DEFENSE ADVISORY COMMITTEE ON MILITARY PERSONNEL TESTING

September 26-27, 2019 Meeting

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

Minutes approved for public release.

Michael A Rodriguez

December 12, 2019

Dr. Michael Rodriguez, Chair, DACMPT **DATE**

DEFENSE ADVISORY COMMITTEE
ON
MILITARY PERSONNEL TESTING

Sonesta Philadelphia, Philadelphia, Pennsylvania
September 26-27, 2019

The meeting of the Defense Advisory Committee on Military Personnel Testing (DACMPT) was held at the Sonesta Rittenhouse Square, Philadelphia, PA on September 26-27, 2019. Dr. Sofiya Velgach (Assistant Director, Accession Policy Directorate [AP]) opened the meeting by stating that it was being held under the provisions of the Federal Advisory Committee Act (FACA) of 1972 (5 USC, Appendix, as amended), the government in the Sunshine Act of 1976 (5 USC, 552b, as amended), and 41 CFR 102-3.140 and 102-3.150 and open to the public. She said the meeting agenda was available and that public comments would be heard at the end of each day. She introduced the new DACMPT committee member, Dr. Nancy Tippins, and then thanked the committee members for their participation and the presenters for their support of the committee's activities. She then directed introductions.

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

1. Accession Policy Update (Tab D)

Ms. Stephanie Miller, Director, AP, presented the briefing.

Ms. Miller began by summarizing the mission of AP, which is to "develop, review, and analyze policies, resources, and plans for Services' enlisted recruiting and officer commissioning programs." She then presented an organization chart detailing the structure and programs within AP. An additional chart summarized the critical items facing the Directorate in the areas of testing, mental health, family readiness, security clearances, medical care, and national service. A table displayed the fiscal year (FY) 2019 recruiting mission for Service (Active Duty, Guard, and Reserve), followed by a table displaying results as of August 2019. For Active Duty recruiting, both the Marine Corps and Air Force have succeeded in meeting mission, while the Army (98.34%) and Navy (99.97%) are slightly behind. All Reserve Component goals have been met except for those of the Navy Reserve, which as of end of August 2019, stands at 86.76% of mission. Recruiting quality goals include accessing 90% high school degree graduates, 60% or more in the Armed Forces Qualification Test (AFQT) I-IIIA range, and 4% or less in Category IV. As of end of August 2019, all Services/Components, have met this goal except the Army National Guard, which accessed 4.1% Category IV recruits.

Ms. Miller then provided a list of Congressional Reports issued by her office. These include a report on the Armed Services Vocational Aptitude Battery (ASVAB) with 10 years of applicant data, demographic information, number of recruits in aptitude Category V, counties scoring in the lowest 5 percent on the ASVAB, and efforts to share information with the Department of Education. An additional report focuses on recruiting of Non-native English Speakers and covers enlistment practices regarding aptitude, academic potential, and academic achievement, as well as marketing efforts, recruiter interactions, and enlistment rate.

As Ms. Miller briefed the list of critical items related to testing (slide 4), Dr. Velgach commented that a cross-service WG was being considered to investigate process improvements for assessing character. Subsequently, a committee member asked whether obtaining verifiable medical data for applicants would violate Health Insurance Portability and Accountability Act (HIPAA) requirements. Ms. Miller explained that, by applying for enlistment, applicants waive their HIPAA rights.

On the FY 2019 mission (slide 5), a committee member asked if the Services' goals were currently set to maintain or to build force levels. Ms. Miller said there had been a slight uptick in recruitment goals in recent years, but that was leveling out. She explained that the goals for each Service are based on a complex formula driven by the mix of forces required for conducting large scale operations in accordance with the national security focus. She added that the formulation considers the number of forces that need to be recruited versus retained.

As Ms. Miller briefed on recruit quality (slide 7), a committee member asked if the first column, percent of high school diploma graduates, included people with General Education Diplomas (GEDs) as well as homeschool diplomas. Dr. Velgach replied that it included those with homeschool diplomas but not GEDs.

In discussing congressional reports (slide 8), Ms. Miller informed the Committee that Congress continues to ask whether the ASVAB should still be required. She said AP attempts to convey how the test has evolved to keep pace with the environment so that it remains a useful selection device. She said they explain that changes to the battery are not made lightly but are based on considerable psychometric evidence. She thanked the committee for helping ensure the battery remains reliable and valid. A committee member then asked if Category IV applicants are accepted, and Ms. Miller replied that they are, but that Category V applicants are not. Dr. Velgach commented that Congress has been interested in the number of applicants that fall into Category V. Ms. Miller attributed this to the Department of Education's (DOE) interest in knowing the percentage of people who cannot achieve Category IV. She said DOE wants to use the ASVAB—due to its immense reputation and the number of people who take the test—as a means of evaluating the performance of high schools. She confirmed, however, that AP's stance is that the ASVAB was not designed for that purpose and therefore, they do not encourage the use of ASVAB for high school evaluations. On hearing this, a committee member noted that not everybody takes the ASVAB, it is voluntary. Ms. Miller agreed and said AP would continue to hold the line and would keep the committee informed.

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

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

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

- New Computer Adaptive Testing (CAT)-ASVAB Item Pools. The objective of this project is to develop CAT-ASVAB item pools 11 – 15 from new items. New form implementation is projected for September 2020.
- Developing New CAT Item Pool for the CEP. The objective of this project is to build a CAT pool from paper-and-pencil Forms 20B, 21 A&B, and 22 A&B for implementation of the Internet CAT-ASVAB (*i*CAT) in the CEP. The new pools will be implemented in the fall of 2019.
- Automated Generation of Arithmetic Reasoning (AR) and Mathematics Knowledge (MK) items. The objective of this effort is to develop procedures for automating AR and MK item generation so that AR and MK pools can be replaced on a more frequent basis. Anticipated completion date is March 2020.
- Automated Generation of General Science (GS) items. The objective of this effort is to develop procedures for automating GS item generation so that GS item pools can be replaced on a frequent basis. The projected completion date is September 2020.
- ASVAB Technical Bulletins. The objective of this project is to develop a series of electronic ASVAB technical bulletins to meet American Psychological Association (APA) standards. The project is ongoing.
- CEP. The objective of this project is to revise/maintain all CEP materials, conduct program evaluation studies, and conduct research studies as needed. The project is ongoing.
- Evaluating New Cognitive Tests.
    - Mental Counters (MCt). The objective of this project is to conduct a validity study to evaluate the benefits of adding MCt to the ASVAB and provide data to establish operational composites that include MCt and operational cut scores for new composites. The Navy is taking the lead. Completion schedule is to be determined (TBD).
    - Cyber Test, formerly the Information/Communications Technology Literacy (ICTL) Test. The goal of this project is to develop and evaluate the Cyber Test. The Air Force is the lead, and the project is ongoing.
    - Nonverbal Reasoning Tests. The objective of this project is to address the ASVAB expert panel's recommendation to investigate the use of a test of fluid intelligence, such as nonverbal reasoning, and to plan and conduct construct validation studies. Project completion is TBD.
- Adding Non-Cognitive Measures to Selection and/or Classification. The objective of this project is to address the ASVAB Expert Panel's recommendation to evaluate the use of non-cognitive measures in the military selection and classification process. The measures being evaluated include the Tailored Adaptive Personality Assessment System (TAPAS); the Work Preferences Assessment (WPA); and Army, Air Force, and Navy interest inventories. The project is ongoing.
- AFCA. The objective of this project is to program the AFCA for Windows-based CAT (WinCAT) administration. Project completion is TBD.
- DLAB. The objective of this project is to transition to all computer-based testing and improve the predictive validity of the DLAB.
- Expanding Test Availability: Web/Cloud Delivery of Special Tests. The objective of this effort is to transition delivery of special tests from the Windows-based platform to a web-based and/or Cloud platform. The anticipated completion date is December 2021.

As Dr. Pommerich briefed progress on developing new CAT-ASVAB item pools (slide 3), she mentioned that DPAC was focused on beating the software freeze. When a committee member

inquired about the meaning of "software freeze," Dr. Pommerich explained that it was required by the move to the Cloud, which she said she would talk more about later. She added that the move to the Cloud was DPAC's highest priority, and that it was affecting much of DPAC's psychometric work. Before moving to the next slide, Dr. Pommerich explained to the new committee member that DPAC sometimes uses the terms "pool" and "form" interchangeably, though the committee prefers the term "pools" in the context of a CAT environment.

After Dr. Pommerich reviewed progress on the automated generation of GS items, a committee member asked if DPAC was considering the use of automation in other areas. Dr. Pommerich replied that DPAC was addressing this through the ASVAB evaluation plan. She explained that DPAC is deriving a ten-point scale for rating the ASVAB subtests for automated item generation (AIG) suitability, and that the most promising tests were those for which AIG is already being applied. She said using AIG for other tests appears to be a tricky matter, but that Paragraph Comprehension (PC) could be a candidate for certain item types, but that generating items that require reasoning or inference might be difficult. She also said technical subtests with graphics content would present challenges. Dr. Pommerich concluded by explaining that DPAC was attempting to rank-order the tests on AIG suitability, but that any further work in AIG would have to wait until after the move to the Cloud.

Commenting on the perceived utility of the AFQT Predictor Test (APT; slide 9), Ms. Miller recalled that recruiters perceive the APT as being less accurate than the Enlistment Screening Test (EST[1]), perhaps because the APT is unproctored. She said she had discussed this matter at a recent recruiter conference, where she tried to explain that the APT was the better predictor. Dr. Velgach replied that the APT, being unproctored, is only as accurate as applicants make it, and that DPAC has been brainstorming on measures to test environment security. Dr. Pommerich then clarified that DPAC does not have the data required to directly compare the tests. She said they have a database of EST scores from the Marine Corps, but it does not include cases in which predictor scores are below 31. She said DPAC needs an EST dataset that is unconstrained, so they can better compare the tests. She also said DPAC has thought about administering the APT under proctored conditions, because evidence indicates that applicants treat unproctored tests differently; she said people sometime get help in unproctored conditions. She concluded by saying that the issue keeps resurfacing, and they have been unable to make headway.

A committee member asked about the differences between WinCAT and the other delivery systems. Dr. Pommerich said WinCAT is the local area network Windows system provided at all 65 Military Entrance Test (MET) sites, but that the Military Entrance Processing Command (MEPCOM) has been directed to decommission WinCAT, which is being planned to occur in conjunction with the move to the Cloud. Dr. Segall clarified that there are a few special tests (i.e., special Service tests, such as the Cyber Test) administered on WinCAT, and that the plan is to move these tests to *i*CAT first and, then, to move all of *i*CAT to the Cloud. Ms. Miller remarked that this was all part of a larger effort to reduce the number or servers used across the Department of Defense (DoD) and to house everything in the Cloud.

As Dr. Pommerich addressed the software freeze (slide 30), a committee member asked if there would be a no-testing period. Dr. Pommerich said testing would continue uninterrupted,

[1]The EST was developed by the Navy around 1990 and was later adopted by all Services.

explaining that DPAC had previously received pushback for shutting down testing for only two and a half days for database upgrades. Dr. Segall reiterated that there was no plan to suspend testing. He explained that decommissioning WinCAT would occur over a three-month period, during which time there would be backups available. Dr. Pommerich called this mandatory "redundancy," and said she was trying to get a schedule update from MEPCOM for decommissioning. Ms. Miller said the current dates are still accurate, but that she was concerned about slippage to the right, which could delay full transition to the Cloud by 2021.

Reflecting on Dr. Pommerich's comments about the difficulty of moving the TAPAS to the Cloud (slide 30), a committee member asked if the final report on the TAPAS study was available. Dr. Velgach said it was still being finalized.

3. Abstract Reasoning Evaluation (Tab F)

Dr. Furong Gao, HumRRO, presented the briefing.

Dr. Gao began by highlighting recommendations from the ASVAB Expert Panel regarding the incorporation of non-verbal reasoning tests in the ASVAB and the characteristics of the Abstract Reasoning Test (ART) that make it suitable for this purpose. Dr. Gao then presented a sample ART item. This was followed by a summary of findings from past research that show a strong relationship between ART and Raven's Advanced Progressive Matrices scores, similar patterns of relationships with ASVAB subtests, and that ART was found to load on ASVAB quantitative reasoning factor with AR, MK, and Mechanical Comprehension (MC). The ART includes 30 items that are scored right or wrong and has a 25-minute time limit. New analyses were carried out on data from 2,162 test takers who were military applicants interested in language training who had already qualified based on AFQT scores and were highly motivated and high ability. The tests were administered between March and September of 2017. Other available data on the test takers included ASVAB scores from their enlistment profiles taken 1-3 years before and scores on the MCt.

