase in scores, as compared to only 5 points improvement without the course (i.e., prior to the introduction of formal prep-courses). She said the targeted instruction is making a difference, but the driving question will be whether accessions completing the prep program perform in Service as well as those who entered without it (i.e., under regular testing practices). Approximately 6,400 persons have graduated from the Army's academic skills development program. A committee member asked about the physical fitness prep-course track, and Dr. Velgach responded that 97-98% are able to make the qualifying physical fitness standards after the course. She said the Navy and Army access people who are up to 6% over bodyfat standards. She said they are looking at behavioral and nutrition habits as well; that is, can they be instilled long-term or not? In terms of academic skills development, Dr. Helland said there is a policy that the Services cannot coach to the test, such that skill development remains the focus.

A committee member had questions regarding sexual assault and impact on recruiting. Dr. Helland responded they had seen an increase in female youth reporting sexual assault as a reason not to join, but it was still not the top reason. The top reasons were risk of physical and emotional injury and leaving family and friends.

Additionally, there was a question regarding accommodations for Non-native English Speakers (NNES). Dr. Velgach commented on additional time or accommodations made for NNES test takers. In the Career Exploration Program (CEP), students are allowed the same accommodations as those allowed by schools for other tests. She said that, when students want to enlist, they must take or retake the test under normal testing practices. She said test takers cannot have additional

support, which is the general policy for testing. Another committee member asked about possibilities for working with the Attention Deficit Hyperactivity Disorder (ADHD) population, noting that population is increasing. Dr. Helland said they would have to look at the risks the Services are willing to take. She said they are looking at accession standards and retention standards, pulling data to look at advances in medical practices and waiver data to evaluate impact on performance and attrition. She said "general prevalence" of a condition would be a factor and stressed the importance of making recommendations based on empirical evidence to senior leadership. She said they are beginning to look at mental health conditions and waivers, because this is one of their biggest concerns due to increases in mental/behavioral health issues in the general population.

2. R&D Milestones Brief – (Tab F)

Dr. Mary Pommerich, Director, DTAC, presented the briefing.

Dr. Pommerich began the presentation with an overview of the projects to be covered in the briefing, including ASVAB development and ASVAB and Enlistment Testing Program (ETP) revision.

- ASVAB and ETP Revision: Evaluating new cognitive tests/composites for the ASVAB including CR, CompT, the Cyber Test, and MCt. Adding non-cognitive measures for selection and/or classification by creating a TAPAS validity framework and joint-Service TAPAS. Other work is focused on the Career Exploration Program (CEP), a military compatibility assessment, and expanding test availability (e.g., web/cloud delivery of ASVAB and special tests and device expansion).
- Ongoing efforts to develop new items for the ASVAB.
- New Computer Adaptive Testing (CAT-ASVAB Item Pools). The objective of this project was to develop CAT-ASVAB item pools 11 – 15 from new items. These forms were implemented in the summer of 2023.
- Developing new paper-and-pencil (P&P) ASVAB forms 29F/G, 30 F/G, 32 F/G, and 32 F/G from new items. Project completion date is to be decided.
- Evaluation and implementation of calculators. The objective of this effort is to move forward with incorporating calculator use on the ASVAB, with completion date to be decided.
- Evaluate CAT-ASVAB methodologies and ways to streamline form development efforts. Completion data is to be decided.
- Evaluate the state of the ASVAB and prepare for the next generation of ASVAB and special purpose tests to be administered on the ASVAB platform in the ETP. These efforts are ongoing.
- Develop a CompT composite score to meet the National Defense Authorization Act (NDAA) requirement to address computational thinking skills. Completion date is to be decided.
- Develop a CAT version of the Cyber Test and program for administration on DTAC's cloud platform. This was completed in June 2023.
- Refine the MCt test of working memory and program for administration on DTAC's cloud platform. This is projected to be completed by fall 2023.
- Evaluate the use of TAPAS in the military selection and classification process. These efforts are ongoing.
- Inform the future development of an evidence-based accessions instrument for military compatibility assessment. These efforts are ongoing.

- Revise and maintain all CEP materials (website and print materials), conduct program evaluation studies and research studies as needed. These efforts are ongoing.
- Program ASVAB and special tests for delivery on DTAC's web-based/cloud-based platform and introduce enhancements. These efforts are ongoing.
- Expand the Internet version of the CAT-ASVAB (*i*CAT) test delivery application to run on additional operating systems and browsers for desktops/laptops. Expand the Pending Internet Computerized Adaptive Test (P*i*CAT) and AFQT Prediction Test to run on tablets and smartphones. Completion date to be decided.

At the end of the briefing, a committee member asked what DTAC is not doing that it should be doing because of resource limitations. Dr. Pommerich replied that DTAC is in a unique situation, one in which they have good resources. They received a funding push for 2017-2023 and the continuation and increase in that funding allowed them to do more work with the Next Generation (NextGen) ASVAB. She also said they received funding for the military compatibility effort allowing them to work on the TAPAS. She said they are in the best shape they have been in 21 years. Dr. Velgach said a primary concern is being able to meet short timelines and explained that more funding does not equate to being able to accomplish tasks faster. She said the number one challenge from leadership is that it takes too long to complete various testing efforts. Dr. Helland agreed. The committee member commented that a very large amount of work is being done. Another committee member noted the large number of deliverables are due in September 2024. Dr. Pommerich commented on the stress those deliverables are inducing, clarifying that, anytime the IT platform is involved, it adds complexity. She said the DTAC-HumRRO team is super, and she is thrilled to have the team; they do very good work. She said if they had more time for certain efforts (e.g., requirements in the NDAA), she would take a slower approach to some things, but the most difficult challenge is balancing utilization of psychometrically correct and rigorous methods and staying on schedule in accordance with the timeline. Dr. Helland said funding levels and priorities can shift quickly. Because they rely on Congress for their budget, they have to be prepared to react to Continuing Resolutions (CRs) or shutdowns, and all of this has implications on funding and timelines. Dr. Helland said the situations looks good at present, but it must be reevaluated every year. Dr. Pommerich described how DTAC had some lean years in the not-so-distant past, which demonstrates how things can fluctuate. She said her predecessors worked very hard to get more funding in place, and the current level of funding is where they should be operating, though unprecedented requirements may result in the need for additional funding.

A committee member asked if DTAC was considering using automated item generation (AIG) in test development. Dr. Pommerich said they had done a lot of work with the Educational Testing Service (ETS) to look at AIG, and it showed some promise but did not materialize into something that can be implemented for the purposes of saving time. She said Complex Reasoning (CR) is an AIG-produced test, as is Assembling Objects (AO). She said they developed 5,000 AO items and, if they all pan out, no more item development would be required for the lifetime of the test. Another committee member mentioned using a prompt engineer rather than item writers, which is having people familiar with content but willing to engage with AI systems. That is, humans learning to ask for the right things so they can take better advantage of the systems, which are continually improving. She said there are improvements in processes and tools and helping people interact with the systems in different ways, which may allow a more effective use of technology. Dr. Pommerich said they would need to make inquiries about what

they can do with AI. She said, though the government is very protective of their systems, DTAC needs to look into methods for using AI to streamline item development. The process would include prompt engineers rather than focusing on item writers and levels of human review. She mentioned improving techniques for drawing information from AI and streamlining the review process, and that a colleague is taking a course on that. Another committee member said he is starting to see academic articles on the subject in the personality testing domain; the items are imperfect and require human intervention, to potentially to include conceptual intervention to ensure the psychometrics are correct. Another committee member reflected that the process will be iterative as both human processes and systems are enhanced.

Dr. Velgach asked for the committee's thoughts on virtual proctoring (VP) and asked if they had experience with it. She noted that all the proctoring by the Department currently is in-person. One committee member mentioned seeing it used in credentialing agencies. S/he said the key question is the reason for its use: what is to be minimized? Cheating, item harvesting? S/he said this matters in respect to which techniques are put in place. Dr. Velgach said they have been talking with other countries, citing New Zealand. She said current OSD policy is to proctor tests used for high stakes decisions, so the U.S. is different than New Zealand in that sense – New Zealand has not historically used proctoring. In the case of New Zealand, there were concerns about whether VP is observing and identifying targeted behaviors. Dr. Velgach mentioned a case in which the use of a cell phone in front of a camera was not picked up, while a person moving their arms was picked up. The committee member said there is a real potential for bias in using scanning technologies, and that certain groups may be more likely to get flagged (or not flagged), which could lead to adverse impact. Another committee member commented about a large-scale test in corporate America and stated that corporate America is giving up on VP and moving forward with *no* proctoring, in particular for entry-level job testing. S/he said this is not necessarily best practice.

In the interest of maintaining item security, Dr. Velgach said she did not think they were in a position to say the item banks are large enough to pursue some of these measures. Ultimately, it will be a policy decision. Dr. Pommerich said they offer testing 24/7. A committee member asked if DTAC was going to do VP on its own. Dr. Pommerich said they are just gathering information to identify the concerns and then they could look at vendors. She said test compromise will always be the primary concern because it could take years to ramp back up if there was compromise from VP. She said they would be vulnerable to compromise. Dr. Helland explained that one driver of the task was consideration of recruiters' time demands. She said recruiters have to drive applicants to a testing site. If applicants could test at home, recruiters would have more time for prospecting.

3. <u>Form Equating Methodology</u> (Tab G)

Dr. Matt Reeder, HumRRO, presented the briefing.

> Dr. Reeder began by presenting a chart showing an overview of the CAT-ASVAB item pool development process, which includes (a) data collection and processing, (b) calibration and scaling, (c) item screening, (d) item enemy identification, (e) consolidating and process analysis results, (f) CAT pool assembly, (g) implementing additional CAT parameters, and (h) equating. He then provided background information on Pools 11-15 which are new CAT-ASVAB pools developed from tryout/seeded items that have passed

content, psychometric, and fairness/sensitivity evaluations. Items are assigned to pools through an optimization algorithm designed to maximize conditional precision levels of each pool and to constrain conditional precision levels to be comparable across pools. Each CAT-ASVAB item pool is unique in that there are no common items across pools. Operational CAT-ASVAB pools are typically repurposed when new pools are implemented, therefore new and prior pools do not share items. Item parameters of items included in pools 11-15 have been rescaled to the operational CAT-ASVAB scale (CAT-ASVAB pool 5-9) prior to the equating study. Standard scores are generated via a linear transformation, where the mean and standard deviation of the standard scores of new pools are matched to those of the reference pool. Linear transformation constants to transform $\hat{\theta}$ to standard scores are estimated during the equating study. Seed parameter values are calibrated using seed response data with the latent distribution of theta fixed to BILOG defaults (0,1). Operational responses from the calibration samples and operational parameter values are used to estimate the latent distribution of theta on the operational scale for the calibration sample. Transformation constants are computed to put the seed parameters on the operational scale. Dr. Reeder then showed a chart summarizing the item parameter rescaling process.

ASVAB forms/pools have historically been equated to a reference form/pool using equipercentile methods to produce equivalent composite distributions across alternate forms/pools. When ASVAB transitioned to Item Response Theory (IRT) scoring, new CAT-ASVAB pools continued to be equated to a reference pool using a linear method that matches the mean and standard deviation of standard scores to a reference pool. The current equating approach relies more heavily on the invariance property of IRT and aims to create equal distributions of scores across alternate pools. Large volumes of applicants qualify on CAT-ASVAB, and small differences between (unequated) pools can potentially have a large impact on the number of qualified applicants. Composite cut scores should achieve the same selection ratio across pools. Effectiveness of score equating is evaluated on the extent to which the pools can be used interchangeably to qualify the same proportion of applicants using selection composites.

The ASVAB score scale allows policy makers to compare current applicant aptitude with past applicants, and to set target qualifications accordingly. The current ASVAB score scale was developed from a nationally representative sample collected during the 1997 Profile of American Youth (PAY97) study. Changes to ASVAB, like introducing new CAT pools, must be introduced in a deliberate, carefully planned manner to ensure the continuity of the interpretation of ASVAB scores. Any given composite cut score should have the same meaning, irrespective of which pool is administered, as it did when standards were originally set. CAT-ASVAB pool 4 was included in the PAY97 norming study. It has subsequently been administered for special purposes only and serves to define the reference scale for future equating studies.

Rigorous equating procedures were developed by DTAC to equate pools 5-9 and put them onto the CAT-ASVAB scale. These procedures have served as a template for the development of future pools, including pools 11-15. They are conducted at the subtest level. Linear equating methods are used to derive constants to transform IRT-based theta scores on pools 11-15 to the scale of the reference pool 4 in a phased approach. A random groups design is used, in which each applicant is assigned a single pool with a one-in-seven probability. The pools in question are pool 4 (only used during equating studies), an operational pool (5) and a new pool (11-15). Subsequently, differences in qualification composite cumulative distribution functions (CDFs) between reference pool 4 and the new pools are evaluated.

Equating is implemented in three phases of operational administration of new pools to military applicants. Each phase includes progressively larger sample sizes. Phase sample sizes are cumulative such that they include all individuals from the previous phases. The intent of the phased design is to maximize the accuracy of reported operational scores. Phase I involves provisional equating based on IRT invariance. In phase 2, data from phase 1 are used to update the combined (across pool) transformation constants. This is also done in phase 3, using phase 2 data. Phase 3 data are then used to estimate the final separate (pool-specific) transformation constants to be applied to applicants testing after the initial operational test and evaluation.

Dr. Reeder then presented a chart showing the targeted and actual number of examinees per phase. All targets were met or exceeded. He continued by discussing the phase 3 analyses. Random group equivalence is assessed to determine if the assignment procedure produced equivalent groups with respect to key

demographic variable. Results suggested that the groups were, in fact, randomly equivalent. Analyses are then performed to determine if the linear transformation is adequate and if the subtest distributions have similar shapes. Some evidence of systematic differences in the shapes of the subtest distributions was found, but this is not problematic for the ASVAB given that qualification decisions are based on composite scores and the composites are likely to be more normal-like. Dr. Reeder then presented charts showing the subtest and composite distributions that supported this finding.

Further analyses are conducted to assess the pool composite equivalence. Composites can have different variances if the pools display different patterns of subtest correlations. Most composites displayed similar distributions across the new and reference pools. However, five composites consistently displayed statistically significant differences between the new and reference pools. However, the differences were comparable to those observed during CAT-ASVAB pool 5-9 equating; they were relatively small and within a tolerable range. Dr. Reeder then presented a series of charts displaying these outcomes.

Analyses indicated that most composite distributions compare reasonably well to the reference pool. However, several composite distributions based on new pools deviate from the reference pool up to 10 percent, conditional composite score. However, a large volume of applicants qualify on CAT-ASVAB and small differences between (unequated) pools can potentially have a large impact on the number of qualified applicants. This equating procedure ensures the equipercentile objective (i.e., distribution matching between a new pool standard scores and the reference pool standard scores), which enhances maximal similarity of qualification rates when a composite score is used for selection purposes. The effectiveness of the equating procedure is evaluated for each new pool against the reference pool, and more importantly, in comparisons of score distributions between the new pool composite and the reference pool composite scores.

Problems were identified with AO items considered for pools 11-15, including evidence of multidimensionality and ceiling effects. Therefore, items from pools 5-9 were reused while these problems were being addressed. Two sets of standard score transformation constants were used, one for the original pools 5-9 equating study and one for the current pools equating study. The AO test is part of very few composites, and the original and new transformation constants produced very similar composite distributions. The new constants are preferable due to slightly improved similarity to the reference pool. Analyses were also carried out to determine if subgroups perform at the same level across pools (e.g., females, Blacks, Hispanics). Results revealed multiple statistically significant pairwise comparisons for females, however the effect sizes were small. For Black examinees there was one significant pairwise comparison, which also had a small effect size. There were no significant pairwise comparisons for Hispanic examinees. These results are similar to those seen in pools 5-9 development and are not a concern given the small effect sized.

Mean differences on operational pool 5 were compared to reference pool 4. Statistically significant mean differences were found in several tests, but effect sizes were small. To answer the question of how closely the provisional equating transformations matched the final transformations, all applicants who took pools 11-15 were rescored using the final transformation constants. Total errors were calculated as the sum of the equating errors and the measurement errors. The total error was then compared with the standard errors of measurement. Dr. Reeder then showed a series of charts summarizing the results, which indicated that the provisional equating transformations closely matched the final transformations.

At the end of the briefing, Dr. Reeder asked if the committee had feedback on reducing Phase 3 or eliminating any phase. A committee member complimented the amount and quality of the work and said the purpose of the additional scale linking was to achieve similar distributions across pools. S/he then said that is done at the price of allowing for potential bias to creep in at the individual level ability estimates. There is potential bias at the individual level upon linking transformation, but that can be examined through a simulation study to measure the amount of bias in the individual theta estimates. It may be very minimal. The simulation study would set

conditions that mimic what has done and reveal the level of bias. S/he said once you have the results, everyone should feel better.

The committee member also asked to what extent bias is introduced in equating at the level of the individual. Is choosing pool 4 problematic? S/he commented that they chose pool 4 as the reference pool, that it shows some systematic differences, and asked if there are any systematic differences between pool 4 and other forms in terms of content or item format? S/he asked about the primary reason for choosing pool 4 instead of a newer pool? Dr. Reeder said the CAT pool 4 was used based on precedent and the need to establish a scale that is tied to PAY97 to allow "apples-to-apples" comparisons. He said pool 4 is the gold standard in that sense. Dr. Pommerich explained that pool 4 is only used for equating purposes and is not exposed. She said the content is largely the same and formatting should be the same, so the only issue might be the age of the items.

The committee member then noted the use of MLE during the linking process and asked if they were using MAP estimates, and if not, why not? Dr. Trippe (DTAC) said scoring is done with MAP, but the metric used in this case to estimate transformation constants is MLE. Dr. Reeder referred to a technical bulletin that he would need to consult before commenting further.

Another committee member said, "fantastic technical work," and asked how the outcomes valued in this effort align with what might be coming in the NextGen ASVAB (new tests, revised tests, closer examination of high school curriculum). S/he noted the large number of moving parts and asked the following questions: Is there a chance you could ever break the scale? What is the vision for NextGen ASVAB and the implications for your processes? Could you change software? In the bigger picture, of the many potential changes of possible changes, what matters? What are the outcomes you value the most? Dr. Pommerich said those were great questions. She said they care about score continuity and the scale and will have to map out the future. . She said some new tests of interests have not been normed yet. She said if they conduct a renorming of existing tests, then that would break the scale. She said they are waiting on a couple of reports that will provide guidance on norming. These reports identify which tests – of which there are many – could have a major impact for the Services, so they will need to decide how to proceed.

Toward the end of the discussion, a committee member raised two potential considerations: First, small differences in scoring can have a large impact on qualifications and selection decisions. Conditional equating on theta could be used to reduce bias. If tests are not perfectly invariant or reliable, these problems will exist. Second, do you worry about parameter drift due to test security or other factors? Dr. Trippe responded that slide 29 shows pool 4 comes out every so often for purposes of equating studies, and Form 5 has been out since 2008-2009. Dr. Trippe said he has not seen much parameter drift, but if it exists, it is small. Dr. Pommerich mentioned simulation studies that found parameter drift over time is much like that seen in national data, noting results did not raise concerns. She said they did not see drift resulting from their methodology.

4. ASVAB Item Development Process – Item Writing (Tab H)

Mr. Jeff Harber (Office of People Analytics [OPA]/DTAC) and Ms. Tiffany Day (HumRRO), presented the briefing.

Mr. Harber began the presentation by explaining that, prior to item development, DTAC determines the number of tryout items to be written for each subtest based on form development goals. DTAC also provides specifications for the item difficulty percentages for each subtest along with the *Guide to Item Writing for the ASVAB, Sensitivity and Bias Guidelines for the ASVAB,* and the taxonomy structure for each subtest (the blueprint) with weights for development. Ms. Day then showed a chart detailing the number of items that have been developed for nine of the subtests (excluding AO) for each year from 2017 to 2022. She went on to explain that HumRRO has an overall project director who (a) manages production of tryout-ready items and images; (b) ensures conformance with security requirements, test specifications, and standard operating procedures; and (c) manages budgets and deliverables. Each subtest has a team composed of one to three editors. The editors for Arithmetic Reasoning (AR), Math Knowledge (MK), Paragraph Comprehension (PC), and Word Knowledge (WK) all have subject matter expertise in the content of those test blueprints. Editors for General Science (GS), Automotive Information (AI), Electronics Information (EI), Mechanical Comprehension (MC), and Shop Information (SI) rely more on consultants and item references. Junior editors manage item writers and conduct initial rounds of copy and content edits. Senior editors conduct final rounds of copy and content edits and act as the point of contact for DTAC for submissions. HumRRO graphic artists render images in Adobe Illustrator for seven subtests (i.e., AI, AR, EI, GS, MC, MK, and SI). They use templates to comply with ASVAB graphics specifications provided by DTAC. Graphic files are saved to the ASVAB item bank. There are 20-40 Subject Matter Experts (SMEs) who serve as item writers. This number varies based on the quantity of items needed. Some write for two or more subtests (e.g., AR, MK) and are responsible for developing draft items and images. Finally, there are 4-6 SMEs who serve as content reviewers by conducting final reviews of the DTAC-approved items with images.

Ms. Day continued by explaining that, to prevent disclosure of controlled items and test materials, all project team members and consultants must sign and comply with the terms of a security/confidentiality agreement. Each SME agrees to submit only original work they author for the ASVAB. All HumRRO project team members are required to complete DoD DHRA contractor training. In addition, for the current task order, item development work for all subtests is conducted in item banks maintained by HumRRO. There is a unique item bank for each test with multi-factor authentication required for access. User permissions are restricted to features authorized for a given role. For instance, item writers can only access and edit the items they draft.

Ms. Day continued by discussing the guidance given to item writers. This includes a tool kit – unique to each subtest – that provides training documents and resources. Subtest-specific guidelines for writing items include (a) the subtest blueprint with categories and subcategories, (b) best practices for writing stems and answer choices, (c) guidance for targeting item difficulty and preventing bias/ensuring fairness, (d) guidelines for citing authoritative references, and (f) an item writer checklist. To be accepted, items must adhere to the guidelines and require only minor to moderate editing. The guidance provided also includes annotated items for each subtest that illustrate best practices.

Item writers are also provided a role-specific ASVAB Item Bank manual. They can only access their current, in-progress draft items which they cannot access once accepted by HumRRO editors. Item writers for some subtests are given additional guidance, including:

- Criteria for including images with items (all subtests except PC and WK)
- More detailed versions of the subtest blueprints (AR, MK, GS, PC)
- Sample stems targeting each blueprint area (AR, MK, GS, PC, GS, SI)
- Criteria for providing rationales for answer choices (AR, MK)
- Best practices for writing PC paragraphs with target lengths of 100-180 words
- Criteria for selecting PC paragraphs from the public domain.

Ms. Day then presented an example of guidance provided to AR item writers regarding targeting item difficulty (e.g., ask the examinee to solve a complicated or multi-step problem). She noted that it is not permissible to combine measures of more than one area of the blueprint in a single item to increase the difficulty level. She also provided examples of guidance given to PC item writers on (a) preventing bias and ensuring fairness, (b) ensuring compliance with best practices for writing paragraphs, and (c) identifying existing sources for paragraphs. Ms. Day then showed a table displaying the percentage of PC stimuli material sourced from the public domain by blueprint code and year from 2017 to 2022. Ms. Day concluded this section of the briefing by listing HumRRO editor resources for reviewing item content, including item writer tool kits, editor checklists, references specific to subtest and item content, and style guidance for copy editing.

Ms. Day continued by outlining the process for item editing. Junior editors assign work to SME item writers, typically a set of 10-20 items to be developed. They evaluate the quality of the items submitted relative to the criteria for acceptance, then accept or reject each item. The reason for rejecting an item is noted in the ASVAB Item Bank (e.g., revise a distractor, provide a reference). When all items in an assignment are accepted, the SME item writer submits an invoice.

Items accepted by HumRRO go through an iterative editing process. Teams of editors use general and subtest-specific guidance to inform revisions to item content (e.g., improve clarity of language, verify accuracy using references). Editors may ask SME item writers about potential substantive edits to ensure they will not alter the item's technical accuracy. Revisions to a set of 10-20 items may be needed to meet the specifications for each 100-item series (e.g., distributions of estimates of difficulty, positions of correct responses). After edits to a set of 10-20 items are completed, senior editors inform DTAC that they are ready for initial review and approval in the Item Bank.

Mr. Harber continued by explaining that DTAC editors review and edit the items for (a) style, (b) content accuracy, (c) stem clarity, (d) absence of clues to the correct answer, (e) appropriate patterns among the distractors, (f) bias and sensitivity, and (g) possible enemy items. Items are approved as is or with edits, or DTAC requests a rewrite or replacement. DTAC informs HumRRO when first reviews have been completed. Ms. Day indicated that HumRRO's senior editor responds to the DTAC review and resubmits items for re-review and approval. The senior editor also monitors the distribution of approved items to ensure subtest specifications are met. For approved items that contain graphics, the senior editor coordinates the development of Adobe Illustrator images and verifies that the images match content and meet graphic specifications. Once a series of 100 items is approved, the senior editor updates the ASVAB item bank to prepare for content review. The senior editor coordinates with the independent SME content reviewer and grants the reviewer temporary access to the item series in the Item Bank. The content reviewer (a) verifies that the content is factually accurate; (b) confirms that the artwork is accurate, understandable, and readable; (c) confirms the indicated choice is the correct answer; and (d) recommends revisions and provides a rationale and/or cites references to support revisions. The senior editor evaluates the content reviewer's edits and feedback and approves and/or applies edits in response to that feedback. Final items and supporting documentation for a series are then submitted to DTAC. DTAC may request SME content review of modified items, identify artwork revisions needed, or accept the final items and artwork as delivered. At a minimum, each tryout ASVAB item is reviewed by seven to eight pairs of eyes at least once (i.e., 1 SME item writer, 1-2 HumRRO junior editors, 1 HumRRO senior editor, 3 DTAC editors, 1 independent SME content reviewer).

At the end of the briefing, a committee member asked if item writing and verification are done by just one person. Ms. Day said they are not. She said item assignments are made at the blueprint level. After an item is written to a competency, junior editors verify that it measures the knowledge. This is followed by reviews by senior editors, a content reviewer, and Mr. Harber's team. She said, they all confirm independently that each item aligns with the blueprint competency. Another committee member had a question regarding centering item writers' thinking about that population at the beginning of the process in order to ensure the use of words and terminology that are common knowledge. S/he asked what processes are used to come to an understanding of what is common knowledge? That is, thinking about examinees and their characteristics, how are reviewers trained to think about who these people are? Dr. Harber

replied that the difficulty lies with what is being introduced in the schools. The age bracket for the item development material is 8th grade thru college, and what students might peripherally have access to (from the Internet). However, he said item writers have to be careful that some topics are not too advanced, such as 3rd or 4th year of college or technical content that students are not going to get until they are in a very specialized area. The committee member then commented, in reference to diversity in the population, that others knew things (e.g., agriculture) that s/he did not know based on where they grew up. Dr. Harber replied that not all item writers are solely teachers, but that some also have technical careers, which allows them to make judgements based on these additional experiences. He said this is the case with HumRRO personnel as well. The committee member commented, "the more eyes the better." Dr. Harber said he has been doing this for years and the information HumRRO uses to develop items is built on broad experience and is the result of a team effort. He said DTAC and HumRRO give feedback to each other, as opposed to just providing items and then signing off.

5. ASVAB Item Development Process – Item Analysis (Tab I)

Dr. Matt Reeder, HumRRO, presented the briefing.

Dr. Reeder began the presentation by displaying a graphic showing the steps in the CAT-ASVAB item pool development process, including (a) data collection and processing, (b) calibration and scaling, (c) item screening, (d) item enemy identification, (e) consolidation of analysis results, (f) CAT pool assembly, (g) application of additional CAT parameters, and (h) equating. He then provided details about tryout item data processing. This includes data cleaning by removing invalid, ineligible, or corrupt records such as non-Service applicants and records with invalid person/item identifiers. Records that suggest potentially unmotivated responding (e.g., high percent of missing responses, anomalous response latencies) are also removed. One record for an applicant who has tested more than once is also selected. A CTT-based pre-calibration check is then run to evaluate patterns of response option selection and option-total correlations. Items with questionable response patterns are flagged (e.g., low/negative item-total correlation for key, positive option-total correlation for non-keyed response). Content SMEs review these items for potential mis-key, multiple correct responses, or other issues. Mis-keyed items are corrected and rescored, as necessary. Items with multiple correct responses or content flaws are removed.

CAT-ASVAB is based on the Three-Parameter Logistic model (3PL). DTAC simulation studies of the calibration process suggest an item-level sample size of 1,000 or more cases is desirable for optimal parameter recovery. Therefore, target-level sample sizes are set at 1,200 to account for data loss. Achieving the target depends on testing volumes but generally takes about 8 months of data collection. Each subtest is calibrated separately using BILOG-MG.

DTAC simulations find that parameter recovery is improved as the number of seeded items administered to each examinee increases. Recovery is relatively poor when 10 or fewer items are administered, therefore each examinee responds to 15 randomly administered tryout items per test according to the seed design. Tryout items are calibrated in seed versions, with 200, 400, or 800 items per calibration. To avoid a sparse response data matrix about 16,000 examinees are needed for AI, AO, EI, SI, and MC, while 32,000 examinees are required for AR, GS, MK, and PC, and 64,000 examinees are needed for WK.

Dr. Reeder then presented a graphic showing a one-page summary of item review information that is generated for each item. It includes item information, eight item-model fit indices, distractor analysis, and differential item functioning data. Many items are automatically eligible for operational status given there are no item quality flags. Some items are automatically ineligible for operational use due to such factors as out of bounds parameter estimates. Some items require psychometric or content review to determine eligibility. This review is a subjective task, where analysts consider all available empirical evidence. Two analysts independently rate each reviewed item as "keep" or "drop." When they agree, the eligibility status

is final. When they disagree, they meet to discuss their ratings and reach a consensus. Occasionally, SMEs are asked for input.

Dr. Reeder continued by discussing bias detection procedures during item review. Item-level sample sizes allow for three item performance differential item functioning (DIF) statistics. Items categorized as "C" or "moderate to severe" DIF according to the ETS framework are reviewed for evidence of bias. Review sessions include members of both focal and reference groups. Reviewers are trained on basic concepts of DIF and construct-irrelevant factors and are provided several examples of items that include construct-irrelevant content. If the reviewers conclude that an item includes construct-irrelevant factors that might plausibly prevent members of a group of test takers from responding to the item in ways that allow appropriate inferences about their knowledge, skills, or abilities, the item is not eligible to be on an operational form. If they conclude that no such construct-irrelevant factors are apparent and the item passes all other psychometric quality screens, the item remains eligible for assignment to a form or pool. Dr. Reeder then presented examples of data/graphics used in DIF screening. Another graphic showed the percentages of items retained during the development of Forms 11-15, and a second set of graphics showed the feedback provided to the content development team based on the results.

Dr. Reeder then turned to a discussion of CAT-ASVAB dimensionality. Each of the 10 CAT-ASVAB subtests is calibrated separately and scored using a unidimensional IRT model. The CAT-ASVAB algorithm implements content balancing for two subtests (AO and GS) based on the outcome of prior dimensionality analysis work. IRT models are robust against minor violations of the unidimensionality assumption. The original developers of CAT-ASVAB considered three approaches to dealing with dimensionality. Current practices continue to rely on the outcome of the original dimensionality evaluation, with construct blueprints and item development practices unchanged. Dr. Reeder then showed a table summarizing the three approaches for dealing with this issue—unidimensional treatment, content balancing, and pool splitting (i.e., separate calibrations of items from each content area). Segall, Moreno, and Hetter (1997) present a decision framework for dimensionality that considers the statistical significance of multidimensionality, the interpretability of factor solutions, the overlap of item difficulties (i.e., academic versus nonacademic content), and the factor correlations. Full information item factor analysis was used to evaluate dimensionality via TESTFACT. Dr. Reeder then showed a table summarizing the dimensionality framework and summarizing the results when applied to the ASVAB subtests.

A previous CAT-ASVAB seeding design involved recalibrating operational items along with tryout items. Recovery of the operational difficulty parameter values is indirect evidence of unidimensionality. Difficulty item parameter values were recovered as expected for all tests except AO when this seeding design was operational. Failure to recover operational difficulty parameter values was important in diagnosing multidimensionality in a generation of AO tryout items.

Recent investigations into ASVAB dimensionality included examining the Cyber Test and the feasibility of combining AR and MK. Dimensionality analyses are rooted in two frameworks: IRT-based (e.g., correlations between theta estimates, item misfit) and item factor analysis rooted in bifactor modeling (e.g., explained common variance, comparison of general factor loadings from one-factor and bifactor models). These and other approaches could be explored in the future for general use in ASVAB item and form development. Evaluating dimensionality of tryout items is not currently part of the pool development process. The primary reason is that each examinee is administered 15 of either 200, 400, or 800 tryout items resulting in a very sparse response data matrix. Covariance and full information maximum likelihood (FIML) approaches either do not work well or are practically challenging. An IRT model-based approach may be an option; it has been applied to a somewhat similar data structure in a related context. For now, we rely on foundational research on dimensionality and the fact that constructs, blueprints, and development procedures remain constant.

At the end of the briefing, Dr. Reeder asked if the committee had any recommendations for evaluating dimensionality under sparse data conditions. A committee member replied that dimensionality assessment is one of the most difficult things to do and, if the results across methods do not agree, there are few options. However, classical statistics are a possibility; like

ITCs, they may provide some dimensionality information. The committee member then asked what the team had identified as strong evidence of multidimensionality and what they planned to do. Dr. Reeder said it depends on circumstances; in the AO, GS, AS situations, it depends on how it is manifesting. Another committee member expressed uncertainty as to what to do with such sparse data and mentioned the potential of machine learning variations. S/he said the practical recommendation had already been made, but when flagging items for review, themes may appear, or when a statistical test shows something is not invariant, the content will typically reveal why that is the case. The committee member asked if that is what Dr. Reeder's team does. Dr. Reeder responded that he cannot say if there are general guidelines across tests. The committee member asked what guided revisions in the case where a large set of items was retained, and subgroup differences were found to be statistically significant. Dr. Trippe replied that it depends on the group, but sometimes the male-female comparison reveals evidence of different socialization. He said a classic example is that men have more sports equipment, and so the team discusses whether that comes into play in regard to whether a specific piece of equipment should be considered obscure, and the question fair. He said these conversations are interesting, but there is not a large theme other than differential socialization. He said, in those cases, they defer to the group that is disadvantaged based on opportunities. An audience member asked if dropped items were revised and then retained, and what would the item look like? Dr. Pommerich said they do not drop many items in that review – perhaps 10 out of 1,000 items – so revision is not worth the effort.

Dr. Velgach noted that an article referenced by Dr. Reeder was the chapter from Dr. Daniel Segall and that it has been posted online.

6. TAPAS Overview – Validity Framework and Joint-Enlistment Composite (Tab J)

Drs. Deirdre Knapp and Dan Putka, HumRRO, presented the briefing.

Dr. Knapp began the presentation by providing an overview of the TAPAS, which was developed by the Drasgow Consulting Group under a Small Business Innovation Research grant with the Army Research Institute for the Behavioral and Social Sciences (ARI). Promising research led to the Army's use of the TAPAS to support enlistment selection decisions. These findings prompted the other Services to initiate their own TAPAS research programs. A RAND report identified some technical concerns which prompted the formation of an independent TAPAS Evaluation Project. One recommendation resulting from this effort was that a theory of action validity argument framework be developed for the TAPAS.

The TAPAS includes DoD-owned statement pools for 27 personality facets, with 13-15 facets typically included on a given TAPAS version. The TAPAS uses multidimensional pairwise preference (MDPP) items, with most items presenting two statements from different personality dimensions. The statements are matched on the strength of the dimension and on the socially desirable nature of the response options. Items are generated on the fly by selecting from pools of pre-calibrated personality statements that measure construct dimensions relevant to performance in the military. The assessment is scored using a multi-unidimensional pairwise preference IRT model.

The purpose of validity argument frameworks is to compile, organize, and review existing evidence related to the use of the assessments. Relevant information is defined broadly, and includes all aspects of the assessments design, development, administration, scoring, and reporting. The outcome is an evaluation of whether available evidence supports the use of the assessment for its intended purpose and the identification of ways to strengthen the evidence supporting its use. The results help inform improvements to the assessment in terms of content, scoring, administration, and/or interpretation. Dr. Knapp then showed

a graphic summarizing the validity argument framework method, which includes (a) developing a Theory of Action (TOA) that identifies major claims made for the assessment, (b) deriving an interpretive argument that identifies specific claims and assumptions, (c) collecting evidence for each assumption, and (d) summarizing the information in a validity argument.

Dr. Knapp continued by displaying a graphic summarizing the TAPAS TOA. Major claims include (a) temperament facets are predictive of performance and continuance intentions/behaviors, (b) TAPAS measures a useful sample of temperament facets, and (c) respondents selected or classified based on TAPAS scores (in combination with other indicators) have a higher likelihood of success within particular military occupations. She then presented an illustration of a specific claim and assumptions. The resulting interpretive argument included 3 major claims, 18 specific claims (6 for selection, 7 for classification, and 5 for both), 47 assumptions (14 selection, 14 classification, 19 both). Dr. Knapp provided a handout summarizing the full interpretive argument.