Dr. Gao presented charts indicating that 80% of the items had a *p* value of .75 or greater, and the mean raw score was 24.1, with a standard deviation of 4.2 and reliability of .803. Raw score distribution comparisons by gender, years of education, and race/ethnicity yielded small effect sizes (-.04 to .23). Additional tables showed that, by contrast, effect sizes seen when comparing ART score distributions by race/ethnicity to those of GS and the AFQT, results were generally small (.23, -.18, -.04). Dr. Gao then turned to analyses examining the relationship between ART, the ASVAB, and MCt. She provided summary statistics for 1,724 cases where scores were available on all three measures. The attenuated correlation between MCt and ART was .52. Correlations with ASVAB subtests ranged from .09 (Auto/Shop) to .43 (AR), and the correlation with AFQT was .46. Similar findings were found for MCt and ASVAB scores. Confirmatory factor analyses were performed with the factor structure superimposed on a subset of the Broad (Stratum II) Cattell-Horn-Carroll ability definitions. Both MCt and ART loaded on a general *g* factor as well as a fluid intelligence/reasoning factor.

IRT analyses were carried out using a three-parameter logistic (3PL) model fitted using BILOG-MG. Overall, 25 items (83%) showed adequate model fit at a significance level of .01. The estimated test-retest reliability was .77. Examination of response times indicated that 98.2% of respondents completed all items, and 99.3% completed the second to last item.

Dr. Gao concluded by acknowledging that the evaluation was done on a limited sample of high-ability test takers. The items appear to be easy to administer. The reliability results likely reflect the lower bound due to the restricted sample. ART appears to measure a unique domain not currently represented in the ASVAB. Test results appear to have small impacts across demographic groups, however the 25-minute time limit may not be adequate. Overall, the results appear promising, but further study is required using more representative samples. Therefore, the recommendations are to (a) conduct further evaluations with more representative samples, (b) increase test time to 30 minutes and reevaluate when more data have been

obtained, (c) develop a similar test of complex reasoning with more difficult items if additional ART results confirm it is too easy, (d) investigate the feasibility of using AIG, and (e) develop a computerized adaptive version of the test.

On slide 6, Dr. Gao mentioned the moderate correlations between the ART and the AR, MK, MCt, and the AFQT. A committee member pointed out that they might expect the ART to show relationships with AR, MK, and the AFQT, because AR and MK are part of the AFQT. Dr. Gao agreed.

On raw score distribution by item difficulty (slide 10), a committee member observed that the last items administered had lower p-values and asked if test-takers may have run out of time. Dr. Gao replied that time constraints may have been a factor, but that the p-values were calculated without including items that were not answered. Noting that the latter items were more difficult, another committee member asked if the more challenging items had been placed at the end by design. Dr. Segall replied that it could have been by design and explained that the test had been developed by Dr. Susan Embretson. Another committee member asked if the raw scores appearing on the left of the raw score distribution chart were from people who did not understand the instructions. Another committee member added that s/he was surprised to see a score of zero. The previous committee member noted that zero was below the guessing rate, and that the lowest scores should have been around six. Another committee member asked if the items were always presented in the same order, and Dr. Gao said they were. Discussion concluded with a committee member's question about whether the reliability estimate was internal consistency reliability; Dr. Gao said that was correct.

On raw score distribution by gender (slide 11), a committee member asked if the sample was the same as what was shown on slide 8 (i.e., military applicants who wanted to take language training). When Dr. Segall said that it was, the committee member asked what had happened to the other 1,000 cases. Dr. Gao said that the sample on slide 11 was limited to those whose scores could be matched to ASVAB scores.

When Dr. Gao briefed the ASVAB GS score distribution by race and ethnicity (slide 14), a committee member asked if Dr. Gao was providing that information because it was representative of other ASVAB subtests. Dr. Pommerich explained that the technical tests, AI, SI, EI, and MC, are those that show the greatest effect sizes for race and ethnicity. She said GS has a greater verbal component, which should lead to moderate effect sizes. She said the point was to show a comparison between the ART, which has a low effect size, and a test that has a larger effect size.

When Dr. Gao explained the factor loadings on Stratum I abilities (slide 22), a committee member sought clarification about the meaning of the stray dot shown for the ART. Dr. Segall said it represented an outlier and clarified that the chart showed the distributions of item factor loadings and that the one item had a loading of close to zero.

Regarding the item-response theory (IRT) analysis (slide 25), a committee member asked how the test-retest reliability estimate was obtained. Dr. Segall said it was generated through simulation in BILOG. The committee member then asked for an explanation of the graph. Dr. Segall said the solid line revealed that the test was easy for the restricted population, and that

they were trying to determine if it would be too easy for an unrestricted population. The committee member pointed to the importance of determining whether the test provides information where it is needed. Dr. Segall concurred but explained that new items were needed in order to make the test adaptive. He said the analysis should identify the levels of difficulty at which new items should be developed. Another committee member emphasized that the test, in its current form, was too easy to be informative. Dr. Segall agreed and explained that the situation was even more complex, because test scores would be used as part of a composite.

Continuing the discussion, a committee member asked if there were any item features indicative of difficulty. Dr. Segall said Susan Embretson had developed the model, but that DPAC was working on a redefined model they hoped to use going forward. Another committee member said it would be interesting to know in what ways the non-fitting items were different, because such a large percentage of items failed to fit the model. The committee member said s/he was not sure what the drivers could have been. Another committee member, returning to the purpose of the test, asked if it was designed for a high or low ability group. Dr. Segall replied that it was for everyone, but that the existing data just happened to be taken from a high ability sample (i.e., language training applicants). He said the test was administered to this sample because, due to its low verbal loading, the test would minimize adverse impact for native Spanish speakers. Ms. Miller confirmed that the test was used to identify candidates for language training, and Dr. Velgach clarified that Assembling Objects (AO) was not performing very well in that capacity. The committee member then asked if the test could be administered to a sample from the intended population. Dr. Segall said that possibility existed. The committee member then remarked that the tail of the distribution made it difficult to tell how the test would perform without a targeted administration.

As Dr. Gao presented the response time analysis (slides 27-28), a committee member said s/he would not expect a response time distribution to be normal, but that with this many items, it appeared to approximate normal. Another committee member questioned why so many test-takers were running out of time, given that the test was so easy. Dr. Gao replied that it was likely due to the large number of items, and said she was recommending that test time be increased. A committee member said that would be important for a broader population.

In response to the recommendations (slide 30), a committee member asked why DPAC was not using the Raven's Advanced Progressive Matrices test. Dr. Segall replied that it would be expensive to use that test, and said they wanted a test that was more secure. He explained that the need for security and the associated need for a larger item pool was the reason they were looking into the use of AIG. Another committee member asked what the test was currently used for, and Dr. Segall said it was not currently used for any operational purpose. He added that the test had been a finalist for use in predicting success in language training—for which it was validated—but that it could be used for predicting success in general military training because it is so *g*-loaded. The committee member said part of the test's value was in reducing adverse impact, but another committee member said the jury would be out on that until it is validated with a representative sample. Dr. Velgach reiterated that using the test would reduce reliance on verbally-loaded tests, and Dr. Segall said it would also be less prone to compromise. He said the PC test is easy to compromise and uses items that are difficult to develop.

As the briefing concluded, a committee member remarked that the test looked pretty good, especially in its capability to reduce adverse impact. S/he restated, however, the issues associated with reliance on a restricted sample. Another committee member expressed concern about how well the various demographic groups were represented in the samples, noting that some groups were small. Dr. Velgach said she thought the current research was a good start. Another committee member raised a final issue, suggesting that a 2PL model might be preferable to the 3PL model the test currently employed. S/he explained that the 3PL model might be causing a misestimation of the *c* parameter. This sentiment was echoed by another committee member, who suggested trying the 2PL approach.

4. Cloud 101 (Tab G)

Mr. Matthew Ellis, Northrup Grumman Systems Corporation, presented the briefing.

Mr. Ellis began by providing a definition of Cloud computing. "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal effort or service provider interaction. The cloud model is composed of five essential characteristics and defines three service models and four deployment models. The three service models are Infrastructure as Service, Platform as Service, and Software as Service. The deployment models are private, public, community, and hybrid. Mr. Ellis continued by presenting a chart showing the benefits of cloud computing. These fall into three categories: (a) efficiency (e.g., improved productivity), (b) agility (e.g., near instantaneous increases and reductions in capacity), and (c) innovation (e.g., better linked to emerging technologies).

Mr. Ellis continued by providing an overview of the DoD Cloud Strategy. The guiding principles are mission first, cloud smart-data smart, leveraging commercial/industry best practices, and creating a culture better suited for modern technology evolution. Within DoD it is acknowledged that every information technology (IT) system includes some element of risk (e.g., information loss, illicit entry, breach of personal information). Therefore, IT systems require approval to operate, which is designated in an Agency Authorizing Office memo. Guidance is provided by the Risk Management Framework from the National Institute of Standards and Technology. The Authorizing Office must be cognizant of the risks introduced by IT systems. The Cloud introduces risks that must be managed, and the Federal Risk and Authorization Management Program (FedRAMP) and Defense Information Security Agency (DISA) have established standards for Cloud computing in DoD. In selecting a Cloud Solution Provider, approvals are required from FedRAMP, DISA, and Office of People Analytics (OPA). The pillars of the secure Cloud computing architecture (SCCA) include the Cloud access point, a Cyber Security Services Provider (including virtual data center management and a virtual data center security stack), and a trusted Cloud credential manager.

The DPAC business case analysis established the reason for moving to the Cloud, which included improved reliability and availability along with increased scalability. It also specified the scope of the effort: the ASVAB and language testing applications. The analysis also established impact level/information security requirements, an included a trade study of Cloud Service Providers. Agreement was reached to migrate a pilot set of applications to Amazon Web Services. It was also agreed that all IL 4 Sensitive data protection requirements can be successfully satisfied through Amazon's Authority to Operate (ATO), DISA's SCCA, and DPAC's shared Information Assurance (IA) responsibility. The goals for the DPAC application migration strategy include expedited cloud migration, reduced licensing costs, reduced system administration, and improved reliability and delivery. Mr. Ellis continued by showing a timeline for Cloud migration that envisions all activities to be completed by September 2020. He concluded by noting some of the challenges associated with this effort, including (a) the fact that migration freezes new feature developments for a period of time and (b) the changes introduced to help desk via the Defense Manpower Data Center (DMDC) and DPAC with cloud-based applications. At the same time, cloud migration provides opportunities for improved

system availability and reliability and reduces Military Entrance Processing Station (MEPS) travel impacts of system outages.

As Mr. Ellis explained the vast amount of storage space available with the Cloud (slide 7), a committee member commented on the size of 15 petabytes, which Mr. Ellis said was fifteen thousand times larger than a terabyte. Dr. Velgach also commented that one of the most important benefits of using the Cloud was the capability of the large providers (e.g., Microsoft) to keep pace with cyber security challenges. She remarked, however, that if the entire system goes down, DoD would be in trouble. She also emphasized the expense DoD would incur if it had to provide the expertise needed to execute the strategy internally.

5. Social Media Project Update (Tab H)

Dr. Tim McGonigle, HumRRO, presented the briefing.

Dr. McGonigle began by providing an overview of the goals of the project. In private industry, social media has changed the way recruiters find and attract applicants. It is used in all phases of the process. Data show that 70% of employers screen candidates using social media. Private industry, as well as the military, currently use social media data idiosyncratically. Currently, military recruiters face significant challenges given historically low unemployment, low propensity to enlist, low eligibility, and a lack of knowledge about military service. As a result, DPAC requested that a project team be formed to consider how social media can help with military recruitment and selection. The team will consider pre-employment steps, including attracting candidates by disseminating positive information about the military, generating high quality lead lists by identifying candidates with favorable characteristics, and making decisions about applicant qualifications. The latter involves many potential technical, psychometric, and legal issues. The plan calls for the team to include individuals with expertise in Industrial/Organizational Psychology, data science, law, and ethics. The team will meet four times to address questions about the use of social media data in military recruiting and selection and advise on a research and development agenda. Dr. McGonigle continued by outlining the qualifications of each of the seven team members.

The team's first meeting occurred on May 14, 2019. In addition to the team members, representatives from the following organizations were in attendance: DPAC, Joint Market Research Studies (JAMRS), AP, the Defense Personnel and Security Research Center (PERSEREC), the ARNG, the U.S. Army Recruiting Command (USAREC), the Air Force Recruiting Service (AFRS), the Marine Corps Recruiting Command (MCRC), the Navy Recruiting Command (NRC), the Army Research Institute for the Behavioral and Social Sciences (ARI), the Air Force Personnel Center (AFPC), and the Defense Science and Technology Laboratory of the United Kingdom (UK DSTL). The agenda focused on project goals and background and current use of social media data in military recruiting. After Dr. Dan Segall provided an overview of the project and its goals, ARNG, USAREC, AFRS, MCRC, and CNRC presented information on their current use of social media in recruiting. Following the meeting, detailed minutes were provided to the project team.

Dr. McGonigle continued by providing highlights from the various presentations.

- Dr. Segall discussed the use of social media in private sector recruiting and the challenges facing military recruiting. He explained that the purpose of the project is to provide input on pre-employment uses of social media, specifically selective recruitment and selection, in a military context. This includes how to use social media and what characteristics are predictable through its use; legal, ethical, public relations, and technical considerations for social media use; and how to supplement lead lists and prioritize leads, but not to disqualify candidates. Dr. Segall expressed interest in obtaining feedback on questions for the project team and research that should be conducted before social media data are used.

- The representative from the ARNG discussed the use of several social media platforms to provide realistic previews and engage potential candidates and answered questions about managing friend requests (they become fans of the page) and connecting candidates to recruiters (call center transfers information to the Army Lead Processing System).
- The representative from USAREC discussed their Virtual Recruiting Center for national leads and the Virtual Recruiting Stations for every battalion, which are used to generate new leads, to try to dispel myths about the military and Army, and to refine leads by gathering social media data. They currently gather social media data manually at the Virtual Recruiting Stations and look for information to start recruiting conversations
- The representative from AFRC indicated that they target paid marketing and human efforts based on social media profiles (of those who have interacted with Air Force content only). They use social media to measure sentiment, value, and scale of posts, and use this information to determine which topics generate the most interest. They are interested in learning about handling Personally Identifying Information (PII), direct messaging techniques, and employment law.
- The representative from the MCRC said they are concerned with identifying candidates who will not drop out of the selection and training process. Social Media posts are image-based and meant to garner interest in service. They create lists of users who like, comment, or reblog Marine Corps social media posts to supplement other lead generation methods. They are interested in information on how to increase the number of leads generated or how to reduce the time spent on recruiting each candidate.
- The representative from the NRC indicated that local recruiters use social media only after they receive leads from the national level or after they meet people in person. He conducted a live demonstration of social media use on Facebook, Instagram, Snapchat, and Reddit. Instagram is currently best for number of leads and interviews generated. He noted that recruiters are given feedback about the quality of their social media activities and supervisors monitor benchmarks. He also clarified that he contacts all leads regardless of whether he believes they are qualified but is more aggressive if he believes that someone would be a good candidate.

Dr, McGonigle continued by providing a sample of the technical and measurement topics that arose during the meeting.

- Predictive models built on one social media platform will have around 80% validity when applied to other platforms.
- What amount and quality of information is required to predict personality using social media data?
- Many of the correlations between social media data and traditional measures are significant but very small; social media is not always the better predictor.
- Test-retest reliability of social media models is about .70 for a six-month interval and about .60 after two years, which is similar to traditional measures.

Some of the legal and ethical topics that surfaced included:

- The effects of gendered language use on text analysis.
- It is not legal to consider protected demographic characteristics when making selection decisions or to use within-group norming.
- How social media information could be used in an effective way given that law and the political environment do not allow the use of social media models in selection decisions. Data cannot be collected from a third party.
- How would the Services access social media data for use in these models given that most platforms limit data scraping?
- Does the fact that potential recruits may be minors change the legal considerations?

Finally, among the "vision" topics addressed were:

- Focus on "screening in" rather than "screening out" candidates.
- Avoid mental health or security risk assessments via social media data.
- Focus on predicting propensity to join the military and scores on selection measures, not on using social media as a replacement for current measures.

> Dr. McGonigle concluded by stating that the next meeting will be on October 22, 2019, and will include presentations from JAMRS, PERSEREC, ARI, and panel members. The third meeting will be held in late 2019 or spring 2020 and will focus on legal and ethical issues. The final meeting will be in the summer of 2020 and will result in recommendations and a research and development roadmap. The final report on the topic will be delivered in the fall of 2020.

As Dr. McGonigle presented the results of the open discussion (slide 13), a committee member asked if the intent was to use social media broadly in prediction, or if there was a primary focus. Dr. McGonigle said the focus was currently on predicting personality characteristics as a precursor to predicting job performance. Dr. Segall clarified that much work has already been done in the area of using social media to predict personality. He stated that the military's focus would be on predicting TAPAS scores, or perhaps ASVAB scores, especially the AFQT. He explained that this could help recruiters prioritize their recruiting efforts.

Dr. Velgach commented that DoD is considering social media as a tool for targeted exploration. She said, for example, attendance at dental conferences may result in identifying people who need to pay off large student loans and that those people could then be targeted. A committee member asked if social media data were processed clinically or via statistical algorithms. Dr. Segall said they were processed algorithmically. The committee member then asked how negative information might be used. Dr. Segall replied that there are differing opinions within OPA on this matter, but that DoD is investigating uses of publicly-available social media for personnel security assessments and background investigations. He added, however, that negative information is not ignored.

Ms. Miller commented on DoD's apparent apprehension about using the information to make decisions. Dr. Segall said his preference would be to use it for that purpose, but that legal had reservations about using non-adjudicated information to make conclusive decisions. Ms. Miller then asked if using social media data might discriminate against non-user populations. Dr. McGonigle replied that there are some very clear ethical/legal issues in this area. A committee member asked if the issues existed in the context of using social media for prioritization purposes. Dr. McGonigle indicated that was a big question the group was facing: what do you do if you have no or limited information? Another committee member commented that some people live in areas that have little or inconsistent access to the Internet or cell reception. The first committee member clarified that some people may not have access, but that others just choose not to use social media. Dr. Velgach pointed out that children of Federal employees have access to training on the proper uses of social media through their parents, which might cause them to use it differently than the general population. Dr. McGonigle replied that these conditions could be partially mitigated by employing a focus on screening in versus screening out. He added, however, that the concern regarding differential use is real.

A committee member then asked if anyone was keeping data in a systematic fashion. Dr. Velgach responded that data are maintained at the Service level, where support exists to help evaluate its uses. A committee member commented that the use of social media data looks like a promising component of the larger campaign to generate interest in the military.

6. Automatic Item Generation (Tab I)

Dr. Isaac Bejar, ETS, presented the briefing.

Dr. Bejar began by providing the names of the project staff working on each of the AIG efforts. He then showed a schematic summarizing how the Word Knowledge (WK) item generator works, indicating that it has been delivered to DPAC. This was followed by an overview of the tasks conducted in developing AIG for MK and AR.

Dr. Bejar went on to explain that an item model is a template that, together with the appropriate software, produces items intended to be of the same difficulty. Each item model includes a set of constraints that limit precisely the items produced by that model. Typically, an item model is based on an existing and calibrated item. When properly authored, the items generated by an item model have similar difficulty and discrimination parameters so that all the items produced from a given model can be pre-calibrated as if they were a single item. Dr. Bejar then displayed a schematic highlighting the major steps in the workflow, along with a sample MK item and item model. He indicated that the MK field test design involved using items from the MK7 and MK3 item pools to develop 50 item models and five items per model with different keys, for a total of 250 items. Four models had graphics which were made part of the model. The results were evaluated at the item level (i.e., How many of the 250 items are within difficulty and discrimination ranges?), at the model level (i.e., What proportion of models behave as expected?), and regarding cost effectiveness (i.e., What is the cost of each generated item?).

Dr. Bejar presented a table showing the *A, B,* and *C* parameters for items in each of the content categories (i.e., algebraic operation and equations, geometry and measurement, number theory, and numeration). These suggested a possible decline in discrimination and increase in difficulty over time. He then listed some possible contributors to parameter estimate variability, including estimation details (i.e., LOGIST vs. BILOG), the composition of the incoming testing population, position and context effects, and curricular trends.

Dr. Bejar continued with an overview of the item modeling results. At the item level, all but 13 of the 250 items met difficulty and discrimination criteria. Dr. Bejar then presented several graphs detailing the results of model analyses and distractor analyses.

Turning to cost-effectiveness, Dr. Bejar stated that the total number of items that could be generated from 50 models is conservatively 500, but the isomorphs of the same model could not appear in the same item pool. Accounting for authoring, reviewing, and graphic modeling, the time per item is .6 hours. If only items from working models are used, this increases to 1.2 hours per item, with actual costs depending on such factors as salaries and overhead. He concluded the MK discussion by indicating that nearly 100% of the items were functional, but the percent of acceptable models (50%) could have been higher. Dr. Bejar said acceptability would increase with what he knows now: (a) pool 3 had many more 5-choice items, (b) the success rate was 60% with 4-choice items, (c) rotating distractors could have an effect on *C* parameters, (d) it is necessary to review the distractor analysis before modeling, (e) avoid modeling items with *a*'s lower than 1, (f) low *b*'s are much less likely to work, and (g) the cutoff for declaring a model unsuccessful requires additional research into whether it is possible to compensate for random variation.

Turning the discussion to AR, Dr. Bejar began by showing a sample item from the test. He said that the AR field test is still ongoing, with a projected completion date of November 30, 2019. In all, 45 models were developed, yielding 225 items. The criteria for evaluating the success of the effort included the item-level yield, the model-level yield (i.e., within model was difficulty held constant, between models was difficulty successfully manipulated), and cost effectiveness.

Dr. Bejar then turned to the GS AIG effort, and displayed a table showing the various tasks to be completed. He indicated that GS covers several domains, however the scope of the field test will be limited to anatomy and physiology and zoology. The approach is to infer the construct from analysis of existing GS items and identify high-level item models. Based on items from pools 3 and 7, several high-level item

models were identified. Generating GS items requires a biology knowledge base or ontology suitable to K-12. After trying one that was developed for the purpose of answering K-12 science questions, ETS determined they would have to generate their own ontology. The approach to knowledge representation involves a semantic web and uses a Resource Description Framework (RDF) involving semantic triples. A top-down approach involves starting from an ontology and using semantic triples (node, relation node) to manually extract the triples. A bottom-up approach uses text patterns to identify subject, predicate, object triples. The steps include identifying sentences with target vocabulary, applying extraction patterns, collecting triples, and conducting SME reviews. The development of a GS ontology requires creating vocabulary, producing SME-generated triples, and experimenting with information extraction. The SME-generated triples are in progress, with a goal of uploading items for field testing in January 2020. Information extraction will be implemented to compare the two approaches.

As Dr. Bejar briefed the approach to modeling MK and AR items (slide 7), a committee member pointed out that changes to the response options, in addition to changes to the stem, may affect the comparability of an item and its parent item. The committee member mentioned, specifically, that the degree of differentiation among response options would factor into comparability. Dr. Bejar then showed a sample item and its parent item, at which point a committee member asked how the items would be generated; that is, would the first two elements (i.e., N1 and N2) be iterated, and the third element (N3) generated. Dr. Bejar said that he was unsure what the item generator did internally—aside from the constraints—but that all the configurations could be produced using those constraints in Excel. The committee member suggested the need for another constraint, which Dr. Bejar clarified as being a constraint on the position of the item key, specifically, that the item key should rotate. Another committee member mentioned that the example included five distractors, but s/he said only three are needed. S/he asked if all the distractors would be equally attractive, such that distractor selection could be random and the resulting sets equally attractive. Dr. Bejar said that was the challenge, and that it does not always happen. He added that there was no distractor analysis component, which meant that attractiveness judgments were a subjective matter. The first committee member asked if the objective was to generate items and place them in the pool without review. Dr. Bejar said that would be the best possible case. He referred to item modeling as being like an art, such as writing. He said that if one can constrain the moving parts, there is hope for constraining the psychometric properties.

On the presentation of MK item set difficulties (slide 13), a committee member observed that the most difficult items dealt with primes, and the least difficult items dealt with place values. S/he then asked if the content areas shown were drawn from the ASVAB. Dr. Bejar said they were.

When Dr. Bejar presented slide 16 and mentioned that he was looking for another data set with different parameter estimates over time, a committee member said he should check with Graduate Record Examination (GRE). Dr. Bejar appreciated the suggestion and then noted the stability of the c parameter. When the committee member said that was a pretty good result, Dr. Bejar replied that he hoped it could be written up.