The evidence strongly supported claim 1, that temperament facets are predictive of performance and continuance intentions/behaviors, particularly when temperament is used for selection purposes. Research evidence regarding temperament and occupational classification is limited by a lack of accumulated evidence or a lack of sufficient relationships. Findings also suggest that as individual trait levels change over time, they change somewhat in tandem across people, thereby maintaining the trait-criterion correlation over time.

Evidence in support of claim 2, that TAPAS measures a useful sample of temperament facets, is generally positive. However, important information is not sufficiently documented to judge, and much of the available evidence could be strengthened with additional research. There is evidence suggesting that TAPAS facets are relevant for selection, but much less so when it comes to classification. There is insufficient documentation to determine whether the TAPAS statement pools sufficiently cover the range of extremity and social desirability parameters to support reliable and accurate measurement, or to critique the CAT algorithm used for statement selection, pairing, and scoring. While there is some evidence of cross-format and cross-version score correspondence, there is insufficient evidence to more broadly conclude that scores are not affected by use of different statement pools and/or different sets of facets included on a given version of the TAPAS. Subgroup differences on TAPAS scores are generally minimal, especially when compared to cognitive ability measures.

Regarding claim 3, that respondents selected or classified based on TAPAS scores (in combination with other indicators) have a higher likelihood of success within particular military occupations, most of the research reviewed addresses notional, rather than current or proposed operational selection and classification decision-making systems. More evidence would be needed to support specific operational applications of the TAPAS. The evidence suggests that multiple TAPAS facet and composite scores (other than Can-Do) tend to show incremental validity over the AFQT, particularly for motivational and retention-related outcomes. Available evidence from controlled studies suggests TAPAS scores display moderate to moderately high test-retest correlations, although results using ad-hoc military retest samples are low. Additional research is needed regarding the reliability of TAPAS. Available evidence shows that TAPAS facets and composites exhibit differing levels of criterion-related validity across occupations. Patterns of incremental validity over ASVAB aptitude area scores are similar to those observed with AFQT scores. There is no direct evidence, however, relevant to the efficacy of using a finite number of TAPAS composites to identify the types of occupations in which enlisted personnel would be most successful.

Dr. Knapp continued by summarizing the validity argument report recommendations. These included (a) expanding and improving TAPAS-related documentation, (b) strengthening content and construct validity evidence, (c) broadening the criterion-related validity investigations, (d) strengthening psychometric evidence, and (e) making administrative improvements. The intent is to integrate the recommendations from the TAPAS Evaluation Project and the current work to produce an updated research and development agenda. Dr. Knapp concluded by highlighting other work in progress, including completing in-progress technical reports on TAPAS development and psychometric properties, and developing (a) design recommendations for a joint-Service TAPAS and a joint-Service selection composite, (b) a research plan to

collect cross-Service criterion-related validation data on the joint-Service composite, and (c) a strategy to evolve the TAPAS validity argument framework.

Dr. Putka began the discussion of the interim joint-Service TAPAS composite by stating that the objective is to develop a composite that can be used to inform general enlisted selection and qualification decisions, complements other measures and data used in the selection process (e.g., AFQT scores, medical/physical/conduct data), and predicts first-term enlisted job performance. The steps to be followed include:

1. Identify first-term enlisted performance dimensions.
2. Capture "overall performance" policy.
3. Define the universe of potential TAPAS facets for the interim composite.
4. Establish interim composite development and validation strategies.
5. Gather archival and SME data to support development and validation.
6. Build and provide initial evaluation of the interim composite.
7. Evaluate the composite based on archival data.

In identifying first-term enlisted performance dimensions, the first focus was on the ten joint-Service dimensions adapted from the joint-Service performance taxonomy for entry-level occupations. Dr. Putka showed a chart listing these dimensions. To gain an understanding of the relative importance the Services place on various performance dimensions when defining overall performance, a policy capturing exercise was conducted. Each Services' representative to the Military Accession Policy Working Group (MAPWG) and key stakeholders and policy representatives were asked to identify individuals who could help define overall performance priorities for first-term enlisted Servicemembers. These individuals were then asked to distribute 100 points across the ten performance dimensions so that the resulting point distribution would reflect how their Service would "effectively" weight these dimensions in defining an overall first-term enlisted performance composite. Overall, the Services tended to give most weight to the *Task Performance, Decision Making, Problem Solving, and Innovation* dimensions, and least weight to the *Counterproductive Work Behavior* dimension. The reliability of the mean, cross-Service profile was relatively high [ICC(C,5) = .76]. The Air Force, Space Force, and Navy profiles were most aligned with the mean profile, while the Army and Marine Corps were least aligned.

Currently, the Services administer different versions of TAPAS during the enlistment application process. Only six facets are common across all versions, and there are 24 facets that appear on at least one of the versions. The current effort aims to build an interim joint-Service composite using a subset of the 24 facets, given limitations in administration time. DTAC and the AP aim to have recommendations and a preliminary evaluation of the interim composite available by the fall of 2023. This is challenging given the lack of data on the job performance criteria of interest and a timeframe that does not allow for execution of a criterion-related validation study. The solution is to ground identification of facets and initial evaluation of the composite in a well-established validation framework from the I/O psychology literature. Dr. Putka presented a graphic of a validation framework and stated that the approach to developing and initially evaluating the interim joint-Service TAPAS composite is based on amassing theory and data to assign labels to predictor scores to establish a linkage between the observed predictor measure and the latent predictor domain. This has been well established through the TAPAS Validity Argument work. The focus of this work will be using logic and judgement based on existing theory and the body of relevant empirical evidence to establish a linkage between the latent predictor domain and the latent criterion domain. This will entail building an AFQT-TAPAS-performance dimension correlation matrix based on archival applicant data and SME correlation estimates. These data will be used to simulate *n* "population" correlation matrices. Frequency distributions of AFQT scores and TAPAS facets will be obtained from archival applicant data. The population matrices and distributions will be used to estimate *n* large samples of AFQT-TAPAS performance dimension data. An overall performance composite will be calculated by applying nominal weights for each performance dimension in each sample. The next step will be to generate full path of Least Absolute Shrinkage and Selection Operator (Lasso) models using the overall composite as the criterion and 25 TAPAS facets as the starting set of predictors. Lasso is a regularized regression model that performs variable selection. The TAPAS facets that tend to remain in the model as

the Lasso constraint becomes more stringent will be identified, and the tradeoff between the number of facets included and model *R* will be identified.

Archival applicant data on AFQT and TAPAS have been obtained. They will allow researchers to empirically estimate AFQT-TAPAS and TAPAS-TAPAS correlations and create AFQT-TAPAS score distributions using very large applicant samples. A group of eleven external PhD researchers with expertise in personality/cognitive ability, job performance relations, and job performance constructs were asked to estimate the intercorrelations of the performance dimensions and the correlations between 1) each of the 24 TAPAS facets and each of the ten performance dimensions, and 2) the AFQT and each of the ten performance dimensions. The SMEs provided "construct" level correlation estimates, assuming the performance dimensions were free from error and had no range restrictions. At the end of this step all data needed to build AFQT-TAPAS-performance dimension correlation matrices had been obtained.

The data gathered were used to simulate data for multiple large samples of individuals. Within each sample, full path of Lasso models were generated to identify which TAPAS facets tended to remain in the model as the Lasso constraint became more stringent. A follow-up evaluation was conducted to evaluate the sensitivity of the results to how performance dimensions were weighted using a Service-specific performance dimension profile as opposed to the cross-Service profile. The change in the best-bet TAPAS facets when a residualized version of the overall performance composite was used as the criterion (removing variance due to AFQT) was evaluated. This helps identify an interim composite that would best increment the validity of the AFQT for predicting overall performance.

A potential final evaluation of the interim joint-Service composite may involve scoring the best-bet composite using archival TAPAS data from applicants to examine the magnitude of subgroup differences on the composite applicant samples (and in turn the potential for adverse impact). An evaluation will be conducted of the criterion-related validity of the composite for predicting other criteria of historical interest (e.g., first-term attrition), and the difference between composition and weighting of the interim joint-Service composite versus a TAPAS composite optimized to predict first-term attrition. Dr. Putka concluded by discussing the possibilities for generating further validity evidence by obtaining administrative data (e.g., training completion, attrition), as well as self-report and peer/instructor/non-commissioned officer ratings generated in the DoD Criterion Measures project.

As Dr. Knapp briefed the validity argument claims, Dr. Velgach commented that the TAPAS work is on-going and is constantly being updated. Because this is being presented publicly, everything must be reviewed from a sensitivity perspective. She said there is more information available, but it could not be included in the presentation in time to complete the sensitivity review.

At the end of the briefing, Dr. Knapp asked the committee if it knew of alternatives to criterion-related validity evidence or ideas for collecting data? A committee member complimented the presentation. S/he then mentioned that using a synthetic validity approach could help make the most out of available information; that is, collecting data and making use of it across a larger array of jobs. The committee member also mentioned collecting validity data retrospectively, to the extent there is not range restriction, because selection was not based on TAPAS scores. S/he compared it to a concurrent validation but using something with a low correlation with the TAPAS. Dr. Knapp said a complicating factor is that operational pre-enlistment testing facet statement pools can only be used at MEPS; they would need to use the original statement pools that are now reserved primarily for research purposes. Another committee member said s/he had no suggestions beyond what had already been offered but asked how composite scores and weights would be determined: would it be by statistical means or input from each Service? Dr. Knapp replied that the Army has established weights based on regression toward criteria of interest, so they could use that approach. She also said that Services have different rankings of

what is most important, so the composite could conceivably be slightly different across Services. Dr. Velgach clarified that the goal is to have a single joint composite with the same facets and weights, but the Services could implement additional composites based on their objectives. Dr. Putka said they could recommend a single weighting scheme, but in this early stage, they want to include the facets that would offer flexibility for future research. So, if there are differences in terms of facets that predict performance as defined by each Service, they want to factor that into the recommendations as well. If they can offer more flexibility for future Joint Service composites, then they will try to do that. Dr. Velgach commented that the initial recommendation may be to use preliminary weights and then investigate the situation further when they have additional data.

7. TAPAS Future Work – Compatibility Composite (Tab K)

Dr. Kevin Bradley, HumRRO, presented the briefing.

Dr. Bradley began by providing background information regarding this effort. In 2021 the President directed the Secretary of Defense to form an Independent Review Commission (IRC) on sexual assault in the military. The goal was to address sexual assault and harassment in the force and make recommendations related to accountability, prevention, climate and culture, and victim care and support. One recommendation emerging from the commission's work was to implement a pre-accession assessment to screen for alignment with military core values (i.e., military compatibility). Prior to the recommendation, the Military Compatibility Research Group was formed to ensure that the men and women selected to serve as members of the military possess traits supportive of, and positively aligned with, military core values. Initial work was performed between 2020 and 2022 by the Defense Personnel and Security Research Center (PERSEREC). They conducted a literature review and identified conceptual predictors of misconduct, counterproductive workplace behaviors, violence, sexual assault, crime, antisocial behavior, and attrition. They also reviewed existing security screening practices in the military, as well as other federal agencies and in law enforcement, which resulted in a comparison of applicant compatibility assessment practices.

Dr. Bradley continued by describing three lines of research to be conducted. The first involves developing an assessment of military compatibility. This will entail (a) providing support to the Military Compatibility Research Group, (b) developing plans for evaluating tests and incorporating clinical assessments, (c) creating a TAPAS compatibility composite, (d) investigating alternative assessments and composites, and (e) refining the TAPAS validity argument. A second focus will be on software engineering to make infrastructure improvements, a plan for modernization of the TAPAS application, and updates to Authority to Operate-related documentation and procedures. The third effort will be to research non-cognitive methodologies directed at the officer population, which will include conducting a best practices forum, reviewing assessments for possible use, and comparing alternative assessments.

Dr. Bradley then turned to the identification of TAPAS Compatibility composites, stating that TAPAS has been identified as the principal tool for assessing enlisted members' personality and character attributes. The goal is to create a Compatibility composite or composites and minimum score(s) to be used in initial operational testing for military compatibility in the enlisted population. The Army's TAPAS Conduct composite will be the baseline. Procedurally, the approach will be very similar to that used to create the joint-Service TAPAS composite. Some unique challenges include specifying the criterion space and conducting research with low base rate criteria.

A literature review will be conducted to identify alternate constructs and assessments and investigate alternate facets, composites, and instruments to predict compatibility in the enlisted population. The review will address the criterion space, including the dark triad, predictors of sexual assault and harassment, predictors of counterproductive workplace behaviors, and approaches to validating assessments designed to predict low base rate behavior (e.g., forensic and clinical assessment literature).

> Another avenue of investigation will involve developing a potential plan and feasibility analysis to incorporate evaluation by a licensed clinician into the enlisted accessions process. This will require a clinical assessment process model and feasibility evaluation given applicant volume, geographic disbursement, and other logistical challenges. The model will address such questions as what credentials will be required of those doing the clinical assessments, which applicants will be assessed, how many can reasonably assessed, and where in the accession process the assessment should occur.
>
> An additional step will be to develop a plan and research design to evaluate applicable test(s) of military compatibility for the enlisted population. This will include a longitudinal research plan to evaluate how well individuals who possess traits incompatible with core military values or are otherwise at risk of committing violent or criminal acts can be identified. Potential screening methods include TAPAS military compatibility composites, alternate personality and psychological assessments, and assessments by licensed clinicians. Challenges include predicting low base rate criteria, the multidimensionality of the counterproductive workplace behavior space, and the availability (or lack thereof) of criterion data.
>
> Dr. Bradley then turned to the topic of military compatibility in regard to officers. He noted that the pathway of officers into the military differs from that of enlisted personnel, potentially requiring a different approach to assessing military compatibility. A Non-Cognitive Assessment for Military Compatibility Best Practices Form has been established to ensure OPA is current on research, possible methods, and technological advances for assessing compatibility, while maintaining best practices in the use of personality and compatibility assessments. The TAPAS and other non-cognitive assessments will be investigated for use across the various officer commissioning sources. Research designs will be developed to compare alternative non-cognitive assessment options.

At the end of the briefing, a committee member asked why a clinical assessment was thought to be more predictive than a measure like TAPAS or another test battery. Dr. Velgach replied that PERSEREC did a holistic review of models used across the spectrum, and found a combined approach was common practice. That is, first use an assessment battery, then a more clinical assessment. Clinical assessment requires more time and may be more appropriate for a smaller population flagged by the initial assessment battery. She said the combined approach provided additional information on whether to disqualify. She said the evaluation plan would address the incremental validity of the approach. She also said it is difficult getting psychological/behavioral health consults as part of the current medical evaluation process, therefore including this as part of the compatibility assessment may prove to be unfeasible. Dr. Velgach added that the incremental validity here is critical. The committee member suggested the validity challenge would be greater trying to predict low base rate events. Dr. Velgach said AP had a similar comment to the IRC, but the IRC described the clinician portion as an additional check for making a decision of that kind.

A committee member inquired about the content of military core values. Explaining that core values vary by Service, Dr. Velgach cited examples as being honor, courage, commitment, and sense of duty. Another committee member suggested that anything that is low base rate will be difficult to predict, but one way to do it would be to identify more proximal behavior, though that approach sacrifices the generality of the trade. Dr. Velgach said the work on the officer side will also inform enlisted work. She stressed the importance of being informed about research, in general, that has implications for all applicants. Another committee member asked whether it was possible to deconstruct CWBs into essential components, for example, assault against an individual versus abuse of property, and then collapse them to address the low base-rate problem. S/he also said one is assuming there are differences among them because you wanted multiple composites. Dr. Bradley said he has looked at the taxonomy including aggression against

individuals or property but, aside from saying it was far from unity, he could not recall the correlation between them. He also said he did not know how much sexual harassment correlated with those variables either.

A committee member commented on the variability among Services in ratings for CWBs and asked if predicting CWBs was going to vary by branch. Dr. Bradley replied that it might be difficult to ask policy members from the Services to identify what is more important, sexual assault or sedition, which he cited as two examples of misconduct behavior. Dr. Velgach confirmed that would be a difficult question to ask. Dr. Bradley explained that there may be several military Compatibility composites, each optimized for separate outcomes. He said he agreed that data reinforce the idea that it is difficult to obtain consensus across the Services. Dr. Velgach said the Services took different approaches to answering the question, and a military compatibility working group is being established to discuss these issues further. Another committee member noted that the issue clearly does not stop at selection and asked about monitoring people real time using new technologies that provide information at the individual level. S/he countered, however, by saying that unit data may mean more, or be more indicative of importance, reinforcing the criticality of aggregating. S/he also asked if some TAPAS content was more relevant than other TAPAS content and commented on the use of traits (as measured by TAPAS) versus behaviors in respect to psychometric modeling and error. That is, in addition to traits as measured by TAPAS, is the team also considering behaviors that might impact the likelihood of future negative behaviors? Dr. Bradley noted one measure being the Dirty Dozen, twelve statements reflecting negative tendencies. He said certain responses to any of those items could be a flag or significant indicator; he reported not knowing whether that is true of the TAPAS. The committee member reflected that question is empirical and can be examined at the item level. Dr. Velgach wrapped up the discussion by commenting that, before making policy, they are planning to collect data to make sure they have a strong basis for recommendations. She said she wants to be very careful of how they approach this matter and the types of policies that may be developed to implement something like this.

8. <u>Complex Reasoning</u> (Tab L)

Dr. Kate Klein, HumRRO, presented the briefing.

Dr. Klein began by defining CR as non-verbal reasoning characterized by the ability to analyze visual information and solve problems using visual reasoning. She continued by noting that fluid intelligence has been found to be a strong predictor of training and job success and complex (non-verbal) reasoning is one element of fluid intelligence. 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 better prediction of training and job success, lower susceptibility to compromise, and increased qualification rates for non-native and non-heritage English speakers.

The objective of the current effort is to develop a CR (non-verbal) testing system to generate items for potential inclusion on the ASVAB. The system should employ non-proprietary AIG capability which will improve item development efficiency and reduce or eliminate field-testing requirements. It should also generate items with targeted properties that are similar to Raven's Progressive Matrices items and are of appropriate difficulty for qualifying applicants into jobs of varying complexity. Dr. Klein then showed examples of transformation items, which include various item features (i.e., types/orientation/size of

shapes, number of shapes, line weighting of shapes) and directions of transformations (i.e., vertical, horizontal, diagonal).

The CR development program involved three lines of effort. The first was to develop, pilot, and evaluate the initial CR capability. Based on the outcomes, recommendations were to use transformation items only, include four response options (with no "none of these is correct" option), and refine the item difficulty model and item selection to ensure appropriate level of difficulty and minimize group score differences by race/ethnicity, where feasible.