Regarding the model level analysis (slides 20-25), a committee member said the separation among root-mean-square deviation (RMSD) curves occurs when something aberrant is happening. S/he said it looked like the *a* parameter was not working well, and that it would be instructive to look at the specific incident to see what was going wrong. Dr. Bejar replied that he had examined the case but had failed to identify any obvious explanation. He said his hypothesis

was that the problem was caused by going from four to five response options. He clarified that the success rate for sets with four options was 60%, while the success rate for sets with five options was only 50%. He said he had been unable to explain what was happening. He then commented that the five-option sets did not appear to provide neat, dichotomous items and suggested that a distractor analysis would be valuable. A committee member said that having the models would allow Dr. Bejar to examine the functioning of the model, hopefully to see that all the distractors were equally plausible at some low end. The committee member said, across variants of a model, he might be able to identify if there was a single model distractor that functioned very well. Dr. Bejar said he would investigate this. The committee member added that the prevailing item writing guideline is that distractors be equally plausible, but that each should catch different errors in problem solving that people actually make. Dr. Pommerich remarked, however, that the items would only look as good as the parent item, and that there was a limited set of parent items. Another committee member observed that, in the bottom row of graphs (on slide 25), the first graph had five distractors and the remaining graphs only had four. Dr. Bejar explained that the analyses shown were limited, and that he wanted to extend it to the rest of the 50 models, in which he hoped to see a pattern of more well-behaved models, that is, models with greater equivalence among distractors.

While presenting his conclusions on the MK effort (slide 28), Dr. Bejar said he would have like to have seen a higher level of model acceptance. Upon further inquiry by a committee member, he explained that he might not have thought carefully enough about parent item selection, but that he wanted to cover the content sufficiently. He added that more thought should be given to what works and what does not work.

As Dr. Bejar spoke on the process of autogenerating AR items (slides 30-32), a committee member commented that varying the numbers in the stem could affect difficulty. Dr. Bejar agreed and proposed, as an example, that a quarter hour might be treated differently than a half hour would be treated. A committee member said s/he thought the intent was to vary difficulty so as not to produce isomorphs. Dr. Bejar explained that they start with fifteen functioning items, and, depending on where the item was to begin with (say medium difficulty), then they make a model that is either easier or harder. He then said it will be interesting to see what happens.

Regarding the auto-generation of GS items (slides 38-44), a committee member said s/he recognized the K-12 effort that Dr. Bejar cited, stating that it was reported in the New York Times. The committee member then said the Resource Description Framework (RDF) was reminiscent of concept maps in science. Dr. Bejar agreed that the concept mapping approach was similar to what he was doing. He described his method as not being “brute force,” and added that, early on, he was told to verify how each item was developed; that is, to verify that the content was based in fact.

As Dr. Bejar presented the current status of the GS effort (slide 44), a committee member asked if his intent was to place the items that work into an item bank and draw them into test forms based on difficulty. Dr. Bejar said that would be one way to do it, but that the notion of tracking instances of a model was challenging. The committee member then raised concern about including two items from the same model in the same test; that is, the presence of the first item would moderate the difficulty of the second item. Dr. Bejar agreed that, in that case, local

dependence would be a problem. Dr. Segall concurred, saying that the intended precision in difficulty would be lost. The committee member asked if the system would mitigate such instances. Dr. Segall said it would and that the item banking system was being refined to track enemy items. He said such tracking is now performed manually. Dr. Pommerich added that enemy items are currently identified at the item writing and item analysis stages. She said they currently use human and algorithmic approaches. Another committee member then raised the previously discussed issue concerning differences between the four- and five-option item formats. Dr. Bejar said he was aware of the problem and was attempting to address it. Another committee member asked if the AIG process could be improved with practice. Dr. Bejar said, over time, it could get better, mainly due to the improved ability to conform to standards. He said his intent was not just to crank out a lot of items, but to increase the extent to which the items meet the standards. He then commented that either the item parameters should be held constant, or they will learn how to manipulate them.

7. CEP Update (Tab J)

Dr. Shannon Salyer, Manager, Career Exploration Center, presented the briefing.

Dr. Salyer began by presenting ASVAB CEP numbers and metrics for school years 2013 through 2019. These showed the number of students tested as ranging from 670,836 in 2013 to 786,807 in 2019. The percentage of schools tested ranged from 55% in 2018 to 60.6% in 2019. For school year 2018-2019, 91 percent of CEP ASVAB administrations used paper-and-pencil (P&P), with 9 percent being computer administered. This represents a 2 percent increase in *i*CAT administrations over the previous year. Dr. Salyer then showed data on the number of leads provided to military recruiters through the CEP from 2014 (492,419) to 2019 (468,003). The number of accessions who used their CEP scores for military entrance ranged from 28,233 in 2014 to 30,257 in 2017. In 2019 the figure was 28,614.

Turning to usage figures for the CEP website, Dr. Salyer displayed numbers that showed increases from 2017-2018 to 2018-2019 in unique visitors, returning visitors, page views, and mobile users, while bounce rates decreased and average time per session and number of pages view per session were relatively static. Dr. Salyer then presented data on CEP website access code use from July 1, 2018 to June 30, 2019. In all there were 251,704 visitors and 101,731 repeat visitors. Data on Careers in the Military (CITM) website usage showed increases in the number of unique visitors (2017-2018 vs. 2018-2019) as well as increases in tablet mobile visitors. From July 1, 2018 to June 30, 2019 there were 1,212 inquiries through the CEP website Contact Us option and 1,469 Bring ASVAB to Your School requests (822 parents/students, 647 counselors). In addition, 2.038 requests for ASVAB scores were received. A total of 99 responses were received through the CITM Contact Us option.

Dr. Salyer then turned to recommendations from the ASVAB CEP Expert Panel that was convened in 2017 and actions taken on those recommendations.

- In response to a suggestion that various program functions be updated/automated, a new contract will examine processes that are currently handled manually, from scheduling to accountability, and gather requirements to determine how they can be accomplished more efficiently.
- The Expert Panel recommended that a review and evaluation of the Find Your Interest (FYI) inventory be conducted to ensure that items encompass critical, occupationally relevant tasks for high school students, and that it is culturally appropriate. A later presentation in this session will address this issue.
- In response to various suggestions from the expert panel regarding the CEP website, a new resource center was reconfigured and implemented to include integration of twitter feed and other social media. A more wide-ranging website reconfiguration is scheduled for 2021-2022.

- The panel believed that more effort should be made to market the CEP in professional journals and textbooks. In response, steps were taken to include the program in the National Career Development Association's publication *A Comprehensive Guide to Career Assessment.*
- In response to concerns that there is a lack of consistency in how post-test interpretations (PTIs) are conducted, training efforts were undertaken.
- The panel recommended identifying strategies to increase the amount of time students spend exploring options, including adding activities that encourage self-reflection. The PTI training included information on the websites. Recruiting commands were invited to participate in the train-the-trainer model which builds multiple opportunities for schools, counselors, and recruiters to use the CEP to inform students about their options.
- To address other recommendations of the panel, work is being done to develop a work values measure that links an individual's work values to occupations. Efforts are also underway to create a Successful Job Search toolbox to be incorporated into the CEP website, and options for providing a uniform credentialed career development training to ASVAB CEP administrators and interpreters are being explored.

Dr. Salyer continued by addressing an issue raised by the DACMPT in previous meetings, which is to identify how states are using the program in response to the Every Student Succeeds Act (ESSA). Ongoing efforts include monitoring state boards and departments of education websites to identify any mention of their use of the ASVAB CEP. Dr. Salyer then presented several charts and tables summarizing the findings of this work. She noted that several states have passed legislation requiring schools to provide the ASVAB CEP to high school students, however the legislation is not worded accurately. The military Services have been speaking to legislators about the program, but many of them are unaware of program updates. A stakeholder meeting was held August 15 that included representatives from AP, the USMEPCOM, military service liaisons, and members of the Defense State Liaison Office. Information on how states are using the CEP was reviewed along with Service initiatives and AP guidance. Dr. Salyer then called attention to an ASVAB CEP orientation event held in Indiana, where the program is included as a pathway to graduation. A three-hour program overview was provided to the 125 attendees, most of whom were unaware of the website offerings. After the event, 70% of attendees said they would use the program in the future. Dr. Salyer also provided a memorandum to the field regarding the appropriate use of the ASVAB CEP.

Dr. Salyer then outlined several reasons for implementing PTI proficiency training, including the fact that new functionality has been added to the CEP and CITM websites and the lack of consistency in which these sessions have been conducted in the past. PTI proficiency requirements include being nominated to become proficient, completing virtual training modules, being observed conducting a PTI effectively, and uploading proof of proficiency to a Moodle. She then outlined the goal of the training and the metrics to gauge success, including increased website usage, increased testing numbers, virtual and in-person training attendance, and additional access opportunities for recruiters. A total of 231 individuals have taken part in in-person training thus far; 1,487 accounts have been created for the virtual training; and 922 nominees have been added since the training was conducted. The virtual training consists of user authentication, learning objectives, multimedia content, concept checks and application activities, and an area to upload supporting documentation. The training portal contains a list of all people who are PTI proficient, an area to assign three-year access codes, a communication system for all people who are conducting PTIs across the country, and the ability to collect information about training needs. Dr. Salyer continued by providing an overview of the topics covered in the virtual training and details about the composition and elements of the in-person training. She concluded this portion of the presentation by summarizing feedback received about the training, including the fact that 75% of attendees did not know much about the CEP website until they participated in the online training, and 93% said they were satisfied or very satisfied with the experience.

Dr. Salyer then introduced #optionready, an online resource to provide information about the various uses of the CEP (e.g., explore post-secondary and career options). The site provides sharable content that school counselors can use to inform their community about the benefits of ASVAB CEP participation and encourage students to sign up. It also includes information about the score release options available to schools. There is a sharing portal for those who wish to upload photos and videos from PTI workshops to encourage peer-to-peer sharing. The goal of the #optionready campaign is to reach 1 million participants in the ASVAB CEP within one academic year, to correct misconceptions about the program and improve its

reputation, as well as to build awareness of the benefits of participation. Dr. Salyer presented a list of metrics that will be used to gauge the success of the campaign (e.g. number of landing page visits, number of downloads). Dr. Salyer then introduced the monthly toolkit, which is designed to make it simple to engage with ASVAB CEP on social media and increase student participation. She concluded by presenting a list of the national events (e.g., conferences, invited addresses) of which ASVAB CEP was a part in 2019.

As Dr. Salyer discussed the number of leads provided by ASVAB-CEP administrations to the Services (slide 6), a committee member asked if students had to grant permission for their results to be sent to the military. Dr. Salyer said that schools set score release options, but parents can opt-in or out, and students are able to let recruiting commands use their scores for enlistment purposes. Another committee member asked what percentage of schools did not agree to provide CEP scores to the military. Dr. Salyer said she would have to follow up on that. The committee member then asked if the test was given to high school sophomores, juniors, and seniors. Dr. Salyer said it was, and in addition to junior college students. She said sophomore results cannot be used for enlistment. Ms. Miller asked if the test can be taken any time during the school year, and Dr. Salyer said it can only be taken on the dates the schools administer it. She added that the Job Corps administers it more frequently than do schools. Another committee member asked if DPAC knows whether a score is from a first or repeated attempt. Dr. Salyer said they do not track that information, because they do not use social security numbers and the tests are administered in P&P format. The committee member replied that a practice effect may be in play for repeat test-takers, but Dr. Segall said that would only apply to AO scores, which are not obtained from P&P administrations. Dr. Velgach mentioned the possibility that a person may be taking the same version of the test several times, which she said could be problematic. Dr. Salyer said repeated administration of the same test was not of great concern, because if a person retests at a school, s/he often has a very low AFQT score, in which case the military is not interested in them anyway. Dr. Salyer also mentioned efforts to control test versions.

On accessions by Service (slide 7), a committee member asked how DPAC knows if a score used for enlistment is an ASVAB-CEP score. Dr. Salyer said it not easy due to privacy concerns, but they can use a person's name, date of birth, and school information. Dr. Velgach then commented that 18% of ASVAB-CEP-based accessions came from schools that selected Option 8 (i.e., schools that declined to allow scores to be sent to the military). Dr. Salyer added that, even if a school does not want to disclose scores to the military, the students know they can share their scores.

Regarding website utilization (slides 8-9), a committee member asked how DPAC tracked the number of unique and returning visitors. Dr. Salyer said each student has a unique access code, by which DPAC can track return visits. She also said the numbers shown include all traffic, to include counselors as well as students. She explained that Google Analytics allows DPAC to track the number of counselors that log in, as well as the types of devices used. Dr. Salyer said she needs to provide an access code for one of the committee members.

As Dr. Salyer presented utilization of CITM (slides 10-12), a committee member asked how many pages were available in CITM. Dr. Salyer replied that OCCU-Find alone includes over 1,200 pages, and that those branch off for each Service. In response to another committee member's question about CITM, Dr. Salyer explained that students receive their scores and percentile scores on each subtest. She said students can see what their scores mean to the

military; that is, how well they might be able to compete for specific jobs. She also mentioned that students who are not going to college can use the information to see if they may qualify for a military job. A committee member then asked about the process for receiving scores. Dr. Salyer said students first take the test and then they receive an access code, which allows them access to scores and resources. Another committee member mentioned that the test is taken in a controlled setting. Dr. Velgach clarified that students take the interest portion on their own in an uncontrolled environment.

Discussion continued as Ms. Miller asked how the ASVAB-CEP could be brought to schools that do not offer it. Dr. Salyer explained that the best course of action is to contact a school counselor to arrange for the student to take the test at a local school that offers the assessment. Ms. Miller then inquired about how PTIs would work in that situation, especially if it is difficult to return to the alternate school a second time. She asked if it was possible to do virtual PTIs. Dr. Salyer said they could, but they prefer to put students in touch with recruiters to leverage face-to-face interactions with military personnel. Ms. Miller said she appreciated that but noted that some Services are retracting their on-the-ground recruiting footprint. Dr. Salyer explained that other options included pre-recorded video providing PTI. Dr. Velgach then explained that schools can work with MEPS to get students to the alternate school, though she said counselors are not always supportive. She also said parents can work directly with recruiters. Dr. Salyer commented that it was difficult to arrange for PTIs at alternate locations during the day, because the students would require an excused absence. She said she needed to continue to work that issue, but that the "Contact Us" link on the website is an outlet for students in schools that chose Option 8.

As Dr. Salyer mentioned the expert panel's recommendation that the FYI item pool be reviewed for job relevance and cultural appropriateness (slide 14), a committee member asked how the panel came to that conclusion and whether its members took the inventory. Dr. Salyer said the panel did take the inventory and there was one item that she recalled as having stimulated the discussion on cultural appropriateness. Dr. Salyer said she would expect to see item drift between 2005 and 2019, which was the impetus for the effort.

Upon hearing recommendations and progress on marketing (slide 16), a committee member commented that Dr. Salyer needed an intern to help with writing articles about the CEP for marketing purposes. Dr. Salyer said the committee had already made that recommendation. She said Dr. Donna Duellberg (Coast Guard) had told her about an intern program she was familiar with, but that the six-month rotations were too short.

Discussion on recommendations closed with a committee member mentioning the Indiana Workforce Development Program as a possible outlet for the ASVAB-CEP. Dr. Salyer asked the committee member to let her know if she had any contacts in the program. She said Indiana wants to use the ASVAB-CEP as part of a grand pathway, though her program is designed to focus on career exploration.

As Dr. Salyer started to brief on state usage (beginning with slide 20), she recalled telling the committee that Texas had put into legislation the requirement that the ASVAB-CEP had to be provided to everyone in the 10th through 12th grades. She said that was great, but "awful" at the

same time, because MEPCOM does not have the resources to support that large of an effort. Dr. Velgach explained that the concern is that states, like Indiana, are not coming to the CEP program for input, which results in laws that are not sustainable. She said students in Indiana are showing up at MEPS and asking to take the ASVAB so they can graduate high school.

Noting the states that reference the ASVAB CEP or career development in legislation (slide 23), Dr. Salyer said that whether the states require it or not, the legislation still results in circumstances that stress the PTI infrastructure. She mentioned that there is currently only one person in Kentucky qualified to conduct PTIs. A committee member asked Dr. Salyer if she had talked with the Council of Chief State School Officers (CCSSO), and Dr. Salyer said she had and was planning to do a presentation for them next year. The committee member said this would be one way to reach most of the states. Dr. Velgach added that DoD has a state liaison office and that Dr. Salyer had met with them to make sure the CEP program was part of the conversation. She also mentioned that the American School Counselor Association has been contacted, which she said reaches about 80% of counselors nationwide. The committee member replied that it would be useful to communicate with counselors, but it was also important to reach state leaders. Dr. Salyer said states sometimes have someone who has been in the military, but that they tell an old story about the military, which is not so helpful. Ms. Miller said some people are treating the ASVAB-CEP as a pretest, and word is out that it is free. She noted, however, that it is not free to DoD. She said working these avenues to obtain needed resources, while not tempering enthusiasm, is challenging. Dr. Velgach responded that legislating a requirement for the ASVAB is not necessarily the right way to go, but if states are going to do that, she wants Dr. Salyer's program to be part of the conversation.

As Dr. Salyer showed the list of states that mention only the ASVAB in legislation (slide 24), Ms. Miller reported that another challenge results from states wanting information about students who are taking the ASVAB-CEP. She said they want the information broken out by school and grade so they will have metrics for the purpose of holding schools accountable. She emphasized that this is not an appropriate use of the scores. A committee member responded by saying this would have been a good problem to have five years ago. Another committee member clarified that the real problem was the mandate. Dr. Salyer agreed, saying it causes more schools to select Option 8. She said schools dislike being told what to do.

About the ASVAB-CEP and the ESSA (slide 27), Dr. Velgach said AP does not lobby states to include the ASVAB CEP through legislation, however, AP and DPAC answer questions and provide assistance when contacted by states. She also said expanded use by states impacts available resources, but states do not provide any monetary resources in support of the program. She explained that several States are educating their counselors and teachers on the available resources, and access codes are provided to them so they can use the resources repeatedly throughout the school year.

Dr. Velgach commented that the orientation program provided in Indiana (slide 28) could serve as a model for doing the same in other states.

As Dr. Salyer briefed the PTI proficiency training program (slides 30-37), a committee member asked how the PTI codes could be misused. Dr. Salyer replied that recruiters who do not

complete the PTI training and, thus, do not receive a code, are using a counselor's code. She said she can identify these cases, and that she sends them an email to tell them to finish the training.

At the end of the briefing, Ms. Miller commented that the ASVAB-CEP program is viewed by the National Commission on the Future of Military and Public Service as something that can be modified for a more general purpose, which includes rebranding to reduce the military flavor. She said AP believes it important to maintain an obvious association with military service. Ms. Miller continued, saying that AP has had to go to the United Nations to explain that the U.S. military is not trying to recruit child soldiers. She told the committee that, if they had any thoughts on that, to please include them in their report. A committee member responded by saying that it would be misleading not to mention the program's association with the military. Ms. Miller replied that the PTI program can be used for other purposes, but that it is important to be transparent about its primary use. She reaffirmed the two aspects of the program: career exploration and recruiting. She asked the committee if AP was being too cautious about that, and a committee member said, no, it starts with the ASVAB; if it was just career exploration, then it would be okay. Dr. Salyer then clarified that she just wanted the committee to be aware of the situation.

8. Evaluation of FYI (Tab K)

Dr. Salyer, Manager, Career Exploration Center, presented the briefing for Dr. Olga Fridman (DPAC).

Dr. Salyer began by explaining that the FYI was developed in 2005 for the ASVAB CEP. It was designed to measure interests in accordance with Holland's theory of career choice which identifies six categories or personality types that characterize people and work environments (Realistic, Investigative, Artistic, Social, Enterprising, and Conventional, or RIASEC). Shanon Salyer created a database of 505,109 records of FYI responses collected from 2015-2017. After removing duplicates, 321,687 records were retained for this analysis. Each record contains 90 responses, with 15 items representing each of the interest categories. For each question there are three response alternatives: Like, Indifferent, Dislike. After 14 years, an evaluation of the FYI is being conducted due to the tremendous changes that have occurred in technology, the economy, and social structures that could make content of the FYI outdated or even irrelevant. The goal of these analyses was to answer three questions. 1.) How well (if at all) does the FYI match Holland's hexagon structure currently? 2.) Does the FYI apply equally to men and women? 3.) What can be done to improve the quality of the FYI? The statistical analyses used to address these questions were multidimenaional scaling (MDS) and factor analysis.

In February 2019, the CEP Expert Panel released a report titled *Initial Research on the Revision of the Find Your Interests Inventory for ASVAB CEP.* The panel made several recommendations, two of which were the focus of these analyses. The first was to minimize differences related to gender and the second was to use MDS to achieve a more robust RIASEC model in a revised FYI. The goal of MDS, which can be considered an alternative to factor analysis, is to detect meaningful underlying dimensions that can explain observed similarities or differences between the investigated objects. One example of an application of MDS is to reconstruct a map from a table of distances between two points on a map. While MDS can recover the relative positions of the cities, it cannot determine absolute location or orientation (e.g., east from west). Dr. Salyer presented results from the expert panel's analyses, which showed males having a Realistic scale pulled away from the others and a Social scale slightly closer to the Realistic scale than the Artistic or Enterprising scales. Both configurations have a gap between the Conventional and Realistic scales. The panel concluded that the FYI inventory items fit the RIASEC model for males poorly.

MDS tries to find points that have a given set of pairwise distances. When no set of points satisfies distance constraints, MDS finds the best solution in the least squares sense. In this case, the distances between objects are expressed in the correlation matrix. Dr. Salyer showed tables displaying the FYI correlation coefficient matrix and dissimilarity matrix for raw scores for males and females. In the current analyses, the two-dimensional MDS calculations done by the expert panel were replicated. A chart of the results displayed six clusters, as expected, but they failed to form a hexagonal structure. Additional data indicated that the difference between female and male dissimilarity matrices is comparable to the errors of the MDS fit. That is, the difference between females and males is lost in the errors of the 2-D MDS method. The 2-D MDS graphs give desirable visualizations of data in some cases (e.g., the distances between two points on a map), but fail to do so in this case where there is a 5-dimensional structure. Calculating the errors for four different dimensionalities lowers the error. The 3-D MDS produces errors that are 30% lower for female respondents and 60% lower for male respondents. Several charts were presented demonstrating these outcomes.

Dr. Salyer concluded by stating that Holland's hexagon is a symbolic illustration of the mutual association of the six interest categories. The analyses showed that mutual associations are consistent with Holland's diagram. The fact that dissimilarities fail to form a 2-D hexagon is not a violation of Holland's model. 2-D MDS is not a convincing tool for analyzing the quality of the FYI inventory, because it is inconclusive and misleading.

Factor analysis helps in determining how well the measured variables (90 items) represent the number of categories in the RIASEC representation. Factor analysis aims to find independent latent variables. Eigenvalues measure the amount of variation in the total sample accounted for by each factor. If a factor has a low eigenvalue, then it is contributing little to the explanation of variances in the variables and may be ignored as less important than factors with high eigenvalues. Dr. Salyer then showed a table that indicated that the first six largest positive eigenvalues that emerged from analysis of the FYI data accounted for 96% of the common variance, suggesting that six factors are present. The results also suggest that the Enterprising category demonstrates excessively strong cross-loadings with the Conventional domain. The bar-plots for male and female factor loadings are similar, but not identical. As an experiment, the factor analysis was re-run after first removing the Enterprising items and then again after removing the Conventional items. In both cases, the factor loadings became remarkably well distinct. Although the male/female factor structures were very similar, there were somewhat different endorsement rates for Realistic items. This led to the conclusion that separate norms may not be needed for males and females, except perhaps for the Realistic domain. Another approach would be to revisit the items on the Realistic scale to see if the differential endorsement rate can be reduced.

When Dr. Salyer explained that the large number of duplicate scores (shown on slide 4) was partly due to parents using their child's access code, Ms. Miller asked if parents were taking the inventory to learn about themselves. Dr. Salyer said, yes, she believed they were curious. A committee member then commented (on slide 9) that the two-dimensional MDS solutions looked good and conformed generally to the shape of a hexagon. Dr. Salyer said she was concerned about the skewness.

As Dr. Salyer presented OPA's factor analysis (slides 22-25), a committee member observed that the Enterprising and Conventional categories were on the same side of the hexagon and said it would be worse if they were not on the same side of the hexagon. Dr. Salyer agreed and said the Enterprising category demonstrates excessively strong cross-loadings with the Conventional domain. She then conveyed the consultants' opinion that the model was not measuring the fullness of the construct, and that there was a different way to measure interest items.

When Dr. Salyer presented the conclusions (slide 26), the committee made an argument in favor of using the MDS analysis for visualizing the model but said factor analyses were sufficient. One

committee member emphasized the importance of being able to see the mean differences between males and females. Another committee member concurred and explained that minimizing male-female (M-F) differences might obscure any relevant real differences, even if they could be attributed to socialization. Another committee member mentioned that Holland and Strong had performed the same analysis and achieved similar results, which indicates that perfect hexagons are hard to produce. A committee member pointed to the spatial component, the geometric representation, that provides information about where the categories sit in multidimensional space; that is, the proximity of the categories. S/he then said the correlations between factors in the factor analysis provide some information about proximity and asked if Dr. Salyer had the correlations. Dr. Salyer said those correlations did not make the slides. She added that the factor analysis was conducted with the same data but said she did not know if the data had been parsed the same way, which meant the mean difference correlations might not be the same. A committee member observed that both analyses told essentially the same story, and then asked how the FYI results were used; that is, if they were normed against particular occupations, which would indicate that the meaningful information would be used to key people to the occupations they are interested in. S/he said it would be nice, from a research perspective, to see the model verified, but that model verification was not as valuable as obtaining accurate outcomes. S/he then asked if the FYI was still recommending students use their top two interest codes to search for occupations. Dr. Salyer confirmed that was correct.