The objective of the second pilot study is to collect data on a refined pool of CR items representative of the population of items with a participant sample representative of military applicants. The goal is to collect sufficient data to evaluate group score differences on CR items and forms. The results will be used to develop CR forms for operational use on the ASVAB platform, select a pool of experimental items for potential use on operational CR forms, and inform future research and development efforts and test maintenance plans. This involved 24 CR items on three static forms, with each form containing the same items in a different fixed order, spiraled by estimated difficulty. Participants are also administered a pre- and post-test questionnaire seeking information on demographics, perceived item difficulty, and test-taking experience. The form includes two CR attention check items and items measuring insufficient effort in responding.

The sample will include non-military participants representative of military applicants—ages 18-35, U.S. citizens, high school degree/GED/less than one year of college. The target is 2,600 participants, or about 866 per form. The CR test is being administered on the Qualtrics platform with participants randomly assigned to one CR form. There is no fixed time limit and the time to completion is being recorded. Participants may use a desktop or laptop only.

Dr. Klein then presented a table showing the data collection status as of June 30, 2023 along with the number of participants in the first pilot who only received transformation items with four response options. A chart displayed score results for each form and the first pilot study. The number of participants in study 2 per form ranged from 451 (form C) to 472 (form A). Mean scores ranged from 14.86 (form C) to 15.59 (Form B), with standard deviations ranging from 5.18 (form B) to 5.64 (form C). The difficulty levels across forms were similar, with *p* values ranging from .62 (forms A and C) to .65 (form C). All values were somewhat higher than those from the first pilot test (mean score 13.89, standard deviation 5.12, average *p* value.58).

Dr. Klein concluded by presenting an overview of the steps involved in the third level of effort. These are develping CR test forms for operational implementation on the ASVAB platform, including four static forms of 24 items each presented in different fixed order spiraled by difficulty, along with a supplemental pool of experimental items. Future research and development and test mainteance plans for CR will also be developed.

At the end of the briefing, Dr. Klein asked the committee members if they thought the completion time should be reduced. She said retaining all 24 items resulted in completion time of 35 minutes at the 97th percentile. A committee member asked how many items would be removed. Dr. Klein said a range of 12 to 20 items would reduce the test to a length of 15 to 25 minutes. She said alphas were lower, but nothing below the 0.7 threshold. A committee member asked what percentage of test takers would complete all items in 35 min. Dr. Klein could not recall but said most completed the test in less than 20 minutes, with an average of 12 to 13 minutes. She said the 99th percentile was around 42 minutes. The committee member said there were no good solutions aside from just having a good rationale for the decision. For example, s/he suggested using a goal such as having 90% of people completing 90% of items; that is, you want them to complete as many items as possible.

Dr. Klein raised as a potential concern that Black respondents had longer completion times and performed better when they used more time. A committee member said she appreciated Dr. Klein raising that concern, because knowing how the test impacts different subgroups is part of the decision process. Dr. Velgach said time limits for other tests are set according to a policy that 99% of test takers complete the assessment. Dr. Pommerich said they are still establishing the test and do not want to close the door by making a hasty decision. She said they can fine tune the test, which is in a conventional format now, by turning it into a CAT format, and that will give them time savings.

Dr. Klein then asked if there are special considerations when making the transition from a static to adaptive form. A committee member asked if item analysis had been done yet, and Dr. Klein said no, only classical test theory. The committee member then recommended doing item level analyses of the data prior to September 2024 when the tests need to be available.

Dr. Klein then asked for the committee's thoughts on types of transformations. She said they tried to estimate prediction by item difficulty, but it got fuzzy when they examined the more difficult items. A committee member said to pay close attention to subgroup differences, because there is literature finding large sub-group differences on tests of fluid intelligence. S/he said Frank Bosco (Bosco, Allen, and Singh, 2015) administered the Raven and found *d* values of around 1, but another measure found lower differences. S/he said to keep on eye on it. Dr. Klein said the numbers they are seeing are lower than 1. Another committee member asked if there was a reason for using multiple forms, and Dr. Klein cited test compromise as the biggest reason. The committee member asked if there were significant differences across forms, and Dr. Klein said there were not. The committee member then asked about other outcomes, such as subgroup differences, perceived performance, and time completion, and Dr. Klein said all differences among the forms were negligible.

9. <u>Computational Thinking (CompT)</u> (Tab M)

Dr. Kimberly Adams, HumRRO, presented the briefing.

> Dr. Adams began the briefing by displaying text from the FY 2021 NDAA which mandated that a special purpose test be developed as an adjunct to the ASVAB to address computational thinking skills relevant to military jobs, including "problem decomposition, abstraction, pattern recognition, analytic ability, the identification of variables involved in data representation, and the ability to create algorithms and solution expressions." As stated in the bill, this was to be accomplished one year following its signing. This date has been adjusted to October 1, 2024 (FY2022 NDAA).
>
> Dr. Adams continued by indicating that a measure of computational thinking does not currently exist within ASVAB or other military testing programs, and the NDAA timeline does not support the creation of a new, valid measure. However, existing ASVAB/military tests potentially measure the six content domains underlying the computational thinking construct, providing a potential means to meet the October 1, 2024 deadline. She continued by outlining the objectives, assumptions, and considerations involved in this effort. The objectives are to develop a CompT composite score from existing ASVAB/military tests that can be used to inform enlisted decisions, to deliver the composite specifications to DTAC by 30 September 2023, and implement the composite by 1 October 2024. The assumptions are that the composite will complement other measures and data that the Services use during applicant screening (e.g., AFQT scores, medical/physical/conduct data), and predict first-term enlisted performance. Considerations include whether (a) the composite will be used in selection, classification, or both; and (b) the weighting of the CompT

domains will be overall or occupation specific. Considerations also include practical concerns regarding required platform modifications and testing time.

The approach taken to develop the CompT composite involved two phases—an alignment study and an empirical evaluation. The alignment study included eight steps. The first of these was to define the CompT construct domains. Dr. Adams displayed two tables listing each construct domain and its definition. The second step was to establish a composite development and validation strategy:

- Build an intercorrelation matrix among six CompT construct domains and a correlation matrix of CompT domains for ASVAB subtests/military tests based on correlation estimates from participating SMEs.
- Build a correlation matrix of all ASVAB subtests and military tests of interest based on observed empirical correlation estimates obtained from prior research or correlation estimates from participating SMEs when empirical data were not available.
- Use these data to simulate $n = 1{,}000$ "population" correlation matrices. These are multiple potential populations that reflect uncertainty due to variation in SME estimates.
- Calculate CompT criterion variable by applying unit weights to each CompT construct domain in each sample.
- Specify predication models. Ordinary Least Square (OLS) finds the regression coefficient that minimizes the sum of squared errors of prediction. Non-Negative Least Squares (NNLS) finds the regression coefficients that minimize the sum of squared errors of prediction when constraining the coefficients to be non-negative (positive). Lasso is a regularized regression model that performs variable selection.
- Run regression models using simulated predictor-criterion correlation data to calculate regression weights with 95% confidence interval.
- Compare models with regard to estimated prediction of the CompT construct.

The third step was gathering ASVAB and military test information and data. Dr. Adams presented a table that listed the tests of interest. These include all ASVAB subtests, the Cyber Test, the Coding Speed test, the MCt test, the CR test, and the Air Force's Electronics Data Processing Test (EDPT). Dr. Adams then presented a sample item from the ASVAB AO test and the Cyber Test.

The fourth step was gathering observed empirical data and SME data to support composite development. Empirical correlation estimates were obtained for the 91 subtest/military test pairings where data existed. A group of eleven Ph.D. researchers with expertise in cognitive ability, job performance relations, and job performance constructs were asked to estimate three sets of correlations: correlations among each of the 6 CompT domains, correlations for the 18 (out of 91) ASVAB subtest/military test pairings that were missing empirical correlation estimates, and correlations between the 14 ASVAB subtests/military tests and 6 CompT construct domains. Dr. Adams then showed a series of tables displaying the correlations described.

The fifth step was specifying the potential prediction models. Two prediction models were tested using OLS and NNLS, one including all tests and one including all tests except EDPT which is not currently on the ASVAB delivery platform with no plans to make it available through the platform. A data-driven selection of predictors was also run using Lasso regression to establish a parsimonious equation for estimating CompT composite scores. This started with 13 predictors (all tests except EDPT). There is the potential to run "constrained" Lasso models based on policy-type decisions, as appropriate (e.g., include a particular predictor to be included or excluded in the model).

The sixth step was to generate and evaluate composites from the prediction models. The ASVAB subtests/military tests that tend to remain in the model as the Lasso constraint becomes more stringent will be identified. Tradeoffs between the number of tests and level of prediction will be compared, taking into consideration such factors as whether tests within the most predictive model(s) are administered to all applicants, are used by all Services for selection/classification, are administered as part of both the ETP and CEP, and whether policy changes need to be considered. Dr. Adams then displayed a chart showing preliminary results of the analyses.

The seventh step is to deliver, integrate and implement composite. Target dates are: Deliver interim composite by September 30, 2023, integrate into ASVAB delivery platform between October 1, 2023 – September 30, 2024, allowing for operational implementation by October 1, 2024.

A final step (8) will be to evaluate the interim composite empirically. The CompT composite scores estimated from the analytically derived equation developed in the alignment study will be validated against CompT marker instruments using military applicants/recruits or a similar population. The CompT composite score will also be evaluated with respect to score distributions, subgroup differences, and other pertinent outcomes yet to be determined. Dr. Adams concluded by seeking committee members' perspectives regarding the potential tradeoffs between OLS (positive and negative) regression weights for predictors versus constraining regression weights to be positive (NNLS), and the tradeoffs between administering the CR test as part of the ASVAB or as a special test.

At the end of the briefing, a committee member asked if the tests being discussed (except for the EDPT) were tests that applicants were already taking. Dr. Adams responded that the Air Force administers the EDPT for some occupations, though it is not on the ASVAB platform. She said they included the test because they thought it might be related and added that the EDPT is also a battery of four tests, perhaps with an EI component. Mr. Andrew Deregla (USAF) said that was accurate so far. The committee member asked, if they are already taking these special purpose tests, like CR and EDTP anyway, why not use it?

Dr. Velgach said CR is not on the platform yet, though it will be. She said Services will participate in CR and other special purpose tests under consideration to different degrees, but that all applicants might not take it. Dr. Adams said it is a fluid situation and that is why they ran the all the different scenarios. If a Service chooses not to use a test, applicants will not have the score on CompT. Dr. Velgach noted the NDAA language describes CompT as a special purpose test usually used for classification, not selection; thus, not having a score for everyone may not be a problem. She said, based on current Service positions on which tests should be administered, not all applicants will have scores on everything.

Dr. Adams asked if the committee had any guidance on including Cyber Test, for the purpose of increasing face validity, given the decrease in multiple R. A committee member suggested that if the Cyber Test is deemed relevant, include it. S/he said the correlations represented a type of formative validity; each component has validity, but then there is the perception that they all have something in common. S/he said, though the practical validity may be based on something other than their perceived commonalities, do not worry about the difference between .79 and .71.

The committee member then commented on the accuracy of SME ratings of correlations, stating a concern regarding overreliance on SME estimates and the need to concentrate on objective, empirical data. Dr. Adams clarified that an empirical evaluation will be conducted on the CompT composite model(s), but this first phase of the project was to conduct an alignment study to identify the viable model(s) to evaluate. She also explained that the alignment study included empirical estimates of correlations between the various ASVAB and special tests, with the exception of 18 test-test correlations. She said exceptions included MCt and Coding Speed with Cyber Test and EDPT, Cyber and EDPT, and the new CR test with all 13 tests included in the study. For these 18 test-test correlations without empirical data available, the SMEs provided estimated correlations. The SMEs also provided correlation estimates for each test with each of the 6 domains of the CompT composite. The average test-test and domain-domain

intercorrelation estimate was calculated across SMEs. The estimated test-domain correlations were averaged across the SMEs and then averaged across the 6 domains to get a test-CompT construct correlation estimate. All these empirical and estimated correlation estimates were used to simulate 1,000 "population" correlation matrices. The CompT composite model(s) specified from the alignment study will undergo an empirical evaluation.

The committee member replied that the more empirical data they could collect, the better. S/he said ratings of correlations are not correlations but only represent what the raters think. The committee member said, when giving the Scholastic Aptitude Test (SAT), it is the verbal test that predicts engineer performance; accordingly, there may be high face validity (e.g., for math), but negligible real validity. The committee member said the team is dealing with a serious challenge and doing well.

Another committee member asked – regarding face validity – if the team perceived the Cyber Test to be part of CompT; s/he suggested the question is whether it should be there. Dr. Velgach said one subdomain specifically related to creating algorithms, and solution expression is measured within the Cyber Test. The committee member said that CompT is more than just the Cyber Test, but there seems to be a component related to the Cyber Test. Dr. Adams said she would assume they would move forward with a composite score with the Cyber Test but asked about the issue of weighting. She said it is simpler to have unit weighting, but maybe it does not matter. A committee member asked Dr. Adams about her perception of the accuracy of the SME estimates. Dr. Adams said there was not a lot of variance in the ratings among SMEs, and that is the best indicator they currently have. She said the SMEs appeared to understand the instructions based on their questions and comments.

Another committee member noted the correlation matrices and asked whether similar constructs are being measured. Dr. Adams said that was all the data they had. The committee member then asked about using the Pearson *r* correlations based on observed data. Dr. Adams said she could not speak to that and that the Services provided the numbers but not how they got the empirical observed score correlations. The committee member said that if the correlations are low because of measurement error and there is a correction method that would be more informative (i.e., a better statistic on the similarity of constructs), it might be good practice to look at dis-attenuated correlations, which may change some of the results. S/he said the important factor is that the correlations between each test and the CompT construct are not the actual observed correlations.

Another committee member returned to the topic of weighting, asking how much it mattered, that is, if the difference was only between .74 and .77, given the intended use: classification. S/he said it is a hard game to play. Dr. Adams said the Phase 2 Evaluation Study will look at that, but they have to specify the composite score ahead of time. For Phase 2, if they can collect data that allow options/flexibility (i.e., evaluate more than one composite model), they would have the option to explore the impact of each (i.e., composite predicts performance on a computational thinking marker instrument). She also said conducting the evaluation study on the applicant population would be preferred, because it would remove threats to validity and provide data on ASVAB/military tests, which would allow the evaluation of various composite models. She said it also may provide an opportunity for future research (e.g., addressing the impact of the current situation in which a composite must be identified immediately). For example, although recruits

may make the cut score for certain jobs with the composite, they would not have evidence of how successful those recruits will be in training until those data are collected. Dr. Pommerich said she hoped to be able to fine tune the composite later with more data, adding that this is uncharted territory to have Congress say to develop an assessment by a specified date. A committee member replied that he respected the timeline issue and that the ongoing work is an admirable effort. Dr. Velgach said, if the initial composite were to be implemented, they can refine it with more data and better information on best use.

Slide 26 provides the results for the ordinary least-square regressions (OLS) as well as the non-negative least-square regressions (NNLS). Dr. Adams shared that during the July 2023 MAPWG meeting, the Services indicated that explaining negative coefficients to laypeople is more difficult, so they prefer to use NNLS (i.e., only have positive weights in the composite equation). Dr. Adams asked if the committee had thoughts on the use of OLS (positive) regression weights versus constraining regression weights to be positive (i.e., NNLS). A committee member commented that the composite is insensitive to whether they are correlated. Due to multicollinearity, OLS can produce negative weights. Dr. Adams said they looked at one construct domain at a time and CR was highly correlated with all 6 domains (as shown on slide 22, it ranged from 0.53 to 0.64 with an average CR-CompT correlation estimate of 0.57). The committee member replied that, out of the tests examined, some had a higher correlation with the CompT construct than others. As a result, s/he said having an equal weight may not be appropriate.

Addressing the tradeoffs between administering the Complex Reasoning (CR) test in the battery versus as a special purpose test on the platform (slide 30), Dr. Adams said a test administration time of 35 minutes would be outside of ASVAB testing time, and that would mean less use by the Services. She asked what tradeoffs the committee members have seen in other testing programs and did they have any advice on this matter. A committee member said a lot of his/her papers have dealt with reducing measure length and it is a matter of balancing (i.e., measuring one construct and going deep, or multiple constructs and going wide). S/he said it involves balancing more than construct coverage, but also reliability, validity, and subgroup differences. The committee members said they were interested in seeing how this plays out.

10. <u>Public Comments</u>

After the end of the first day of presentations, Dr. Velgach opened the floor to public comments and asked participants to limit their comments to no more than 5 minutes per person. There were no comments.

11. <u>High School Curriculum Study</u> (Tab N)

Dr. Peter Ramsberger, HumRRO, presented the briefing.

> Dr. Ramsberger began by presenting the goals of the high school curriculum study which are to design research to determine how ASVAB subtests align with content taught in high schools, explore how ASVAB content is taught, and map ASVAB content to other relevant sources. The study design should include (a) a review of previous high school curriculum and high school assessment alignment studies with ASVAB content, (b) a review of previous mappings between ASVAB and other tests, (c) a review of any

available National Assessment of Educational Progress (NAEP) transcript studies, and (d) a method for assessing if there are differences in course-taking behavior patterns between military applicants and the general high school population.

Dr. Ramsberger continued by providing an overview of current trends in teaching practices. The development that had the most significant potential impact on educational approaches in the past 20 years was the introduction of the Common Core State Standards (CCSS) for English Language Arts and Mathematics in 2009 and the Next Generation Science Standards (NGSS) in 2011. The common core recommended an emphasis on complex texts and writing assignments that called for the use of evidence to support arguments. In regard to math, the goal was to encourage teaching practices that support gaining a conceptual understanding of underlying principles. Two consortia were established to develop assessments that align with the Common Core—the Partnership for Readiness for College and Careers (PARCC) and Smarter Balanced. The Common Core was initially adopted by 46 states—Minnesota only adopted the English Language Arts standards. In the interim the standards have been altered or replace by several states. The NGSS place an emphasis on developing an understanding of core underlying principles, using that information to generate and apply models to explain various phenomenon, and treating science as a progression that builds throughout a student's time in school.

Dr. Ramsberger continued by summarizing results of studies examining the impact of the CCSS.

- A survey of teachers found that the majority reported employing instructional practices recommended by the Common Core. The survey data were analyzed in conjunction with PARCC and Smarter Balanced test results and a positive relationship was found between student math scores and teachers' self-reports that they had been observed and coached and received professional development.
- Another study found a small, positive relationship between adoption of the Common Core and NAEP math and reading scores.
- However, and additional study found a negative relationship between Common Core adoption and 4th grade reading and 8th grade math NAEP scores.

There have also been several studies that examined the impact of the NGSS.

- Results from a survey of science teachers and school administrators indicated that most districts had taken steps to implement the standards, with more progress being made in elementary and middle schools. Some issues being confronted were a lack of instructional resources and credentialed teachers. There was also a negative impact of COVID, during which adopting the standards took a back seat.
- Other studies relying on teacher and administrator self-report data found an increase in the quality of science learning and student engagement.

Dr. Ramsberger next addressed other trends in teaching practices including integrated instruction, where content is blended within and across disciplines. Several states have adopted some form of integrated instruction in science and math. Here, too, there are reported difficulties finding qualified teachers and supporting materials. A meta-analysis of evaluation studies examining integrated science, technology, engineering, and mathematics (STEM) instruction found mixed results, with positive outcomes more likely at the elementary school level. Issues include the fact that the approach is often not aligned with standardized tests, difficulties in implementing teacher collaboration, and finding instructional materials that are geared towards an integrated approach.

A review of evaluations of integrated instruction did find some positive outcomes, but again there were issues with finding teachers with cross-disciplinary qualifications.

Other trends in teaching practices include:

- Identifying and applying learning progressions, which starts by specifying the ultimate learning objective and moves backward to identify all the required prerequisites. The Department of

Education funded a study in which SMEs identified seven reading and writing strands and six major math strands.

- Microlearning involves breaking instructional material into small chunks and incorporating assessments throughout to ensure that students understand fundamental content before moving to more complex content. A recent review found this was largely driven by mobile technologies and is found more often in higher education.
- Flipped instruction moves the introduction of content outside the classroom so that class time can be spent discussing and developing an understanding of it. One study in which flipped instruction was implemented in a chemistry class found that students in the experimental group actually performed less well on the final exam.
- Project-based instruction involves having students, individually or in groups, apply what is learned in the classroom and what they discover through their own research to develop solutions to real-world problems. In one study a set of high school economics teachers was provided professional development in implementing project-based instruction. Students in the project-based classrooms outperformed those in the control group on end-of-course tests.
- An NCES-funded study of the use of technology in the classroom found that 47% of schools reported using technology based instructional materials to a moderate or great extent, and 84% of schools indicated that technology was being used for activities normally done in the classroom, with 54% suggesting that the activities would not be possible without employing technology.

Dr. Ramsberger continued by addressing the implications of this work for the ASVAB. He noted that, given the decentralized status of public schools, keeping up with various trends would be difficult. For instance, some states adopted the Common Core and then later abandoned or amended them, and New York moved to implement an integrated math curriculum, but later switched back to a traditional format.

Perhaps the biggest implication may be in the way knowledge is assessed. A recent comparison of ASVAB and Smarter Balanced math items found that the latter required students to demonstrate skills in a more diverse and language intense context. Smarter Balanced items often involve lengthy passages with multiple questions related to each. For instance, identify an inference that can be drawn from a passage and then select the portion of the text that supports your answer. Smarter Balanced items also often involve open-ended questions.

More complex item types could be added to the ASVAB. Examinees could be presented a passage that offers a particular point of view on a topic, with the instruction being that it must be shortened. The examinee is asked to identify the most critical points and arrange them in a coherent manner. However, this would involve challenges. If open-ended items are incorporated into the ASVAB, it would require a valid and reliable automated scoring system, given the volume of testing. It is likely that item development costs would increase, and significant programming efforts would be needed. Additionally, there is the possibility that testing times would increase.

Dr. Ramsberger then turned to prior ASVAB alignment studies. A 1997 study focused on GS and the technical tests. Researchers examined 1990 high school transcript data and conducted an exposure-to-content survey of recruits. Both sources indicated a higher level of exposure to GS content than the technical tests. The survey results suggest that the recruit sample was technically better prepared for military training, which was attributed to a selection effect. Also surveyed were military SMEs, most of whom were found to judge ASVAB content to be relevant to military training.

A 2015 investigation compared the ASVAB test blueprints with other relevant assessment programs, such as NAEP, Scholastic Aptitude Test (SAT) and American College Testing Test (ACT). Researchers found there was a good deal of overlap between them, particularly the non-technical tests. They used the results to generate more detailed taxonomies for the ASVAB subtests, which they felt could increase the breadth of the subject matter covered.

The results of this research and a more recent replication of the military SME survey regarding ASVAB content indicate that the ASVAB science and technical tests are relevant to military training and jobs. Although overlap between the content of the non-technical tests and other assessments was found in the 2015 investigation, there was less overlap for the technical tests.

Dr. Ramsberger then discussed studies examining high school course taking behavior. These largely fell into one of four broad categories: (a) course-taking behavior and changes in course taking over time, (b) the impact of course taking on future outcomes, (c) changes in and the impact of Career and Technical Education (CTE) course taking and (d) methodological studies. Much of the research is based on NCES-funded studies, including the high school longitudinal studies (HSLSs) and the high school transcript studies (HSTSs).

Overall, the results suggest that, over time, students were earning more credits and pursuing more challenging curricula. However, there is evidence that course titles may not accurately reflect content. In one study SMEs reviewed textbooks and rated their content. They found that 73% of students who took an honors algebra class and 62% of those who took an honors biology class actually received instruction at the intermediate level. Another study rated the curriculum students received and found that only 12% were found to be rigorous, while 23% were below standard.

In regard to the impact of course taking, several studies have found that students who do well in middle school math and science are more likely to take advanced classes in high school. Further, students who take Algebra 1 before 9th grade are more likely to go to a 4-year college than those who take it in a later grade.

Studies of CTE course-taking indicate that most high school students earn at least some CTE credits, although the number of credits has declined over time. CTE course-taking patterns have also shifted, with less focus on fields such as agriculture and business and more on engineering, technology, health care, and hospitality. There have been consistent male-female differences in CTE course taking, with more males earning credits in areas such as architecture, construction, engineering, and more females earning health care and human services credits. Some differences have diminished over time, for instance business and marketing. Longitudinal studies suggest that high school graduation rates among CTE course takers have risen, and the limited data suggest there is no relationship between CTE course taking and attending post-secondary institutions.

Dr. Ramsberger next addressed methodological studies related to course taking and course outcomes. One such study examined HSLS 2009 data and found that self-reports were generally accurate regarding courses taken, although less so when it came to when they were taken, and grades received. Students getting higher grades were more accurate in their reporting. A 2020 NCES study compared student self-reports on courses taken with high school transcripts and found that, overall, a higher percentage of students reported taking math classes than was indicated by their transcripts.

Several approaches are being taken to achieve the goals of the current research. One is to explore HSTS 2019 data to see if there are relevant findings that have not already been reported. The alignment work done in 2015 is being reexamined to see if there have been shifts in the sources used that indicate a greater or lesser alignment with the ASVAB. Another type of alignment study will be conducted in which course catalogs from a sample of high schools across the country will be collected, and SMEs will be asked to review ASVAB test blueprints along with relevant high school courses and make judgments regarding the degree to which the ASVAB content is covered. The course descriptions often mention instructional methods, so SMEs will also be asked to indicate if particular methods are used, for instance integrated approaches, project-based learning, or technology-based learning. Dr. Ramsberger then showed sample pages from a randomly selected high school course catalog.

A final approach being taken is to include a question in the Futures Survey conducted by the Joint Advertising Market Research and Studies (JAMRS) branch, asking respondents to indicate courses taken. These results can then be compared with data from the 2019 HSTS. Analyses can also be run to compare results for respondents who indicate a propensity for enlisting to those who are not propensed to see if there are differences in course taking. Another question will focus on extracurricular activities that may be relevant, such participation in clubs or special interest groups. Assuming space is limited, this may have to be on a subsequent survey. Dr. Ramsberger then showed a mockup of the question regarding course taking and extracurricular activities.

> The sampling plan for the course catalog portion of the work involved (a) randomly selecting one state from each of the nine Census regions; (b) creating an extract of data from the Common Core of Data for each state that lists all schools in each state; (c) sorting the schools by level and eliminating Pre-K, elementary, and middle schools; (d) sorting schools by type and eliminating special education, unknown, and alternative schools; and (e) generating random numbers to select five schools from each state. The results of this process led to an underrepresentation of City/Large schools given that three of the selected states had no City/Large schools. As a result, one City/Large school was randomly chosen from the other four. The websites for the selected schools were reviewed for course catalogs, which were found in 30 of 49 cases. The schools that did not supply catalogs typically were quite small. Additional samples within the state/size jurisdiction groups were drawn until course catalogs were located. This could mean that smaller schools will be underrepresented in the sample.
>
> Dr. Ramsberger concluded by updating the committee on the current status of the project. NCES has indicated that HSTS:19 datafiles will be released by the end of September. The review of the Waugh et al. report, and sources used has begun. Ratings materials for the alignment study are being created and SMEs identified. Discussions are underway with JAMRS. Given limited space on the survey, the questions are being revised to include only courses and activities that are likely to have variance in terms of participation and some relation to the ASVAB.

At the end of the briefing, a committee member asked, what do you do if there is a gap between what high schools are covering and what the ASVAB is covering – that is, what the Services need? Can the Services influence curriculum, or do they have to adapt? Dr. Ramsberger said the question is really whether the content is available and if students take the classes. If the content is not available, then that could be problematic; if it is available, then it is a question of the frequency at which courses are taken. Dr. Velgach said the Services could use preparatory courses to close gaps in skill development across schools, but that process has not been fully implemented or evaluated. Another committee member said there are courses listed in catalogs that may or may not be offered. S/he said students can take classes such as environmental science, but it depends on whether a teacher is available. S/he said that may be impacting the study's results. That is, "taking" versus "listed" may provide different results. S/he said it also may be important to know why some courses are taken and others are not, and that may be informative for designing the NextGen ASVAB. The committee member then asked several questions about the timeline for the NextGen ASVAB, to include how the current study fits into it, as well as how do the Computational Thinking and CR tests fit. S/he asked how it all comes together and if that could be addressed in the next meeting.

A committee member thanked Dr. Ramsberger for the briefing and expressed appreciation for hearing about the reality of course offerings across schools. S/he said a moderating variable is the availability of educational resources and asked what type of multilevel analyses on the school, state, or other level could inform those differences. The committee member suggested a hypothesis might be that schools with more resources could provide more courses, and though that is not a sufficient condition, it is a necessary condition. S/he then made another point: in the broader context, aligning skills with the ASVAB could be used in a promotional sense, as well as a way to assess individual schools from a workforce development perspective. S/he said, in Texas, there is a big push to have graduates who are employment ready and then mentioned the importance of considering diversity, equity, and inclusion in making sure school to work transitions are successful. Another committee member asked if schools with the right curriculums and resources offered better recruiting environments than poorer education systems. Dr. Ramsberger said the two top states for recruiting are Florida and Texas. He said they

obtained a sample of catalogs from those states to see if they had courses more aligned with the ASVAB and will look at this. He also said they want to have SMEs look at course descriptions and identify content that is not covered by the ASVAB currently. Dr. Helland said they looked at estimated aptitude scores and state propensity and eligibility rates, and Florida and Texas had more propensity but lower eligibility (i.e., lower estimated aptitude scores). Dr. Ramsberger said that JAMRS data indicate that higher propensed students are less likely to take baccalaureate courses than less propensed students.

12. <u>Non-Native English Speakers Analysis</u> (Tab O)

Dr. Bill Walton, HumRRO, presented the briefing.

Dr. Walton began the presentation by stating that the objective of the study was to address concerns raised in a conference report accompanying the FY 2020 NDAA that potentially high-quality recruits were being denied entry into the miliary because they do not speak English as their native language. HumRRO (a) examined practices in civilian education regarding English Language Learners (ELLs), (b) surveyed best practices in English as a Second Language (ESL) instruction, (c) reviewed past efforts to recruit NNES, and (d) conducted analyses of existing data to investigate various aspects of the issue.

Dr. Walton continued by discussing assessment of ELLs in the civilian academic sector. English Language Learners are defined in the *Every Student Succeeds* Act of 2015 as students whose native language is not English and whose level of English fluency is low enough to make it difficult to achieve success in school and society. States have individual procedures for identifying ELLs, typically involving brief screeners. The home language survey assesses whether students come from an environment where a language other than English is present or prevalent. Full assessments establish English Language Proficiency (ELP) across the entire scale. Dr. Walton cited several commonly used assessments.

Dr. Walton continued by presenting tables showing the number of English Learner (EL) students enrolled in public elementary and secondary schools and their percentage of total enrollment over the years 2000 to 2017. Another table presented the number of EL students by grade in 2017 and their percentage of enrollment in that grade. Dr. Walton pointed out that the percentage steadily declines from kindergarten (15.9%) to grade 12 (4.6%) as students ELP increases and they leave EL status. Additional tables presented the most frequently spoken languages in 2017, with Spanish being spoken by nearly three-quarters of ELs. By state, California (19.2%), Texas (18%) and Florida (10%) had the highest percentage of EL students in 2017-2018.

Dr. Walton then presented a chart showing the percentage of 4^{th}, 5^{th}, 6^{th}, 7^{th}, and 8^{th} grade EL students scoring at various levels on the English Language Arts (ELA) portion of the California Assessment of Student Performance and Progress (CAASPP). At each grade level, the majority of students failed to meet standard. Another table presented the percent of U.S. public school students scoring at the NAEP Basic level or above in 12^{th}-grade reading by EL status in 2015. Over three-quarters of EL students scored below basic compared to just over one-quarter on non-EL students. In contrast, while nearly a third of non-EL students scored at the proficient level, only 4 percent of EL students did so.

Dr. Walton continued by providing lists of the accommodations offered to EL students when taking NAEP and CAASPP. These include bilingual dictionary without definitions, extended time, and translated test directions. Another table showed the various ELP assessments used by states in 2019-2020, with 36 states employing the WIDA ACCESS for ELLs. The content covered by ELP assessments can include reading, writing, listening, and speaking. Dr. Walton then discussed the assessments used by colleges and universities to assess the ELP of international students applying for admission. He pointed out that the number of students using an F1 or J1 visa to study in the U.S. more than doubled from 1990 to 2014, reaching a total of 1.1 million in the 2016-2017 school year. A table summarized information on the most commonly used assessments and their characteristics.

Dr. Walton then presented charts summarizing best practices in EL instruction for young adult ELs and for adult ELs. For young adults, these include developing English skills and vocabulary as part of subject matter learning, providing opportunities for extended discussion of text meaning and interpretation, and providing small-group instructional support for struggling students. For adult English learners, best practices include providing courses of varied intensity and duration with flexible schedules, stressing the importance of interaction with peers and others, and providing ongoing opportunities for language assessment to measure progress and provide motivation.

There is no direct measure of the size of the recruiting market that are NNES. However, NCES tracks the number of ELLs in American schools, and these data can be cross-referenced with student population census data to get an estimate of the percentage of students who are ELLs, by race, ethnicity, and age range. Dr. Walton showed a table presenting these results which indicated that approximately 4.75 percent of 11th and 12th grade 16–19-year-olds are ELLs, with the highest percentage of these (16.6%) being Hispanic. These findings were examined in conjunction with data from the JAMRS Futures Survey, which assesses the propensity of youth to join the military and also collects race/ethnicity information. Dr. Walton showed a table indicating that, overall, less than 1 percent of students in the 11th and 12th grades would be NNES and propensed to join the military.

Dr. Walton then turned to results of analyses of ASVAB data. He noted that studies conducted to identify the most common reasons that applicants do not qualify to serve show that only about 2 percent are disqualified based on AFQT scores alone. An examination of non-qualifying scores by race/ethnicity mapped to the percentage of NNES within each racial/ethnic group revealed no clear patterns. For instance, while an estimated 16.6 percent of 11th and 12th grade 16–19-year-old Hispanics are ELLs, only 4 percent of Hispanic applicants had an AFQT score below 10. Dr. Walton then showed charts presenting subtest scaled scores for the AFQT tests and AO by racial/ethnic group, which again demonstrated no clear pattern suggesting ELP as a factor in test outcomes. Another set of charts displayed completion times for these same subtests by racial/ethnic group, with no indication that those groups with likely higher percentages of ELLs were taking more time to complete the subtests.

Another approach to examining ASVAB performance and ELP is to examine performance of applicants who are U.S. citizens versus those who are not, with the assumption being that a higher proportion of the latter group would be NNES. Dr. Walton then presented tables comparing citizens and non-citizens on (a) AFQT percentile and total time to complete the AFQT tests and (b) scores on the individual AFQT tests and AO and time to complete each of these tests. Again, the results showed no patterns that suggested non-citizens were disadvantaged in taking the ASVAB. Finally, Dr. Walton presented results of an analysis of non-qualifying AFQT scores to identify the percentage of test takers who had MK and AR scores above 31 and Verbal Expression (VE) composite scores (WK + PC) below 31 to identify the population for which verbal ability was the barrier to qualification. For the years 2015-2019, this percent ranged from 0.08 to 0.14, indicating that verbal ability alone played a very small role in determining ineligibility to serve.

Dr. Walton next discussed past effort to recruit NNES. He indicated that interest in doing so increased in the early 1980s in the face of a shrinking recruiting pool. One study examined three Navy ESL programs, which varied in their characteristics and outcomes. One effort established the English Technical Language School at Camp Santiago, Puerto Rico. Fewer than half of those attending achieved a score of 70 or above on the English Language Comprehension Level (ECL) test following training. A more successful program at the Defense Language Institute English Language Center (DLIELC) at Lackland Air Force Base, TX was self-paced and individualized and incorporated a secondary emphasis on military training. The Verbal Skills Curriculum instituted at recruit training centers in Orlando and San Diego had positive outcomes but limited capacity. An Army program targeted to NNES, also conducted at DLIELC, provided up to 24 weeks of residential instruction for those scoring less than 70 on the ECL test. Findings suggested that those with lower entry scores exhibited the largest gains in proficiency.

The Army's Foreign Language Recruiting Initiative (FLRI) began in 2002 as a 2-year pilot program. It was originally targeted to Spanish speakers but broadened to include all native languages. The criteria for entry include having an AFQT score in the IVA range, scoring between 40 and 74 on the ECL, and having a score of 54 or above on the ASVAB AO test. The 8-24-week training was conducted at Fort Allen for Puerto Rican

recruits and DLIELC for others. To graduate, trainees must achieve a passing score on the ECL or the American Language Course Placement Test. An evaluation of FLRI was conducted between 2006-2010. Approximately 91 percent of participants graduated from ESL training. Fort Allen graduates had higher average score gains (18.4 points) than did DLIELC graduates (4.9 points), which was attributed to participants at Fort Allen also taking a General Technical Proficiency Course. Between 2004 and 2008, the 12-month attrition rate for FLRI graduates was 13.2 percent, which was comparable to overall attrition rates. The three-year attrition rate for FLRI graduates was 19.5 percent, less than the overall figure of 29-33 percent. Dr. Walton then showed a table displaying the number of FLRI active duty, reserve, and guard participants from 2003 to 2019. The highest number was 621 in 2011, with the 257 participating in 2019. Dr. Walton concluded that, even though the program is successful, it will not result in a large number of new accessions.