Continuing the discussion, a committee member asked Dr. Salyer if she was worried about gender differences, and if she really thought men and women should get the same results. Dr. Salyer said she was concerned, rather, about washing gender differences away, when research and the analyses showed that gender differences were still present in the data, though it would make her job easier. She then talked about how some people say they are gender fluid, but that the FYI requires a gender, whereas industry is moving away from that practice. She then asked for the committee's recommendations about how to proceed.

Dr. Pommerich recalled slide 25, which she said showed different endorsement rates on Realistic for males and females. Dr. Segall said the existing items had already been screened to eliminate gender differences, and Dr. Pommerich replied that the only area she would worry about was Realistic. A committee member said s/he may not agree that the differences present a concern. Another committee member asked how much difference entering male versus female makes. Dr. Salyer said it might make a difference on Realistic, and because people can only search on two categories at a time, they might be rerouted. She added that a person could be provided one combination of categories as a male and another combination as a female.

Dr. Pommerich then raised the requirement that the inventory should not be discriminatory. Dr. Velgach replied, however, that the inventory does not provide *a standard* for joining particular career fields, but that it simply *indicates interests*. She said that maybe the important piece is how the inventory is presented. Dr. Salyer said the inventory already provides clear guidance on the interpretation of gender differences. This prompted a committee member to ask if the inventory asked for sex, or identity, or was it ambiguous? Dr. Salyer said people could choose based on how they felt that day, for example, what would you like to explore as today, male or female? She said people could also see their results based on combined norms. Dr. Velgach said that wording recommendations came from AP's legal department.

Dr. Salyer again asked the committee what they recommended regarding the Realistic category. One committee member said s/he had not seen evidence that the items were deficient, clarifying that there may be a true difference between how males and females respond. Dr. Salyer rephrased, asking if they should report gender differences if there is a real difference. The committee member explained that the longstanding challenge in this line of work is that men and women are represented differently in different jobs, so there might really be bias in suggesting that a female pursue jobs in which there are not many women. S/he said the main purpose of an interest inventory is to facilitate exploration of pathways that might not otherwise be explored.

A committee member asked if the profiles say something like, "compared to men, you are high, but compared to women, you are low." Dr. Salyer said the FYI does not put it exactly like that. Another committee member suggested that people receive scores that are compared to norms for males and females. S/he said the inference would be, for example, that because my score is high compared to males or females, then I should explore either this or that. Dr. Segall clarified that there are three sets of norms: male, female, and combined. He said that an individual only sees results in two of these: the gender they selected and combined. A committee member then suggested that, because the inventory is to be used as a guide rather than a prescription, there should not be an issue. At that point, Dr. Salyer asked for confirmation that the committee was recommending sticking with displaying the combined and gender norms, and the committee said yes. A short exchange then occurred as an audience member asked if the inventory could provide output on all three designations (i.e., male, female, and combined). A committee member clarified that showing all three would not help people make choices. Dr. Salyer said it is difficult enough to understand two, which is why there are people available to help students interpret their results.

The discussion closed with a question from Dr. Fechter about the importance of the context included in the items. That is, she asked if the context might be causing items to be more appealing to a male or female. She used the example of auditing a construction site versus auditing a restaurant. Dr. Salyer said it might, suggesting that the context of some items should be revisited. A committee member then asked if Dr. Salyer was proposing a difference analysis. Dr. Salyer said, yes, but asked how many items they would need for that, to which the committee member responded, "a lot." Dr. Mark Rose (AFRS) then commented that there may be something that separates the context from the activity, so looking at the activity free of context may be a purer way of estimating interest. A committee member, however, argued against that approach and explained that removing the context would eliminate much of the concreteness, which s/he said is what makes the items work. S/he said that removing the context would make the items more challenging. Dr. Salyer concluded the discussion by noting the existence of geographic differences. She described a case in New York, where students asked, "what is a riding lawnmower?" She then said, if these were easy items to write, they would have written more.

Dr. Velgach closed the day's sessions by asking for public comments, of which there were none, and then adjourned the meeting for the day.

9. ASVAB Validity Framework (Tab L)

Dr. Art Thacker, HumRRO, presented the briefing.

Dr. Thacker began the briefing by explaining that the validity of an assessment cannot be summarized through a single statistic or coefficient. Validation depends on the assessment's purpose, the inferences made from assessment results, and the uses of those results. Argument-based validation tests the underlying claims that must be true to support the inferences made from assessment information (scores). An assessment score may be valid for multiple purposes, in which case it is rare that the assessment is equally valid for all of them. Evidence is collected for a validity argument to support claims in a chain of reasoning, where any claim in the chain found to be weak may undermine subsequent claims. For instance, poor item quality can undermine all results from an assessment and poor year-to-year equating can undermine cross-year comparisons of scores. If multiple inferences are drawn from a single assessment score (or event), each inference may have its own unique validity argument.

Regarding the ASVAB, the most important inferences are admission in to the military (AFQT) and placement into training programs or advanced educational opportunities (ASVAB). The ASVAB primarily relies on an informal reasoned approach, and evidence is not currently tied to organized claims or assumptions. A Theory of Action (TOA), or something similar, is required to frame interpretive and validity arguments for the ASVAB. The current work does not address admittance to specific training programs or occupational specialties, as each would require its own body of evidence, which is beyond the present scope.

A TOA addresses all the things the test and test scores are expected to be used for and the expected advantages of using the test for those purposes. An Interpretive Argument is a description of the inferences that the test scores support, while a Validity Argument summarizes the evidence providing justification for the inferences in the interpretive argument. The draft AFQT TOA states that the AFQT measures *g*, and because *g* is predictive of a broad range of future performance, the AFQT will broadly predict candidates' success in military occupations. With this formulation as the basis, claims can be developed that must be supported for each step in the TOA to be true.

Dr. Thacker then identified the claims associated with the TOA for the AFQT.

I. **AFQT Subtests Measures *G***
   1. *g* is a broad stable construct underlying cognitive test scores.
   2. Each of the four AFQT scores is a reliable measure of its intended construct.
   3. The four AFQT subtests are the best options for a *g* proxy.

II. **Derivation of the AFQT Category Scores Supports Their Use for Recruit Selection**
   4. AFQT scores measure *g* in the intended population.
   5. The predictive nature of *g* is continuous for nearly the full scale of the AFQT (i.e. higher scores always yield a better predicted outcome, irrespective of the area of the scale the score falls in).
   6. The AFQT categories represent important differentiators among applicants.
   7. AFQT scores have high overall reliability and lower error, especially near the cut points for the categories.
   8. AFQT scores have high classification accuracy.
   9. AFQT scores are unbiased regarding race/ethnicity, gender, etc.

III. ***g* and the AFQT Predict Important Training and Job Performance Criteria**
   10. AFQT is a measure of *g*, so AFQT scores should demonstrate a pattern of correlations with different types of job and training performance criteria similar to *g*'s predictive pattern.
   11. AFQT category scores are associated with important outcomes.

IV. ***g* and the AFQT Scores Yield Similar Patterns for Subgroup Differences**
   12. *g* and the AFQT scores yield similar patterns for subgroup differences.

V. **Contextual Factors Support the Use of the AFQT**
   13. Users of the AFQT scores and/or the AFQT category scores understand the meaning and use/outcome of the scores.

14. The ASVAB is administered appropriately.

VI. **Candidates Scores are Interchangeable Irrespective of the Version of the AFQT they Take**

15. The P&P and CAT versions of the AFQT yield interchangeable scores.
16. Unproctored verified and proctored versions of the AFQT yield interchangeable scores.
17. AFQT delivery on other devices (e.g. tablets, cell phones) yield interchangeable scores compared to proctored CAT versions. Currently being evaluated.

Dr. Thacker then presented an excerpt from the draft AFQT validity argument that listed the assumption, the claim resulting from that assumption, and the evidence in support of that claim. He explained that each claim still includes a link to more specific documentation about that claim. All claim links are organized using a common structure. The links include (a) restatement of the claim, (b) evidence categories or the main headings for organizing evidence, (c) a summary for the validity argument, (d) a literature review, and (e) a reference list. Dr. Thacker then showed an excerpt from a claim link. Next steps include (a) revising the TOAs to better reflect the logic model underling AFQT and ASVAB, which is an iterative process; (b) defining and revising the assumptions associated with the revised TOA; (c) developing or revising specific claims that support the assumptions; (d) identifying the required evidence necessary to support the validity claims; (e) referencing the evidence for specific validity claims from the literature and ASVAB documentation; (f) identifying evidence gaps or weaknesses and conducting studies to address them; and (g) maintaining and updating the validity argument as necessary.

Dr. Thacker then presented a draft ASVAB technical test TOA. The model begins with the assumptions that the ASVAB subtests measure many knowledge, skills, and abilities (KSAs), and that job analysis can identify KSAs required for job or training success. Experts then match job/training KSAs to ASVAB subtests such that ASVAB subtests predict job/training performance, and recruits selected based on subtest scores succeed in training and on the job. The model relies on clear linkages between KSAs required for military training/jobs and KSAs measured by the ASVAB. A second draft ASVAB technical test TOA begins with the proposition that AFQT and ASVAB measure *g*. In addition, ASVAB technical tests may reflect course-taking patterns or life experiences of candidates (e.g., took shop class, worked in automotive repair). ASVAB technical tests may act as an interest inventory, with candidates who spend time on an interest related to a subtest scoring higher. Preparation may improve subtest scores and could be associated with conscientiousness or motivation.

Dr. Thacker then mentioned several unvalidated potential uses of the ASVAB, including use as (a) an indicator of student readiness for careers, (b) school-level accountability measures, (c) alternate evidence of high school preparation for students who do not pass the state's graduation assessment, and (d) alternative language versions of the ASVAB as a better *g* measure for non-native English speakers. These uses must be validated before inferences can be made.

Dr. Thacker presented several ongoing challenges in the validity argument work. These include the fact that the ASVAB was not created using Evidence Centered Design or a similar approach based on claims and references. The ASVAB also has a 50-year history, multiple users, varied score information, and multiple inferences that need to be supported. There is also a lack of models from comparable assessment systems. Finally, discerning which ASVAB literature is relevant for the validity argument is not always straightforward, and the literature itself is not always unbiased.

Dr. Thacker concluded by providing a preview of some recommendations that could result from this work. These include establishing an AFQT/ASVAB technical manual that starts from the "ASVAB Standards" from the field test checklist to create categories for the technical manual. This should indicate how often each type of evidence in the technical manual should be updated, with updates occurring only when a substantive change is made in the assessment. Additional recommendations would be to estimate the classification accuracy at the selection decision cut score and investigate comparability of devices as new technology becomes available (e.g. phones, tablets).

Early in the briefing, Dr. Thacker referred to the importance of Kane's (1990) ideas on validity arguments. A committee member noted that Cronbach (1988) had previously addressed the subject. Dr. Thacker then mentioned Cronbach's work with Meehl from the 1950's. Dr. Thacker subsequently commented, on the draft TOA (slide 7), that the first two elements in the graphic would be reversed, because the designers of the ASVAB obviously realized that *g* was important. The committee agreed the modification was appropriate.

As Dr. Thacker described the last of six AFQT assumptions (slides 8-10), a committee member asked him to repeat what he had said about the fourth assumption (i.e., that *g* and the AFQT scores yield similar patterns for subgroup differences). Dr. Thacker replied that subgroups sometimes score differently on measures of *g*. He said he would like to see the ASVAB show similar differences. A committee member remarked on the large number of impact studies and asked if Dr. Thacker meant that the AFQT should show differences similar to those shown in the National Assessment of Educational Progress (NAEP). Dr. Thacker replied that he would worry more if the ASVAB did *not* show the differences shown by other reputable tests. The committee member then said the assumptions regarding classification accuracy and lack of bias on slide 9 were questionable. Dr. Thacker explained that he had evidence of similar differential impact between the AFQT and other tests. Another committee member suggested that the point may not be necessary at all, raising the matter of mean differences versus predictive differences. The first committee member asked Dr. Thacker if he had a plan for providing evidence. Dr. Thacker said he did, and that it would show the tests are similar. The committee member said that was not evidence that the AFQT was unbiased and might indicate that all the included tests shared the same biases, and Dr. Thacker agreed. Another committee member reported that test developers get a lot of grief from political factions among stakeholders for having subgroup differences; s/he said they think the differences invariably show bias in the test. S/he went on to point out that producing similar results would likely lead to similar criticism. Dr. Thacker said in reply, "right, we are admitting to bias then." The committee member said that it really turns on whether the differences reflect some real differentiation on the constructs or whether they are due to some other irrelevant factor.