An additional remedial program was run by the Navy under different guises starting in World War II. Most commonly known as Fundamental Applied Skills Training (FAST), it involved a two-week course on literacy and a three-week course covering verbal skills. The ASVAB VE score was used to identify participants. Evaluations found that, controlling for education and AFQT, FAST graduates were anywhere from 1.9 to 2.88 times as likely to advance to E-4 within three years and had lower training attrition and first-year attrition rates. However, the program was suspended in 2014 due to diminishing enrollment and minimum return on investment.

Dr. Walton continued by summarizing lessons learned based on past experience and best practices. ESL and other remedial training programs can be effective and are more so when targeted to an individual's initial skill level and are flexible to adjust to different rates of progress. An emphasis on conversational English with opportunities for peer interaction and small group instruction also enhance effectiveness. However, without a significant investment of resources, ESL programs will not result in large numbers of recruits. A review of the English language training provided at DLIELC indicated that it incorporates many of the best practices recommended based on past research and experience, including individualized and small group learning, frequent assessment of progress, and individual remediation where needed.

Dr. Walton then turned to potential methods for screening NNES who may be effective Servicemembers. These include screening for ELP prior to testing content and cognitive domains and following the Army's procedures for identifying applicants who may have failed to qualify due to English comprehension issues (i.e., ECL score between 40 and 70, AFQT Category IVA, and AO score 54 or above). A working group of testing accommodation and military testing professionals could consider the implementation of accommodations for NNES by addressing questions of whether such a move would necessitate an equating study, what the implications would be in terms of the quality of accessions and the costs involved, how this would impact the CEP, and would this open the door for offering accommodations to other groups.

Dr. Walton concluded by summarizing the outcomes of this work. It does not appear that the ASVAB is screening out large numbers of potentially qualified recruits due to their not being native English speakers. Although ELS instructional practices employed by DoD seem to comport with best practices in the field, there are questions about the ECL test concerning whether it taps language skills relevant to military service and whether it is a comprehensive measure of required language skills.

As Dr. Walton briefed on the size of the NNES population (slide 16), a committee member asked him to restate the goal of the research. Dr. Walton said it included multiple pieces, but the issue was trying to access more NNES into the military in general, as well as how to mitigate related issues. He said one aspect of the research was to determine how many people fall into that NNES category and could be part of the recruiting population. He said because there was no direct measure, they got creative by drawing on the two data sources, NCES and Current Population Survey (CPS).

After Dr. Walton discussed the percentage of VE-driven non-qualifiers (slide 22), a committee member said it was very informative to know that VE scores were responsible for disqualifying

so few applicants and that it is helpful to know how many students are really being caught by the cut score. Dr. Walton said there were similar results with AO. Another committee member asked if only 2% are disqualified because of their AFQT scores alone, and Dr. Walton said yes. The committee member then asked if there was evidence that inability to speak English is related to other factors that disqualify them. Dr. Helland responded that 44% are ineligible for multiple reasons, however, overlap more frequently occurs between medical and drug use rather than medical and AFQT scores. She said those data can be pulled, and current data shows that it is now only 1% of applicants who are disqualified for AFQT scores only.

A committee member complimented the work, saying "nicely done." S/he mentioned seeing a published paper that looked at native/nonnative academic professionals in science fields, which found that non-native speakers spend less time with activities and are less desirous of presenting at conferences. The committee member said it looked like the team had done the best they could, given the lack of available information. S/he said it was interesting to learn that Latinx persons are more likely to join the military than most in the overall population.

Another committee member remarked that the presentation was "tremendously interesting" and said she wanted to connect the findings to the NextGen ASVAB effort. S/he asked, what level of proficiency is sufficient for service? Saying that was the key question, s/he asked how Dr. Ramsberger's earlier comment about a common core course would fit in. S/he also said there was a criterion issue to solve, and that it would be important to identify the specific skills that must be performed in English. The committee member also noted the existence of a fairness issue, and that is, does everyone have the opportunity to demonstrate what they could do if selected for service. S/he also suggested avoiding the establishment of accommodations for NNES; that is, it is not about providing advantages that others are not receiving but giving them the opportunity to demonstrate what they can do. S/he asked if there are measures that could be taken to help applicants improve their skills sufficient to prove they can be successful.

Another committee member said the presentation was excellent and commented on how to think about ESL in the military context. Specifically, s/he said having a better grip on work-relevant language may lead to more efficiency in learning. For example, technical language or idioms relevant to various MOS are likely not included in formal instruction prior to service. S/he suggested thinking about job redesign and technology, especially the use of translators and artificial intelligence (AI) tools. The committee member also mentioned that assignment geographical location may offer a point of leverage. Dr. Walton said they are working with language centers, but the task is difficult: identifying criteria and how to measure it, as well as how to train the skills is all very challenging.

Dr. Velgach concluded the discussion by explaining that the analysis that Dr. Walton had briefed was requested by Congress and the feedback they received from Congress was that this report was one of the best reports they have received. She said Congress believes they got the answer to their question: Are we missing out on potential recruits? The answer appears to be "no."

13. <u>ASVAB CEP Update/Demo</u> (Tab P)

Dr. Irina Rader, OPA/DTAC and Ms. Kate McLean (Written, LLC), presented the briefing.

Dr. Rader began by stating that DoD sponsors the ASVAB CEP at no cost to schools with the mission of increasing participant exposure to both civilian and military career options, providing quality leads to military recruiting services (if student information is released by the participating school), and enabling 11thgrade and above students to use their scores for enlistment up to two years after taking the ASVAB. DTAC executes the program's technical development, maintenance, and evaluation, while the United States Military Entrance Processing Command (USMEPCOM) administers the program.

Dr. Rader continued by displaying a table showing participation rates and recruiting leads for the 2018-2019 through 2021-2022 school years. The latest figures show a rebound from the COVID-19 shutdowns, with 631,045 students in 13,224 schools taking part in the program, leading to 504,114 leads provided to the military Services. Additional charts showed the number of P&P forms administered and the number of CEP *i*CAT tests taken, demonstrating a continued growth in computer-based testing (111,728 tests in 2021-2022 versus 72,299 in the pre-COVID period of 2018-2019). Dr. Rader then presented tables showing the number of accessions based on CEP ASVAB scores from 2016-2017 through 2021-2022, by Service and overall, as well as website utilization numbers from July 1, 2022, through June 30, 2023 (e.g., users, returning users, page views) for both the ASVAB CEP and Careers in the Military (CTM) sites. She also presented numbers of inquires received through both websites in the 2022-2023 school years, including score requests, requests to bring the program to a school, and Service-specific inquiries through the CTM.

Dr. Rader then turned to areas of focus for the program in the 2023-2024 school year. These include (a) technology modernization and ensuring DoD compliance regarding online activities, (b) growing propensity, (c) workforce multipliers by expanding training for individuals involved in administering the program, (d) monitoring state legislative activity and expanding state contacts, (e) focusing on underserved populations to include participation in DoDEA schools, and (f) improving occupational data and content capture and maintenance. Dr. Rader noted that 35 states have initiated some form of college and career readiness mandate, and that some states have either mandated the use of the ASVAB by making it available to all students or authorized the use of the ASVAB as an alternative graduation credential. Other uses include an option to meet graduation requirements and serving as a military and career readiness indicator. Dr. Rader then presented charts summarizing the status of the ASVAB CEP by state.

Post-test interpretation (PTI) sessions are conducted by qualified professionals to help students understand the meaning of their ASVAB scores and capitalize on the resources available to use those scores in conjunction with their Find Your Interests (FYI) inventory results to explore potential civilian and military careers. The PTI proficiency training standardizes the way in which the sessions are delivered and serves as a workforce multiplier by employing a train-the-trainer model. Various metrics are used to gauge the success of these efforts, including increased testing numbers and website traffic.

Dr. Rader then presented a chart showing 2023 marketing events attended by CEP representatives and other efforts at stakeholder engagement. She stated that the overall marketing goal is to reach 1 million participants in one academic school year, resulting in 500,000 leads to the military services. This will be achieved by improving the program's reputation through contact with target audiences, gaining insight into users' needs, and establishing thought leadership through presentations and training. Brand awareness will be built through content marketing, improving website performance, advertising, and effective use of social media. Strategic search engine optimization, content marketing, and enhanced website performance drive traffic and build brand awareness to increase testing participation and leads to recruiters. Dr. Rader showed charts demonstrating organic website traffic growth and participation and leads over several years. Dr. Rader next presented information on the number of CTM page views and the top jobs about which there were inquiries by Service.

Dr. Rader stated that integrated strategic marketing represents a comprehensive strategy that leverages multiple channels to get the most out of investments aimed at achieving increased participation. This includes

owned media (e.g., website, email marketing, training, and events), paid media (e.g., social media ads, conference ads and exhibits), and earned media (social mentions, testimonials). She then listed several paid media sources, along with their intended audience and messaging. The briefing concluded with a demonstration of the enhanced ASVAB CEP website and its new features.

During the briefing, a committee member asked for clarification in regard to the percentage of schools that request PTIs. Ms. McLean said about one third of schools request PTIs, but that they want that number to be 100%. Schools have to request the service, and the program must do a better job explaining the benefits. Another committee member asked how the CEP facilitates recruiting to military academies. Dr. Velgach said there is information on the site about the academies and how to apply. That information is provided from a post-secondary education perspective. She said the ASVAB score is not used to apply for academy admission.

At the end of the briefing, a committee member said the website demonstration helped put things in perspective. Another committee member noted that the number of leads provided to the Services was a much smaller number than the number of students participating in the CEP (slide 7) and asked what constitutes a lead. Dr. Velgach said each school makes the decision about whether scores can be provided to recruiters, which is one reason the number of leads is lower than the number of participants. Mr. David Davis added that, though 10^{th} through 12^{th} graders are tested, only 11^{th} and 12^{th} graders can be classified as a lead. Two groups of participants are excluded: 10^{th} graders and those who do not want scores to go directly to recruiters. Students whose scores are not automatically sent to recruiters are able to share their scores for the purpose of enlistment at the individual level.

Another committee member noted that the number of leads had increased through the end of April, which is good. Dr. Velgach said they saw a significant drop during the COVID pandemic, and that, though they are not back at the pre-COVID levels, they are on the right trajectory. She said they hope to be fully back by next year. Prior to the pandemic, the enlisted testing program and the ASVAB CEP together were testing close to one million persons per year. She said the ASVAB CEP was at almost 800,000 per year prior to COVID. Another committee member complimented the website and asked if there were plans to identify the areas where people spend the most time. Dr. Rader said, yes, and that they also want to know how many times they come back to the site. She said it is possible to explore multiple plans, so people can stop and restart at a later time. She said it is still very new. The committee member replied that it would be great to learn more about the program. Dr. Rader said there is a lot more to talk about, to include classroom activities that were not covered in the present briefing. Another committee member commended the amount of data available and data planning tools. S/he said in small towns, it is empowerment to be able to make plans for the future. The committee member then asked if queries and results were filtered by interest preferences (i.e., were the resulting plans interest-driven?). Ms. McLean said they were. The committee member asked what if it did not filter by interest? Ms. McLean said the tool is designed to provide options on how to explore careers, using some or all the elements provided. Another committee member asked what percentage of people use their scores to enlist. Dr. Velgach said about 15% of the accession population use an ASVAB CEP score to enlist.

14. <u>Future Topics</u> (Tab Q)

Dr. Mary Pommerich, DTAC, presented the briefing.

Dr. Pommerich presented a list of potential topics for future DAC meetings:

- CAT-ASVAB/Form development methodology
  - Using machine learning methods to streamline the form assembly process
  - Item calibration sample size reduction study
- Unproctored testing
  - P*i*CAT/Verification Test (Vtest) updates
  - AFQT Prediction Test (APT)
- Adding new non-cognitive measures
  - TAPAS/personality measures
    - Joint-Service TAPAS effort
    - Service TAPAS efforts (Army, AirForce/Space Force, Marine Corps, Navy)
  - Interest measures
- Social media effort
- Adding new cognitive tests/composites
  - Cyber Test
  - MCt
  - CR update
  - CompT update
- Next generation testing
  - ASVAB evaluations
  - Roadmap
  - Norming investigations
- ASVAB validity
  - Criterion domain/performance metrics
- Impact of COVID/score trends for ETP and CEP
- Calculator effort

As Dr. Pommerich talked through the list of potential future topics, she mentioned the roadmap to the NextGen ASVAB, equating, and recent topics (e.g., CR). She then asked the committee for their input on what should be covered in the future.

A committee member reiterated the importance of the NextGen ASVAB, saying the roadmap should address how all the pieces fit together and get prioritized; for example, how does CR fit in? Dr. Pommerich said those were good questions, because everyone needs a better understanding of the complexity of the situation, to include what it takes to make changes and how long the process takes. She said, in the past, is has not been until changes were directed that they were made. She explained that it is difficult to make changes in their environment where so many stakeholders have different objectives. She said the path DTAC is on is not entirely clear, but it will attempt to investigate and implement the requested changes and then track the situation to determine if the objectives are met.

Another committee member commented on the roadmap, emphasizing that the testing landscape is changing very fast, with technology being the primary driver. S/he said the ASVAB is considered one of the most technologically advanced testing programs and it is important to take a proactive approach with respect to getting ready for the NextGen ASVAB. The committee member asked if DTAC was considering doing more systematic planning of needs, goals, strategies, and resources, and whether it might identify use cases for research in AI, machine learning, etc. not just for item development but for item banking, management, virtual proctoring, and more. S/he said it is important to take a systematic approach to planning. Dr. Pommerich replied that she was not sure they could address all those things in the next meeting,

but the point is taken. Another committee member likened the situation to an octopus, in it being so unwieldy with so many tests, initiatives, testing technologies, administration technologies, and so many moving parts. S/he said the situation requires stepping back and looking at the big picture, which is what is required for the committee to contribute.

Dr. Pommerich said they have been thinking about the roadmap for a while and know they need to take all these things into account. She said the timing is difficult because some components are closer than others to being ready. She suggested a visual, though difficult to construct, might be useful. The committee member reiterated that the amount of work being performed is amazing and is to be commended, and that she is not being critical in any sense. Dr. Pommerich said the DTAC team members are planners; it has a complex IT platform in place, and DoD has complex cyber requirements for maintaining the IT system. She said DTAC has an IT roadmap for the future but also a lot of current initiatives that would improve the testing experience in the current environment. To think about where they are headed with AI is a difficult scenario to envision. She pointed out the need to be proactive, but called attention to everything that must be considered: the size of the IT system and everything they have to support (e.g., registrations, testing, scores, score reporting, database, applications, tests within applications); it is extremely complicated. She said their orientation is toward gaining an understanding of the situation sufficient to allow them to be proactive versus reactive.

Another committee member said understanding the big picture is helpful, especially as an outsider, and that it would help the committee to know what components exist and how they differ from each other. She recommended starting with an overview of all the moving parts, for example, what are the differences between the various tests.

Dr. Velgach mentioned that in the future they would like committee's advice on norming, what they are doing with calculators, super-scoring, how does the program differ from what universities are doing.

As closing remarks, a committee member reaffirmed that there are many pieces to this complex puzzle, spanning from technology to substantive sections of the test. S/he said he wanted to mention that testing is under challenge these days and asked if there is a way to engage in that conversation productively moving forward on a broad array of whole person assessments. S/he said the areas in which standardized testing is critical are:

- Recruiting (learning about engagement and commitment to developing young people and older under-employed people).
- Selection, being the largest area where testing is most often used.
- Development, which often gets lost in the testing conversations. Tests can help identify development needs and emphasize to applicants the importance of learning about themselves. That speaks to training and probably many other areas, and this may be an opportunity to show why testing is so important.
- Classification, driving the decisions that ultimately lead to performance and job satisfaction as a corollary.

15. Public Comments

At the end of the second day, Dr. Velgach opened the floor to public comments and asked participants to limit their comments to 5 minutes per person.

There were no public comments. Dr. Velgach closed the meeting by saying that it had been very successful in providing information on where we are and how to proceed into the future.