Upon hearing more about the evidence summaries and links to additional documentation, (slide 13), several committee members said they liked how the effort was coming together. Dr. Thacker said he was trying to focus on literature most relevant to the ASVAB and AFQT. A committee member said that was important, because the key to validity is the theoretical basis—inference and support. The committee member said Dr. Thacker appeared to be setting a high bar. Dr. Thacker remarked on the high quality of the Test of English as a Foreign Language (TOEFL) work. The committee member said the quality of the present work is close, if not a step ahead of the TOEFL work.

When Dr. Thacker said, regarding the job analysis model and why the ASVAB predicts success (slides 18-19), that the TOA relies on a regular refreshing of job analyses as new jobs come onboard, a committee member observed that the effort was looking only at technical tests and noted that components of *g* were present in every test. Dr. Thacker replied that he did not have a good answer, currently, for dealing with that. He said the Services use other parts of the ASVAB to qualify people for different jobs, and *g* is important, but that the current approach is the best he had in relation to determining how the technical tests apply. The committee member then stated

that the knowledge, skills, abilities, and other characteristics (KSAOs) captured by non-AFQT tests were matched to training and asked whether some of that domain was not also being captured by the AFQT. Another committee member agreed, while another asked if one of the KSAOs might be general cognitive ability. The first committee member suggested that it might be possible to validate the entire ASVAB, instead of just the AFQT. Another committee member said people are sometimes more likely to succeed based on other characteristics. Dr. Thacker said he had not yet addressed the "are they more likely to succeed" aspect of the argument but said it may require more than just determining that people who are successful have the KSAs. He said, ultimately, there are many reasons people may succeed at a job or training, and they rely on *g*, motivation, prior experience and expertise, and other factors. He said, for this effort, the AFQT is characterized as predictive because it is a good proxy for *g*. To get to more nuanced predictions for specific jobs or job categories, the Services use subtests from ASVAB (including those from the AFQT) in various combinations and weightings. He said there may be an interest inventory application as well; that is, interest may cause better test scores. A committee member asked why Dr. Thacker cared about that, and if he needed to make a claim about interest from ASVAB scores. Dr. Segall explained, however, that looking at technical tests that were not the main focus would be thinking ahead. He said that developing a TOA for a technical test would run into candidate factors, and that they may find the primary factor to be *g*. He added, however, that the technical tests add incremental validity, and that may be because they are acting as measures of interest. He said it could be important to explain why those tests added predictive validity. Dr. Thacker said that if he could show a link between course patterns and ASVAB scores, or if he had evidence that taking a course as a junior or senior in high school might help a person qualify, it might be useful. A committee member warned, however, that *g* might affect the impact of taking a course, and people should not be encouraged to take a course that is too hard for them. Dr. Thacker agreed, saying that it is not a confirmative causal inference that if one takes a course, s/he will get a good score on the ASVAB.

When Dr. Thacker presented the list of unvalidated potential uses of the ASVAB (slide 20), a committee member suggested that maybe the ASVAB could predict success for nonmilitary careers. Dr. Thacker said that he did not want to go that far. He then suggested that the CEP has some use in that area, perhaps in predicting "option ready." The committee member asked if Dr. Thacker was looking at that, to which Dr. Thacker replied, no, but that he could keep the door open. Another committee member asked if another TOA belonged here. Dr. Thacker said, yes, a separate TOA would be required, but that this just addresses inferences for which there is not support for a TOA. Dr. Salyer said some states are doing research on using the CEP to predict student readiness, but Dr. Thacker said the evidence he had seen was weak. He said he had looked at concordance-type evidence, which did not come close to saying people would be career ready based on ASVAB scores, even if there were correlations with some other work readiness indicators. He said some of those indicators were weak.

As Dr. Thacker talked through the rest of the unvalidated potential uses, he said that some states want to use the ASVAB as an accountability measure. He said they think they can use this "on the cheap," but that he is trying to outline why that is probably a bad idea and that they should do more research first. He also said some states want to use the ASVAB as part of the path to graduation, but that there is no evidence indicating that was a reasonable course of action. He said, however, that he still wanted to address it. A committee member then asked if Dr.

Thacker's intent was to say there is no evidence for these uses and, therefore, either avoid it or collect data first? Dr. Thacker said that was his intent. The committee member then asked to whom that information would be directed. Dr. Thacker said the states, or whoever the user happens to be. The committee member said it is usually the test's owner. Another committee member explained that test owners should be responsible for making interpretations that can be supported. Dr. Velgach replied, however, that there were different cases, with one being the push to allow the use of calculators on the current test. She said AP was receiving pressure to modify the battery for calculator use and to make inferences about test scores without analyses to support them. She explained that AP was trying to warn that this is not appropriate. The committee member said there will always be pressure to use tests for alternate purposes, and that, as the owner, one can say s/he needs to collect data on it or do not do it. Dr. Segall replied that DPAC communicates that to the states, but they do it anyway. Ms. Miller concurred. Dr. Segall said that the current work would help DPAC because they conduct careful analyses that should carry some weight. The committee member said s/he thought the approach was great. Ms. Miller said it was a good mechanism for explaining why people should not use the test for certain purposes and for making an argument as to why the introduction of calculators would not be appropriate with the test as currently designed. Dr. Velgach told Dr. Thacker that it would be helpful if he would incorporate his perceptions on whether certain targeted uses are appropriate, and if not, why.

Regarding ongoing challenges (slide 21), a committee member asked if the claims and inferences were not implicit. Dr. Thacker said they were, but that, as the ASVAB has evolved, he must work backwards to figure them out. Another committee member said to take care in how that is documented; that is, avoid saying that the test did not have initial claims and inferences, because Evidence Centered Design was not around 50 years ago. Dr. Thacker agreed that he could say it was based on logical thought, but not specific documentation. The committee member then asked Dr. Pommerich what was the earliest technical documentation available for the ASVAB. Dr. Pommerich said DPAC has technical bulletins that address form development, but that those focus on DPAC's contribution to the ASVAB Testing Program. She said there is also a whole domain of literature that spans the lifetime of the ASVAB, but that it has not been consolidated like one would see with the American College Testing (ACT) or Scholastic Assessment Test (SAT). Dr. Thacker said he would recommend the production of a consolidated technical report on the ASVAB. Dr. Segall remarked that the ASVAB was created to consolidate the tests used by each Service, and that the content goes back to WWII.

Following Dr. Thacker's presentation, the committee said that he had done great work.

10. <u>Criterion Measurement</u> (Tab M)

Dr. Laura Ford, HumRRO, presented the briefing.

> Dr. Ford began by stating that the objective of this project is to document the test evaluation criteria unique to each Service's accession and classification testing efforts, and to identify and/or develop a unified set of criterion instruments which can be used by all Services. Standardizing measures across Services is desirable to facilitate (a) interpretation of validation studies and (b) generalizability of results. To facilitate the work, a Criterion Measures Advisory Panel (CMAP) was formed, made up of representatives from each of the Services and the Office of the Under Secretary for Personnel and Readiness. The first step taken was

to develop a taxonomy to define the job domain of first-term enlisted Service members. The three criterion domains were:

- Job performance: individual behaviors that are relevant to the Services' goals and that can be scaled in terms of each individual's proficiency. The four factors at the highest level were technical proficiency, organizational citizenship and peer leadership, psychosocial well-being, and physical performance. There are 12 mid-level performance dimensions and 33 sub-dimensions.
- Attitudes: cognitions that are relevant to individuals' job plans and performance (e.g., commitment, satisfaction, career intentions).
- Organizational outcomes: outcomes that are important to the Services at an organizational level, such as reducing attrition and enhancing reenlistment.

The job performance taxonomy was developed by (a) conducting a literature review, (b) developing an initial taxonomy, (c) conducting a retranslation exercise, (d) finalizing the taxonomy and definitions, and (e) obtaining subject matter expert (SME) evaluations. The goal was for the taxonomy to be comprehensive, efficient, hierarchical, and relevant. The 8-dimension Campbell model served as a scaffold, and content from other models was incorporated. The process resulted in 33 draft dimensions. A retranslation exercise was conducted to evaluate the clarity of the 33 performance dimensions and determine the appropriate hierarchical structure. The procedure involved having 17 researchers with extensive experience in performance measurement and/or military criterion development evaluate a 2-dimension (can-do, will-do), 4-dimension (technical, counterproductive work behavior, citizenship and peer leadership, physical), and 10-dimension structure. They either rated (2-dimension) or sorted (4-, 10-dimension) the 33 performance dimensions. Reliability and agreement were high. The results suggested that the 2-dimension solution did not work well, but the 4- and 10-dimensions did. The ratings were then used to make refinements. Some dimensions were moved to other categories, definitions were refined, and dimensions that did not sort well into any of the 10 dimensions were broken into their own. The final model has three levels; 4 categories, 11 dimensions, and 33 sub-dimensions. Dr. Ford then showed a graphic summarizing the final taxonomy.

The purpose of the SME review was to evaluate the dimensions regarding their generalizability (extent to which constructs measured are important DoD-wide) and relevance (extent to which the constructs measures are relevant to the organizations' occupations and goals). Military experts with broad knowledge of Service job requirements were asked to rate each sub-dimension on its importance across enlisted first-term occupations and criticality to accomplishing the Service's mission. Sufficient data were received from the Army, Navy, Marine Corps, and Air Force to proceed with analysis. The interrater reliabilities within Service ranged from .73 to .98. Dr. Ford then displayed an example of the data collection instrument. This was followed by a chart displaying the average importance/criticality rating for the technical, organizational citizenship and peer leadership, psychosocial well-being, and physical dimensions by Service. She then summarized the results of the SME evaluation.

- With one exception, broad performance factors were relevant and generalizable across Services.
- Eight of 33 performance sub-dimensions had an average importance/criticality rating less than 3.0.
- Psychosocial Well-Being was rated highest across Services (i.e., most generalizable).
- Physical Performance sub-dimensions were rated least highly, also were the most variable across Services.
- Organizational Citizenship/Peer Leadership and Technical Performance were rated comparably to one another overall, but there was substantial variation across elements (e.g., job-specific proficiency and safety were rated highly, nonverbal and written communication were rated lower in the Technical Performance category).

The overall conclusion was that the 4-dimension and 11-dimension levels are relevant and generalizable across Services.

The next step was to develop a criterion database containing descriptive information needed for available criterion instruments. This involved conducting a criterion review, in which parameters were set for collection of existing criterion measures, "operational" and "exploratory" criteria were identified, and the CMAP met to finalize the list. To facilitate this process, an online data entry tool was created and populated with over 230 criteria. The criterion instruments were then mapped against the taxonomy and gaps or problem areas were identified. The parameters were that the measures should be (a) applicable across

Services, (b) developed since 1980, (c) focused on first-term enlisted outcomes, and (d) constructed with an operational emphasis. Dr. Ford then showed an example of the on-line survey tool used to collect the information and a screen shot of the database itself. Through this process, 74 current criterion measures were identified, including 13 job performance rating scales, 13 performance tests (including knowledge, physical, and situational judgment tests [SJTs]), 20 attitudinal surveys, and 28 variables from administrative data. The measures were then mapped to the three taxonomic domains, and those with few or no instruments were identified. In formulating recommendations for criterion instrument use, the goals were that they be:

- Relevant and generalizable;
- Relatively easy to develop, administer, score, manage, and maintain;
- Psychometrically sound;
- Flexible enough to enhance Service-specific use;
- Future oriented; and
- Utilitarian (i.e., making the most of the Services' current practices and procedures).

Recommendations were developed in four steps.

1. The generalizability and relevance of the criterion constructs were assessed through a CMAP-supported survey of internal staff with extensive military testing experience who rated the importance and criticality of the 33 job performance sub-dimensions.
2. The psychometric quality and the feasibility of the criterion measurement methods were assessed.
3. Criterion composites were computed and applied to the criterion instruments in the database.
4. Draft recommendations were assembled and reviewed and discussed with contract monitors, a panel of internal military measurement experts, and the CMAP.

The recommendations centered on administrative data, attitude measurement tools, and performance measurement tools, and included:

- The focus should be on available administrative data with outcome criteria aligned across Services.
- Cross-Service attitude assessments should be developed and aligned with existing personnel assessments and surveys.
- Standardized job performance rating scales, job knowledge assessments, and SJTs should be developed.