Tab A

LIST OF ATTENDEES

Defense Advisory Committee on Military Personnel Testing (DACMPT)
August 16-17, 2023

Name	Position	Organization
Dr. Nancy Tippins	Owner and Manager	DACMPT (Chair), Nancy Tippins Group, LLC
Dr. Sonia Esquivel	Professor	DACMPT, US Air Force Academy
Dr. Won-Chan Lee	Professor	DACMPT, University of Iowa
Dr. Osvaldo Morera	Professor	DACMPT, University of Texas El Paso
Dr. Fred Oswald	Professor	DACMPT, Rice University
Dr. April Zenisky	Associate Professor	DACMPT, University of Massachusetts, Amherst
Dr. Sofiya Velgach	Designated Federal Officer (attendance req'd by FACA)	Office of Accession Policy (AP)
Dr. Katherine Helland	Director	AP
Mr. Christopher Graves	Senior Staff Scientist	Human Resources Research Organization (HumRRO)
Ms. Sachi Phillips	Project Manager	HumRRO
Dr. Mary Pommerich	Director	Defense Testing and Assessment Center (DTAC)
Dr. Matthew Trippe	Supervisory Personnel Psychologist	DTAC
Mr. Jeff Harber	Personnel Research Psychologist	DTAC
Dr. Tia Fechter	Supervisory Personnel Research Psychologist	DTAC
Dr. Irina Rader	ASVAB CEP National Director	DTAC
LTC Charles Manning	US Military Entrance Processing Command (USMEPCOM) Liaison Officer	AP
CPT Ryan Helm	Operations Research Analyst	US Marine Corps, Manpower Plans and

		Policies
Ms. Cheryl Fitzgerald	Branch Head	US Marine Corps, Manpower Plans and Policies
Dr. Jennifer Tucker	Assessment Branch Chief	US Space Force
SGM Alan Myers	Senior Retention & Accessions Policy Manager	US Army HQDA, G1
Dr. Erin O'Brien	Research Psychologist	US Army Research Institute (ARI)
Dr. Amanda Mouton	Personnel Research Psychologist	US Air Force
Dr. Sophie Romay	Senior Personnel Research Psychologist	US Air Force Personnel Center
Dr. Bobbie Dirr	Personnel Research Psychologist	US Air Force Personnel Center
Mr. Andrew Deregla	Personnel Research Psychologist	US Air Force Personnel Center
Dr. John Trent	Senior Personnel Research Psychologist	US Air Force Personnel Center
Mr. James Johnson	Director, Selection and Classification	US Navy, OPNAV N132
Mr. Robert Tiegs	Testing Director	US Military Entrance Processing Command (USMEPCOM)
Mr. David Davis	Chief, Testing Division	USMEPCOM
Mr. Jaime Clayton	Enlistment Testing Program Manager	USMEPCOM
Dr. Scott Oppler	Principal Scientist	HumRRO
Dr. Claire Vincent	Program Manager	HumRRO
Dr. Deirdre Knapp	Principal Scientist	HumRRO
Dr. Peter Ramsberger	Senior Staff Scientist	HumRRO
Dr. Kimberly Adams	Program Manager	HumRRO
Dr. Bill Walton	Program Manager	HumRRO
Ms. Tiffany Day	Senior Staff Scientist	HumRRO
Dr. Katherine Klein	Senior Scientist	HumRRO

Dr. Matthew Reeder	Senior Staff Scientist	HumRRO
Dr. Dan Putka	Principal Scientist	HumRRO
Dr. Kevin Bradley	Senior Staff Scientist	HumRRO
Dr. Tim McGonigle	Division Director	HumRRO
Ms. Kate McLean	Owner	Written, LLC

Tab B

AGENDA

Defense Advisory Committee on Military Personnel Testing (DACMPT)
August 16-17, 2023

August 16, 2023 (Central Time)

8:30 a.m. – 8:45 a.m.	Welcome and Opening Remarks	Dr. Sofiya Velgach (OASD(M&RA)/AP)
8:45 a.m. – 9:15 a.m.	Accession Policy Introduction	Dr. Katherine Helland (OASD(M&RA)/AP)
9:15 a.m. – 10:00 a.m.	R&D Milestones Brief	Dr. Mary Pommerich (OPA/DTAC)
10:00 a.m. – 10:15 a.m.	*Break*	
10:15 a.m. – 11:15 a.m.	Form Equating Methodology	Dr. Matt Reeder (HumRRO)
11:15 a.m. – 12:15 p.m.	ASVAB Item Development Process a. Item Writing b. Item Analysis	 Dr. Jeff Harber (OPA/DTAC) Ms. Tiffany Day (HumRRO) Dr. Matt Reeder (HumRRO)
12:15 p.m. – 1:45 p.m.	*Lunch*	
1:45 p.m. – 2:45 p.m.	TAPAS Overview/Validity Framework	Dr. Deirdre Knapp (HumRRO)
2:45 p.m. – 3:45 p.m.	TAPAS Future Work a. Joint Enlistment Composite b. Compatibility Composite	 Dr. Dan Putka (HumRRO) Dr. Kevin Bradley (HumRRO)
3:45 p.m. – 4:00 p.m.	*Break*	
4:00 p.m. – 4:30 p.m.	Complex Reasoning	Dr. Kate Klein (HumRRO)
4:30 p.m. – 5:15 p.m.	Computational Thinking	Dr. Kimberly Adams (HumRRO)
5:15 p.m. – 5:30 p.m.	*Public Comments*	

August 17, 2023 (Central Time)

8:30 a.m. – 9:00 a.m.	High School Curriculum Study	Dr. Peter Ramsberger (HumRRO)
9:00 a.m. – 10:00 a.m.	Non-Native English Speakers Analysis	Dr. Bill Walton (HumRRO)
10:00 a.m. – 10:15 a.m.	*Break*	
10:15 a.m. – 11:30 a.m.	ASVAB CEP Update/Demo	Dr. Irina Rader (OPA/DTAC)
11:30 p.m. – 12:00 p.m.	Future Topics	Dr. Mary Pommerich (OPA/DTAC)
12:00 p.m. – 12:15 p.m.	*Public Comments*	
12:15 p.m. – 12:30 p.m.	Closing Comments	Dr. Nancy Tippins (Chair, DACMPT)
12:30 p.m. – 2:00 p.m.	*Working Lunch (Administrative Items)*	

ABBREVIATIONS KEY:
ASVAB - Armed Services Vocational Aptitude Battery
ASVAB CEP - ASVAB Career Exploration Program, student testing 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
DACMPT – Defense Advisory Committee on Military Personnel Testing
HumRRO - Human Resources Research Organization
OASD(M&RA)/AP - Office of the Assistant Secretary of Defense (Manpower & Reserve Affairs)/Accession Policy
OPA/DTAC - Office of People Analytics/Defense Testing and Assessment Center
TAPAS – Tailored Adaptive Personality Assessment System

Tab C

LIST OF ACRONYMS

3PL	Three-Parameter Logistic Model
ACT	American College Testing Test
AFQT	Armed Forces Qualification Test
AI	Automotive Information
AIG	Automated Item Generation
AO	Assembling Objects
AP	Accession Policy
APT	AFQT Prediction Test
AR	Arithmetic Reasoning
ARI	U.S. Army Research Institute for the Behavioral and Social Sciences
AS	Auto & Shop
ASVAB	Armed Services Vocational Aptitude Battery
CAASPP	California Assessment of Student Performance and Progress
CAT-ASVAB	Computerized Adaptive Testing ASVAB
CCSS	Common Core State Standards
CDF	Cumulative Distribution Functions
CEP	Career Exploration Program
CompT	Computational Thinking
COVID-19	Coronavirus Disease 2019
CTE	Career and Technical Education
CTM	Careers in the Military
DACMPT	Defense Advisory Committee on Military Personnel Testing
DIF	Differential Item Functioning
DLIELC	Defense Language Institute English Language Center
DoD	Department of Defense
DTAC	Defense Testing and Assessment Center
ECL	English Language Comprehension Level
EDPT	Electronics Data Processing Test
EI	Electronics Information
EL	English Learner
ELA	English Language Arts
ELP	English Language Proficiency
ESL	English as a Second Language
ETP	Enlistment Testing Program
FACA	Federal Advisory Committee Act
FAST	Fundamental Applied Skills Training
FIML	Full Information Maximum Likelihood
FLRI	Foreign Language Recruiting Initiative
FY	Fiscal Year
FYI	Find Your Interests

GED	General Educational Diploma
GS	General Science
HSLS	High School Longitudinal Study
HSTS	High School Transcript Study
HumRRO	Human Resources Research Organization
*i*CAT	Internet version of the CAT-ASVAB
IRC	Independent Review Commission
IRT	Item Response Theory
JAMRS	Joint Advertising Market Research and Studies
Lasso	Least Absolute Shrinkage and Selection Operator
MAPWG	Military Accession Policy Working Group
MC	Mechanical Comprehension
MDPP	Multidimensional Pairwise Preference
MEPS	Military Entrance Processing Stations
METs	Military Enlistment Testing Site
MK	Mathematics Knowledge
M&RA	Manpower & Reserve Affairs
NAEP	National Assessment of Educational Progress
NDAA	National Defense Authorization Act
NGSS	Next Generation Science Standards
NLSY	National Longitudinal Survey of Youth
NNES	Non-Native English Speakers
NNLS	Non-Negative Least Squares
OASD	Office of the Assistant Secretary of Defense
OLS	Ordinary Least Square
OPA	Office of People Analytics
OSD P&R	Under Secretary of Defense for Personnel and Readiness
P&P	Paper-and-Pencil
PARCC	Partnership for Readiness for College and Careers
PAY97	1997 Profile of American Youth
PC	Paragraph Comprehension
PERSEREC	Personnel Security and Research Center
P*i*CAT	Pending Internet Computerized Adaptive Test
PTI	Post-Test Interpretation
SAT	Scholastic Aptitude Test
SGM	Sergeant Major
SI	Shop Information
SME	Subject Matter Expert
STEM	Science, Technology, Engineering, and Mathematics
TAPAS	Tailored Adaptive Personality Assessment System
TOA	Theory of Action
USC	U. S. Code
USMEPCOM	U.S. Military Entrance Processing Command

VE	Verbal Expression
VTest	Verification Test
WK	Word Knowledge

Tab D

NANCY T TIPPINS
GROUP

November 15, 2023

Katherine Helland, Ph.D.
Director, Accession Policy
Accession Policy
Room 3D1066
4000 Defense Pentagon
Washington DC 20301-4000

Dear Dr. Helland,

The Defense Advisory Committee on Personnel Testing (DACMPT) is pleased to provide this report on our meeting of August 16-17, 2023, in Rosemont, Illinois. We found this meeting to be particularly informative, productive, and well-facilitated. The interactions among the presenters and Committee members as well as other participants in the meeting who are interested in military personnel testing, were useful. In addition to myself, the DACMPT Committee members are Drs. April Zenisky, Fred Oswald, Won-Chan Lee, Osvaldo Morera, and Sonia Esquivel. All but one member of the DACMPT (Dr. Esquivel) were able to attend the meeting in person. Dr. Esquivel attended virtually as available.

The meeting began with opening remarks from Dr. Sofiya Velgach (Assistant Director for Testing Standards, Office of the Under Secretary of Defense for Personnel and Readiness/M&RA/MPP(AP))and Dr. Nancy Tippins (Chair of the DACMPT). Dr. Velgach reviewed the agenda and facilitated introductions of the members of the DACPT and other attendees, including those from the HumRRO staff, from the Defense Personnel Assessment Center (DPAC), and various military units.

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

Accession Policy Brief

After introductions, Dr. Katherine Helland, Director of Accession Policy (AP) provided a presentation on AP's organizational structure and then discussed the current recruiting environment and mission, as well as planned actions and testing priorities. Recruiting continues to face challenges due to a number of factors, including a shrinking pool of qualified youth and a declining propensity to serve. A central question for AP is what risks should be addressed with the testing program. For example, should calculators be allowed? How long should ASVAB scores be valid? Three key actions for AP include increasing propensity, expanding eligibility, and improving candidate processing. AP will be focusing on five key testing initiatives:

- Expansion of ASVAB to alternative devices
- Development of TAPAS-based Joint Enlistment Composite
- Development of TAPAS-based Compatibility Composite
- Development of new special purpose test: Mental Counters
- Development of new special purpose test: Complex Reasoning

During the presentation, the members of the DACMPT posed several questions regarding preparation courses and their effectiveness in raising both scores and performance. Other questions regarded the

effect of publicity about sexual assault cases on recruiting, and accommodations for Non-Native speakers and individuals with ADHD and mental health problems. Although the sexual assault cases are sometimes provided by candidates as a reason for not enlisting, it is not one of the primary reasons. The Services will need to determine the level of risk they are willing to take with respect to candidates' ADHD and mental health issues and determine their potential impact on performance and attrition.

R&D Milestone Brief

Dr. Mary Pommerich of the Defense Testing Assessment Center (DTAC), Office of People Analytics (OPA), provided an overview of the R&D efforts related to the ASVAB. Dr. Pommerich first covered research projects on ASVAB development, including item development efforts, item pools and forms, and the effects of calculator use on scores and performance. She then discussed ongoing research on the ASVAB and the Enlistment Testing Program (ETP), the Career Exploration Program, Military Compatibility Assessment, and the ASVAB and Enlistment Testing Program Revision.

The members of the DACMPT expressed their admiration for the amount of work that was being done in relatively short time frames. When asked if there were other things that should be done that have not yet been undertaken, Dr. Pommerich explained that the DTAC was currently well-funded. The primary issues her unit faces are the deadlines for the work. Many of these projects take a long time to complete, and more money does not equate to shorter time frames.

Another topic discussed was the use of technology in test development, such as automated item generation (AIG) or generative artificial intelligence (GAI). All agreed that such technology could not completely replace human item writers at the present time. However, it was noted that GAI was rapidly developing, and users have been learning how to prompt GAI (e.g., Chat GPT) for information to support item writing efforts.

Virtual proctoring was also discussed. The current OSD policy is to proctor high-stakes tests with humans. The main advantage of virtual proctoring is the ability to administer the test in the applicant's home which alleviates the need for the recruiter to take the candidate to a testing center and allows the recruiter more time to look for other potential candidates. In addition to cheating and harvesting item content, concerns about virtual proctoring, such as misreading and missing signals indicating malfeasant behavior and bias against some groups resulting from scanning technologies, were also mentioned.

Recommendations

The DACMPT appreciated the detailed information Dr. Pommerich provided and wishes to be updated on the results of the research efforts being conducted and the plans for new research. The DACMPT also recommends that DTAC monitor developments in GAI to determine if it will be a useful tool at some point in the future. DTAC should also stay up to date on innovations in virtual proctoring and continue to research other countries' positions to determine what input to give to policymakers who will make decisions regarding the use of virtual proctoring.

Form Equating Methodology

Dr. Matt Reeder of HumRRO provided background and results of the CAT-ASVAB equating study for the newly developed CAT-ASVAB pools 11-15. The item parameters of items in pools 11-15 underwent

rescaling to the operational CAT-ASVAB scale, and subsequently, standard score equating was conducted to align the standard scores (SS) for the new pools with the mean and standard deviation of the SS of the reference pool. The equating approach relies on the measurement invariance assumption of item response theory (IRT) and aims to create equal distributions of scores across alternate pools.

The equating study was conducted using the random groups design and implemented in three phases of operational administration of new pools. Each phase included a progressively larger sample size. Equating results were evaluated in terms of the differences between the refence and new pools in the cumulative distribution functions of the qualification composite. The comparison of distributions of key demographic variables suggested that groups were randomly equivalent. Most composites displayed similar distributions across the reference and new pools. Five composites consistently showed statistically significant differences; however, those differences were comparable to what was observed using CAT-ASVAB pools 5-9 equating and were within a tolerable range. The equipercentile objective of equating was revisited by comparing the results solely based on the IRT invariance property with those based on pool-specific transformation constants to match the mean and standard deviation of the reference pool. In general, the pool-specific transformation yielded composite distributions closer to the reference pool. The results of subgroup performance analysis were similar to those seen during pools 5-9 development.

Recommendations

The DACMPT acknowledged the outstanding technical work and comprehensive information provided. The committee recognized the importance of using the pool-specific scale transformation, in addition to relying on the IRT measurement invariance property, for the purpose of improving the congruity of composite distributions and qualification rates across different pools at a group level. However, the committee recommended examining the potential bias that could arise from the pool-specific scale transformation when estimating applicants' abilities at the individual level. The committee suggested that a simulation study relevant to the question be designed to explore this issue. The DACMPT also raised a question regarding the consistency of using the same operational IRT scoring method also used in scaling, equating, and other psychometric analyses. Additional rationale may be necessary if consistency was not maintained. The committee also highlighted the importance of contemplating the implications of the project's outcomes that align with potential developments of NextGen ASVAB.

ASVAB Item Development Process – Item Writing

Dr. Jeff Harber (DTAC) and Dr. Tiffany Day (HumRRO) described the ASVAB Item Development Process. Between 2018 and 2022, an average of 5,860 items per year have been developed. After defining the number of "easy," "medium," and "hard" tryout items needed, DTAC and HumRRO partnered to develop and revise items. Item writing teams composed of subject matter experts, junior editors, senior editors, and graphic artists created and reviewed the items, using tools, including (a) item writing guides, (b) guidelines for sensitivity and bias reviews, (c) a blueprint for each subtest and weights for development, and (d) items for future use on the ASVAB. HumRRO editors also used tool kits, editor checklists, and style guidance for copy editing to ensure that all items meet test specifications. Steps to prevent disclosure of items were comprehensive, including signed confidentiality agreements among all team members, secure access to test specifications, completion of mandatory DoD security training by contractors, and item banks secured by HumRRO.

DTAC editors reviewed and edited the items independently for style, content accuracy, stem clarity, enemy items, bias, and use of distractors. Items could be sent back to HumRRO with edits or requests for replacement. Once a series of 100 items was approved, the senior editor updated the ASVAB Item Bank and readied these items for content review. An independent subject matter expert reviewed DTAC-approved item. Independent revisions to items with a rationale were provided to the HumRRO Senior Editor, who responded to the feedback, applied needed edits, and submitted final items to DTAC. DTAC could still request another content review of the items at this stage of the process.

In summary, each potential ASVAB item is reviewed by at least seven individuals at least once (one subject matter expert writer, one or two HumRRO junior editors, one HumRRO senior editor, three DTAC editors and one independent subject matter expert content reviewer). Safeguards are in place to ensure item safety and test bank safety.

Recommendations

In their discussion of this presentation, the DACMPT members recognized the careful item writing process that results in secure, high-quality items written to specifications. The DACMPT has no recommendations to improve this process.

ASVAB Item Development Process – Item Analysis

Dr. Matt Reeder of HumRRO provided the overview of the item analysis as a part of the ASVAB item pool development process. The tryout item data underwent cleaning and a pre-calibration key check. During the cleaning process, the records that were invalid, ineligible, or reflected potentially unmotivated participants were removed. The pre-calibration key check involved CTT-based analysis, evaluation of response patterns, review by content SMEs, and the removal of items with multiple correct responses or content flaws. Item parameters for the 3PL model were calibrated using BILOG-MG. Following calibration, psychometric quality analyses were conducted, including the assessment of item information, item-model fit, distractor analysis, differential item functioning (DIF), and screening rubric. The percentages of items retained during Forms 11-15 development were reported by subtest. Finally, the methods and results of the CAT-ASVAB dimensionality assessment were presented. Recent investigations into ASVAB dimensionality included research on sparse data dimensionality assessment with the Cyber test and the feasibility of combining AR and MK ASVAB subtests, and exploration of both IRT and item factor analytic approaches.

Recommendations

The DACMPT acknowledged the challenge of identifying suitable methods for evaluating dimensionality of ASVAB tryout items under sparse data conditions and proposed the potential use of basic CTT-based statistics, such as item-total correlations as a viable option. The committee also noted that planned missingness can be acceptable when researching the overall dimensionality (correlational structure) of measures; however, planned missingness is definitely not recommended when using scores for estimating individual scores in operational settings. Suggested solutions included the potential use of machine learning and inspection of the content of items to identify themes.

TAPAS Validity Framework and Joint Enlistment Composite

Dr. Deirdre Knapp and Dr. Daniel Putka began the briefing with an overview of the TAPAS (Tailored Adaptive Personality Assessment System), including its developmental history and actions taken to establish the validity of TAPAS for several different selection and classification purposes. One of these actions was the development of a theory of action / validity argument framework, with the overarching goal of collating research and findings across Services and TAPAS use cases to advance a coherent validity argument for the TAPAS as well as identify specific directions for additional validity research. Over the past several years, the methodology employed consisted of the development of a theory of action (ToA), leading to an interpretive argument, a validity argument, and finally a validity argument summary. For the TAPAS ToA, three major claims were identified:

- Temperament factors are predictive of performance and continuance intentions/behavior
- TAPAS measures a useful sample of temperamental facets
- Respondents selected or classified based on TAPAS scores (in combination with other indicators) have a higher likelihood of success within particular military occupations

Dr. Knapp and Dr. Putka illustrated the specific claims and associated assumptions and summarized the interpretive argument (3 major claims; 18 specific claims for selection, classification, and selection and classification; and 47 assumptions). They further detailed the ongoing work to organize validity evidence for TAPAS (i.e., the technical reports produced since 2020).

They also provided an update on the development steps for a joint-service TAPAS composite to be used for general enlistment selection and qualification decisions, along with a research plan for gathering criterion-related validity evidence for this composite. Part of this plan is to determine which facets are common and necessary for a composite test that would be required across the Services.

Recommendations

The DACMPT suggested that the feasibility of a synthetic validity approach should be explored as a way to make the most of the available data given their variability and sparseness. A further suggestion made was to consider strategies to collect validity data retrospectively (i.e., concurrent validity). The committee also asked about the use of the TAPAS composite scores and the weights for its multiple components. For the purpose of the Joint Services Composite, the weights might be common across all services, but individual services might build additional composites and each assign unique weighting schemes. The DOD is tasked with producing the weightings. Another suggestion was to include other TAPAS facets for future research.

TAPAS for Military Compatibility

Dr. Kevin Bradley presented a briefing on a new proposed use of TAPAS, to assess military compatibility initially among the enlisted population, although there is potential interest in this use for prospective officers in the future. This initiative emerged from a directive concerning a multifaceted approach to addressing sexual assault and harassment in the Services. This proposed compatibility assessment is intended to fulfill a recommendation for a pre-accession instrument to screen for alignment with military core values. The intent of this instrument is to assess compatibility and then use that

information to identify people who are at risk of exhibiting "counterproductive work behaviors" (CWBs). The identification of TAPAS facets that in combination may be predictive of one or more such behaviors is exploratory in nature. Identifying those facets will include a literature review of the dark tetrad research and an exploration of the feasibility of a licensed clinician carrying out this kind of assessment. Dr. Bradley reinforced the idea that any TAPAS-based noncognitive assessment used to evaluate military compatibility may be used in conjunction with other approaches/indicators.

The DACMPT inquired as to the rationale for clinical assessment rather than a battery of assessments. Dr. Bradley noted that a clinical approach was just one approach of several being investigated, but he pointed out the logistics and person-hours required would likely make a clinical assessment prohibitively difficult. A related challenge to this endeavor that was noted is the low base-rate of the kinds of serious activities that the services are interested in minimizing/preventing.

Recommendations

The members of the DACMPT had a number of questions about this research and made several suggestions on overcoming the challenges inherent in it. One question involved the definition of military core values, and the extent to which they are incompatible with counterproductive behaviors, which are also difficult to define and measure. Military core values vary across branches of the Services, but they generally refer to constructs such as honor, courage, commitment, sense of duty, and so forth. Another member of the committee suggested that the challenge of measurement might be addressed by identifying a criterion more proximal to the actual counterproductive behaviors (if those were specifically elaborated) which would sacrifice generalizability for fidelity to specific trait identification/prediction. The committee also suggested considering the possibility of deconstructing counterproductive work behaviors into essential components (e.g., making verbal comments as a prelude to physical altercations) as a strategy to address the low base-rate issue. A great deal of variability has been found among the Services in terms of ratings of counterproductive work behaviors, and there is a general lack of consensus on the importance of specific negative behaviors (e.g., sedition, aggression, harassment). A further question was raised about the relative stability of the characteristics to be assessed and the extent to which pre-accession assessment of these constructs might be useful for the prediction of later behaviors. Multi-level unit of measuring these constructs over time was suggested as a possible alternative.

The DACMPT expressed a great deal of concern about what is being measured at what specificity, and what level of reliance on the data is appropriate. At present, while the infrastructure for TAPAS exists in MEPS, making TAPAS a logical administrative choice as an instrument to measure these CWBs, there remain a number of significant questions outstanding about the extent to which TAPAS could defensibly predict CWBs, adherence to military core values, and military compatibility in the general case or at a more specific, granular level targeting more clearly articulated CWBs. The ongoing work to establish a validity argument for TAPAS for varied purposes and uses suggests that the outcomes associated with TAPAS use are variable, and considerable work will need to be done around construct definition (including specificity), the stability of the construct at pre-accession and over time for various examinee groups (such as enlisted vs. officers, and demographic considerations like male/female, race/ethnicity), the validity argument for the use of this measure for purposes such as disqualifying enlistment candidates or identifying potential issues, and interpretation and use generally. The DACMPT

recommends that considerable attention be paid to determining what should be measured in a compatibility assessment for articulated specific purposes. In addition, Accession Policy should be open to instruments other than TAPAS that provide targeted information that could predict counterproductive work behaviors in general or specific counterproductive work behaviors, adherence to military core values, and military compatibility.

One final suggestion involved the use of a clinical assessment to follow-up on high scores on facets predictive of counterproductive work behaviors. This two-stage process could save money by limiting the clinical evaluation to high scorers only.

Complex Reasoning

Dr. Kate Klein of HumRRO defined *complex reasoning* (CR) as abstract problem-solving that is nonverbal in nature (e.g., visual, spatial). CR measures therefore do not require job-specific knowledge and skill, and they have no language requirements; as such, they have been found to be predictive of performance across a wide range of jobs.

The CR measure tested involved matrix reasoning, where each item is a 3x3 matrix of shapes, with a shape missing in the lower right corner. Respondents must select the tile that completes the matrix from a set of four presented. Based on a 24-item CR measure and a large (N = ~ 2,600) and age-relevant (18-35 years old) non-military sample, results indicated the CR measure had high reliability and less race/ethnicity-based and gender-based adverse impact than is found in knowledge-based tests.

Recommendations

The DACMPT made several suggestions:

- *Measure development*: Determine why CR scores were "spiked" at a score of 11 across the three forms (this is unlikely to be coincidence). Continue expanding the item bank: Given that only 24 items were developed here, the items content might be leaked to examinees who then cheat. Fortunately, this can be remedied, because the quick generation of literally thousands of items is a virtue of the item format.
- *Nomological net:* Correlate CR with ASVAB subtests to understand the nature of CR, where shared and unique sources of variance occur between the measures.
- *Validation*: Support the CR measure further with validity evidence drawn from sources such as past military studies involving similar CR measures or the research literature when the results are generalizable to the military setting as well as from new studies with the current CR measure.
- Locate existing military data with CR-related data, in addition to conducting new validation work on the current CR measure (both selection- and classification-oriented validation). Although some military tests involving CR have not demonstrated incremental validity (see Besetsny et al., 1993[1]), there is clearly more work to be done under a broader research framework. To this end,

[1] Besetsny, L. K., Ree, M. J., & Earles, J. A. (1993). Special test for computer programmers? Not needed: The predictive efficiency of the Electronic Data Processing Test for a sample of Air Force recruits. *Educational and Psychological Measurement*, *53*(2), 507-511. https://doi.org/10.1177/0013164493053002020

job analyses, O*NET data, and other resources may speak clearly to the need for an agenda for CR research across a wide range of MOS.

- *Profile-driven analyses*: Future research might consider how CR might work in tandem with a recruit or enlistee's profile of ASVAB scores. For example, specific ability tests are known to be more correlated (less differentiated) for those with lower general cognitive ability (see Detterman & Daniel, 1989[2]), and those with higher cognitive ability may be more trainable for MOSs that do not fit their ASVAB subtest profile. These points have implications for classification that considers each enlistees' current interests and future goals alongside broader recruiting and labor demands.

Computational Thinking

Dr. Kimberly Adams of HumRRO reported that c*omputational thinking* (CT) was identified in the recent National Defense Authorization Acts (NDAAs, 2021 and 2022) in connection with "skills relevant to military applications, including problem decomposition, abstraction, pattern recognition, analytical ability, the identification of variables involved in data representation, and the ability to create algorithms and solution expressions."[3] With this definition comes the critical requirement that a measure of these six dimensions of CT be established by October 1, 2024. Because this timeline prevents the development of a new measure designed solely for the purpose of assessing CT, the committee agrees—and appreciates—that meeting this deadline necessitates combining a great deal of expert judgment on an appropriate approach with an enormous amount of effort in vetting different candidate measures, to include a new measure of complex reasoning (CR). This was fully evident in the presentation on CT development to date.

To estimate CT, SMEs were asked to estimate (a) correlations among the six dimensions of CT mentioned above and (b) correlations of those six dimensions with other tests: the ASVAB subtests, Mental Counters, EDPT, and the new Complex Reasoning test. SME correlation estimates were then averaged; these averages were used in subsequent regression analyses that predicted the CT composite (six dimensions above combined) from these other tests. Because all correlations (zero-order validities and predictor correlations) were high and positive, the R^2 value (joint contribution of the predictors) tended to be high, while the predictor regression weights (unique contributions of each predictor) tended to fluctuate somewhat unpredictably. Therefore, more parsimonious modeling is possible, where predictors can be dropped without a meaningful drop in R^2 for predicting CT. Constrained lasso models and non-negative least squares (NNLS) modeling were applied to this end.

[2] Detterman, D. K., & Daniel, M. H. (1989). Correlations of mental tests with each other and with cognitive variables are highest for low IQ groups. *Intelligence*, *13*(4), 349-359. https://doi.org/10.1016/S0160-2896(89)80007-8

[3] National Defense Authorization Act for Fiscal Year 2021. H.R. 6395. 116th Cong. (2021).

Recommendations

The DACMPT offered several recommendations:

- *Validation*: Given that a new measure solely designed to assess CT is not being developed, it could be useful in the time allowed to consider approaches that might refine the validation of CT composite further. For example, in a two-stage process, you might find the weights that estimate the six components of CT separately in stage 1; then in stage 2, you create a composite of the six CT predicted scores depending on the MOS (SMEs rate the importance of CT components for each MOS).
- *Fairness*: A question that is important to the Services is, "Will selection/classification outcomes based on CT be fair to race/ethnicity and gender subgroups, in terms of minimal adverse impact?" This information was not provided, but given that there are some subgroup mean differences on ASVAB and other cognitive tests examined here, subtest composites can increase these mean differences.
- *EDPT*: Given that components of EDPT look like ASVAB + CR subtests and given that EDPT will not be given to all enlistees, consider removing EDPT from further research.

High School Curriculum Study

Dr. Peter Ramsberger of HumRRO presented a research plan to study the alignment of the ASVAB subtests with the curricula in high schools and the ways ASVAB content is taught in schools. Dr. Ramsberger noted the decentralization of public education and the difficulty of identifying trends and adapting to them and suggested it would probably be easier to adapt the ASVAB to the curriculum than change the course of public secondary education. He also pointed out that there are two potential problems, the first of which is the availability of courses included in the content measured by the ASVAB subtests. The second problem is the willingness of students to take the courses related to ASVAB subtests when offered. The DACMPT also noted that course catalogs may contain a class that is not necessarily offered because a teacher is unavailable.

Recommendations:

The DACMPT would like to hear more about this research and understand how the NextGen ASVAB and the Critical Thinking and Complex Reasoning Tests support alignment with common high school curricula. The DACMPT also suggested that researchers consider multilevel analyses on variables like school and state to test the hypothesis that schools with more resources provide more courses. Another suggestion was to consider the extent to which such information could be used to assess schools from a workforce development perspective. Another possibility to investigate was whether or not schools with offering curricula aligned with ASVAB subtests and better resources offered better recruiting environments and produced more eligible students with a propensity for military service.

Non-Native English Speakers Analysis

Dr. Bill Walton of HumRRO provided a summary of analyses that were designed to determine whether the use of ASVAB may be limiting the ability of English Language Learners (ELLs) to enlist in the military. According to the Every Student Succeeds Act of 2015 definition, an ELL is someone whose native

language is not English and whose level of English fluency makes it difficult to perform well in school or society. States differ in the ways they identify someone as an ELL, so commonly used assessments like the World Class Instructional Design and Assessment (WIDA) were used to determine who is an ELL. The percentage of ELL students overall has increased over time (from 8.1% in 2000 to 10.1% in 2017). In addition, the percentage of ELL students by grade decreases as students get older. For example, in Fall 2017, 15.9% of kindergarteners were ELL, but only 4.6% of high school seniors were ELL. The most frequent language spoken among ELLs in 2017 was Spanish, as almost 75% of ELL students spoke Spanish at home.

Dr. Walton's team examined differences on the verbal scores of non-qualifying AFQT scores between native and non-native English speakers and between citizens and non-citizens within ethnic/racial groups (e.g., Hispanic citizen versus Hispanic non-citizen). Dr. Walton estimated that less than 1% of students in the 11th and 12th grade are both non-Native English speakers (NNES) and propensed to join the military. Moreover, only 1% of all applicants are disqualified from joining the military based on AFQT scores alone. Across five years of assessments from 2015-2019, 0.11% of applicants had math knowledge and arithmetic reasoning scores that met standards and verbal expressions scores that were not qualifying, indicating that the population of ELL examinees affected by their language abilities was very small.

Recruiting efforts that summarized the military's attempts to recruit NNES individuals since the 1980s was nicely summarized. Lessons learned from these prior attempts to recruit NNES individuals were also presented and recommendations for military screening were also presented. In addition, the Defense Language Institute English Language Center (DLIELC) provides world-wide English language training, in which training is geared toward an individual's English Comprehension Level (ECL).

To conclude, the ASVAB does not eliminate a large number of highly qualified recruits who are not English proficient. There were questions about the use of the English Comprehension Level exam with respect to tapping into language relevant skills for military service and whether it was comprehensive in measuring all required language skills. ESL instructional practices employed by DoD were aligned with best practices but incorporating increased military subject matter into training was discussed.

During the presentation, the DACMPT had a robust discussion of the findings and the goal of this exercise in creating this survey, remarking on the low base rates of disqualification due to issues involving language and the relationship of English proficiency to other factors that may disqualify an examinee from enlistment. Dr. Helland indicated that only 1% of applicants are now disqualified due to AFQT scores alone. Dr. Velgach added that this work was requested by Congress, and the feedback received indicated that the report based on this presentation was one of the best Congress had seen.

Recommendations:

The DACMPT recommends considering how this report informs the development of the NextGen ASVAB. In addition, it may be useful to determine what level of proficiency is needed for military service. For example, how do work-relevant language and technical language lead to effective learning? What idioms might be important to functioning in and an MOS that is not included in formal assessments (e.g., due to work culture, due to geographic assignment)? How might job redesign and technology (e.g., AI tools, translators) be used to improve language facility for ELL, or in fact all enlistees? Given these and other considerations, appropriate MOS-relevant levels of language proficiency, and criteria for

measuring those levels should be revisited for the benefit of expanding recruitment and enlistment efforts.

ASVAB CEP Update/Demo

Dr. Irina Rader, OPA/DTAC and Ms. Kate McLean (Written, LLC) provided an overview of the ASVAB Career Exploration Program (CEP), which aims to enhance career literacy among students by providing them exposure to career field entry requirements and planning tools for future-oriented career development. The program creates an action plan that offers career exploration services to students and can be shared with parents and educators and generates qualified leads for military recruiters. Dr. Rader began by summarizing the current state of the program, the year-to-date usage metrics, the priorities for the 2023-2024 calendar year, the influence of state legislation and activities on the program, proficiency training for post-test interpretation (PTI), national events and marketing efforts, and finally, a demonstration of CEP 2.0. During the 2021-2022 school year, the program tested 607,324 students in grades 10-12 across 12,907 participating schools, providing 494,981 leads to the military services. Furthermore, over 1.3 million people accessed asvabprogram.com and over 560,000 users looked at careersinthemilitary.com.

Recommendations

Following the overview, the DACMPT complimented the tool and made a recommendation to identify ways to evaluate user engagement that goes beyond merely counts of accessing the website, such as by measuring frequency of return users. The committee also endorsed the idea of better explaining the program, so that more participants take advantage of the Post-Test Interpretation service.

Future Topics

Dr. Pommerich presented a list of potential future topics and led a discussion of them with the DACMPT. Members of the DACMPT believed that all the suggestions for future research were worthy of attention.

Recommendations

The DACMPT recommends future meetings incorporate briefings on the following topics:

- Overview of the various tests that highlights similarities and differences among tests (e.g., Cyber test vs. EDTP)
- Another review of the equating procedures
- Overview of NextGen and how the pieces (e.g., Complex Reasoning) fit together
- Overview of the process for planning that takes into account a rapidly changing testing landscape (especially important given the rapid influx of AI technologies that affect testing)
- Norming procedures
- Allowing the use of calculators
- Reviewing the nature, pros, and cons of super-scoring

DACMPT Terms of Reference (TOR)

DACMPT TOR stipulates that during the first year after date of the ToR, DACMPT will:

- Review the Department's current military accession testing capabilities to select, classify, and provide career exploration information to the accession population; identify gaps based on best practices from academia and private industry; and recommend changes leveraging private sector best practices.
- Review the Department's approach and methodology to develop, administer, and make decisions based on applicable accession instruments; and recommend modernization techniques to ensure all instruments are reliable, valid, and fair to all demographic populations and used for appropriate accession decisions. Recommendations must leverage the latest theory and standards used within the realm of test development and uniform guidelines for employee selection.

As a result of the first two sessions in FY 2023 (December 2022 and August 2023) the committee has concluded the review of two objectives above. Overall, the committee finds that the Department has a comprehensive testing program to select, classify, and provide career exploration information. Critical gaps in capabilities have not been identified. The Committee agrees with the Department's plans for future development and measurement efforts and feels they will aid in delivering a whole person assessment.

Furthermore, the committee concurs with the methodology being used to develop, administer, and make decisions based on applicable accession instruments. The above recommendations provide advice for further refinement of research and management practices and techniques helping to ensure all instruments are reliable, valid, fair, and used for appropriate decisions.

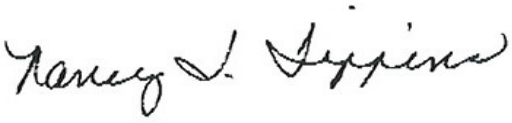
Summary

The DACMPT remains impressed by the scope of activities of Accession Policy and DTAC. Especially admirable are the extent and quality of research on the ASVAB and other assessments. Funding levels to date have allowed Accession Policy to build and maintain a testing research program that reflects the state of the art in assessments and has no equal. We applaud the high standards that are maintained and encourage the continuation of this research. Only with these extraordinary efforts can the Department of Defense's talent management strategies be maximally effective and sustain the U.S. as a world-leading military power. Without these efforts, we manage talent much less effectively and pay a serious cost for that.

Overall, the DACMPT meeting was very informative and useful, deepening our appreciation of the work of the dedicated experts and our understanding of the intensive research supporting the military's assessment program. The important and informative results of this program clearly increase military effectiveness while informing future research to the same end. The DACMPT appreciates the efforts of Accession Policy and DPAC staff and the research staff of each of the services as well as the consultants who provide their services to the DoD. As always, the DACMPT is interested in supporting these efforts,

as they provide strong, well-informed, and timely justification for the intended interpretations and uses of the ASVAB. We look forward to our next meeting.

Sincerely,



Nancy T. Tippins, Ph.D.
Principal, The Nancy T. Tippins Group, LLC
Chair, Defense Advisory Committee on Military Personnel Testing