Dr. Ford concluded by showing a series of slides depicting the steps necessary to (a) align outcome criteria across Services, (b) develop cross-Service self-assessments, (c) develop performance rating scales, (d) develop DoD-wide SJTs, and (e) prepare a research plan to pilot test/validate the measures.

As Dr. Ford described the job performance taxonomy (slides 7 and 34-37), a committee member asked if its purpose was to capture only the essential elements of performance. Dr. Ford said that was true, but that performance had to be represented at a relatively high level. Another committee member referred to slide 34, noting the breadth of the performance domain. On slide 10, a committee member remarked on bi-directional arrows connecting the predictors and criterion constructs. Dr. Ford said she had replaced the uni-directional arrows because the criteria were not driven specifically by the predictors[2].

As Dr. Ford explained the SME ratings of construct generalizability and relevance (slides 11-14), a committee member asked how many SMEs performed the rating task. Dr. Ford said the Army provided nine SMEs, the Navy three, the Air Force four, and the Marine Corps ten. A committee member then remarked that the means were not particularly stable, based on those numbers, and asked Dr. Ford how certain she was that there were no differences across Services. Dr. Ford said her team had relied on inter-Service reliability estimates, which she said were high. She also

[2] The committee had recommended against identifying criteria based on specific predictors at its last meeting.

reiterated the purpose of the ratings, which she said was only to determine what would be included in the taxonomy, as opposed to identifying constructs by Service. She emphasized that, of the twelve dimensions, only one did not make the cut, and that was physical endurance. Another committee member asked about the comparability of SMEs among Services. Dr. Ford said all the SMEs were familiar with first-term performance and had years of experience in assessment within their respective Service. Dr. Cristina Kirkendall (ARI) said the Army used researchers familiar with first-term performance at ARI as well as senior non-commissioned officers (NCOs) from Army schoolhouses. Dr. Steve Watson (Navy Selection and Classification Policy) said the Navy used three retired E9s with between 20 and 30 years of experience. He said these NCOs were the best they could find.

As Dr. Ford described the search for existing criterion measures (slide 15), a committee member asked if the measures examined had been previously implemented. Dr. Ford explained that many had, but said their intent was to identify as many measures as possible, while excluding job-specific as well as officer and NCO measures, with a few exceptions. She also said they limited their search to measures that had been used from 1980 to present. Finally, Dr. Ford explained that they had expanded the search to measures beyond those with an operational emphasis, especially research-based measures documented in publications such as The Journal of Military Psychology and at the International Military Testing Association (IMTA). Another committee member asked if Dr. Ford's team had looked at rating scales, and Dr. Ford said they had. Finally, a committee member asked if any of the measures had been used across Services, to which Dr. Ford replied that some surveys had been used in that manner.

Regarding the mapping of measures against the taxonomy (slide 19), a committee member asked if any of the rating scales dealt with psychosocial wellbeing. Dr. Ford replied that very few measures dealt with that construct. Dr. Velgach asked specifically about the frequency of measures addressing counterproductive work behaviors, and Dr. Ford said they did not find many. Another committee member asked Dr. Ford if her team accounted for the quality of the measures. Dr. Ford said that the evaluation took that into account, as shown on the next slide (20). A third committee member asked if the team had done a gap analysis by Service, and commented that, if all the Services eventually buy into the concept, some Services would have more work to do than others. Dr. Ford replied that her team was developing measures to fill those gaps.

As Dr. Ford described the research plan (slide 27), a committee member asked if measure validation would require Internal Review Board (IRB) approval. Dr. Ford said IRB approval would be required whenever respondents were asked to provide information about themselves. She said validation of the criterion measures against predictors would require a method to link responses on criterion measures to performance on predictors, which would require some level of personally identifiable information, such as name, date-of-birth, and/or Social Security Number.

11. <u>Navy Validation Business Model</u> (Tab N)

Dr. Stephen Watson, Director, Navy Selection and Classification, presented the briefing.

> Dr. Watson began by providing an overview of the Navy's accession process, which includes ability and moral/financial/educational screens; taking the ASVAB; conducting medical, mental, and moral checks; providing the applicant job and career information; providing a school guarantee and ship date; swearing

in; time in the delayed entry program; and shipping to the recruit training center. Another slide pictured the process by which expected work is defined, curricula are derived from job information, and rating entry standards are developed. He went on to explain that the AFQT, made up of four ASVAB subtests that measure general intelligence, is used in selection. Navy classification composites are measures of specific intelligence and are used in classification. Dr. Watson then displayed a chart listing the ASVAB subtests and special tests (i.e., Coding Speed, the DLAB, the Navy Advanced Placement Test, MCt, and the Cyber Test) used/under development by the Navy. An additional slide presented the various composites formed from the ASVAB and special tests for classification. The ASVAB institutional controls include the Office of the Under Secretary of Defense for Personnel and Readiness Accession Policy Directorate, DPAC, MEPCOM, and the Navy Education and Training Command (NETC).

Dr. Watson continued by showing a schematic depicting the high-level business process for developing and validating ASVAB selection and rating entry standards. He indicated that validation studies are carried out annually for every rating. The priority for a rating may change based on observed predictive validity changes, ad hoc requests from stakeholders, observations from the rating priority index, or the introduction of new tests. Validation studies may be automated based on historical training data or may be more in-depth such that they include information from schoolhouses. The automated procedure involves processing accession data using training outcomes. Correlations between test scores and training success are run for both operational and alternative ASVAB composites, with correction for range restriction and calculations to identify adverse impact. Candidate alternative composites are identified. The evaluation metrics include qualification rate and adverse impact using historical data, training success using regression-based models, and cross-validation with the latest available data. The test candidate alternative composites or cut scores are evaluated through a full-year, whole-Navy assignment simulation, and the results are compared against operational standards.

A predictive validity tool calculates correlation estimates and population-level group-score differences of ASVAB composites or special tests against training success metrics for each rating. It is used to select the best composites for minimizing academic setbacks and failure rates and reducing adverse impact based on gender and race/ethnicity. Spreadsheets are used to calculate the projected impact of alternative cut scores on qualification rates and setback/attrition rates. In both cases the projections are made based on the total population and by gender and race/ethnicity. Dr. Watson concluded by presenting an example of the application of these tools which resulted in the recommendation of using an alternative standard that provided higher qualification rates and improved adverse impact ratios without increasing training setback rates.

As Dr. Watson briefed the business process flow chart (slide 11), a committee member asked him what criteria were used currently in validation studies. Dr. Watson said first-pass yield, or first-pass pipeline success. He said the binary nature of these measures presented challenges, and that he would like to add school grades and number of setbacks, but that those data would not be as available, pristine, or accurate, as first-pass yield data. Another committee member asked about the range of sample sizes across ratings. Dr. Watson said, for nuclear jobs, thousands, but for crypto jobs, it is only two to three hundred. Tom Blanco (S&T Consulting) clarified that the sample sizes ranged between 200 and 15,000. The committee member then asked what percent defines success, and Mr. Blanco replied that the success rate is now close to 90% overall in the Navy, but that it varies by occupation. Dr. Watson said that the rate is lower in nuclear power and SEAL jobs, and that the rate for culinary specialists is close to 100%. The committee member replied that it is difficult to predict a dichotomous criterion if there is only a small proportion of failures. Mr. Blanco replied that, setback/failure rates are very high, up to 30%, for explosive ordinance device and air traffic control jobs. Dr. Watson summarized by saying he thinks the system works well, but there is room for additional enhancements.

As Dr. Watson explained the uses of the projected impact spreadsheets (slides 17-18), a committee member asked if spreadsheet generation was automated, and Dr. Watson said it was. As he talked through the example uses (slides 19-23), a committee member asked if the validities shown on slide 20 had been corrected for range restriction. Mr. Blanco said they had, because they were for the accessed population rather than the applicant population. He said they were trying to get access to the applicant population. Upon hearing that the correlations were biserial, a committee member said they were "really good." Mr. Blanco noted that the current operational composite was AR + Verbal Expression (VE), but that the data indicated the MK + VE composite would perform better. He said the latter composite minimized setback and failure rates, as well as reduced adverse impact. He referred to the graph on slide 21 as something that could be shown to stakeholders. A committee member asked if composites could be changed every year, and Mr. Blanco said that they could, if warranted.

When a committee member asked about the criteria for making composite changes, Dr. Watson explained that they present findings to stakeholders who make policy decisions, and he described the process as one that helps leadership understand the data and provide input. He said if they saw a 10% difference, for example, they would red-flag it and bring it to leadership's attention. This number, he said, had previously been set at 3% and 5%, but that it had changed based on input from leadership. Dr. Velgach explained that the data inform the risk associated with lowering the standard for ratings that have difficulty contracting sufficient population. The committee member asked what decision rule would be followed if an increased qualification rate was accompanied by decreased setbacks but increased adverse impact. Dr. Watson said they would try not to make the situation worse. Mr. Blanco added that there is a lot of pressure to open up qualification rates, but to do so without affecting training success. Dr. Watson said there is usually room for compromise, and that they talk to the training components about the impact of increased setback rates. Another committee member asked if the results shown had been cross-validated, and Dr. Watson said that they had.

At the end of the briefing, Ms. Miller said AP owed a great deal of thanks to the Navy for their work, which she said was part of the Force of the Future (FOTF) program. A committee member asked if the results had been published. Dr. Watson said he had spoken with Dr. Velgach about publishing. He explained that the Navy previously had a research lab called, Navy Personnel Research, Studies and Technology (NPRST), that produced technical notes, but that it had been defunded. He said they are currently looking for a practical way to share their processes and results. The committee member also asked if the Navy's work regularly underwent IRB approval. Dr. Watson said it did not, and Mr. Blanco explained that they had received an exemption. The committee member then asked if informed consent was obtained from participants. Dr. Watson replied that, legally, consent is not required for the use of administrative data. When the committee member said that it is in other fields, Dr. Watson said he would take another look, but that their interpretation was that use of personal administrative data was a condition of employment. Dr. Velgach confirmed that they would review their process to make sure they were conformant. Dr. Segall remarked that DPAC had exemption determination officials. The committee members then discussed the internal procedures at their institutions.

12. Standards Settings for ASVAB Technical Tests (Tab O)

Dr. Tia Fechter, DPAC, presented the briefing.

Dr. Fechter began by explaining that in 2015 DPAC generated a 23-task list called *Plans for Evaluating Current ASVAB Tests*, to guide a holistic evaluation of the ASVAB with respect to its relevancy for this and future generations. One task involves evaluating the usefulness/appropriateness of existing tests with the current population. This will be accomplished by tracking test scores over years (1984-2014), and then evaluating what fraction of the population possesses the knowledge and skills assessed by the test over time. The context for this is criterion-referenced, meaning a measure of performance against a fixed set of predetermined learning standards. ASVAB subtests do not have cut scores on a test score scale that establish a demarcation that categorizes examinees into those with significant exposure to the content that the subtest measures and those who do not. Implementing a standard setting would establish these cuts and allow for calculation of the proportion of the population that could be considered "significantly exposed" to test content. Cut scores will help to determine the proportion of applicants significantly exposed to the content of interest by subtest and by year (1984-2014). This will help in identifying trends in the proportion of applicants with significant exposure over time and thereby evaluate one potential criteria for the continued usefulness of subtests for making classification decisions with the current population of youth.

Dr. Fechter continued by identifying the subtests of interest as the ASVAB science and technical tests: Automotive Information (AI), Shop Information (SI), Electronics Information (EI), Mechanical Comprehension (MC), and General Science (GS), as well as the special test, Cyber Test. These tests are used to develop composite scores for classifying military personnel into occupations. Performance on ASVAB subtests are weighted, as appropriate, to match the skills and abilities required for successful performance in training schools and on the job. Each of the Services is responsible for validating their own composite scores. The ASVAB has included AI, EI, SI, and GS since 1968. In the intervening years, military occupations have expanded to include fields such as cyber security. DPAC would like to determine whether AI, EI, MC, and SI are still dominant technical areas to which high school students are exposed, as well as whether these technical areas represent the bulk of current-day vocational interest and needs. In addition, DPAC would like to determine if areas such as computer science may be more prevalent and relevant. GS will be used as a baseline for the technical tests and Cyber Test comparisons, as it is assumed that high school students are significantly exposed to areas of General Science.

Dr. Fechter continued by outlining the proposed methodology for this work using the bookmark method. For the bookmark method, test items are ordered in a booklet by difficulty, with the easiest items placed first and the hardest last. Standard setting panelists select a test item in the booklet that represents the "spot" where applicants have mastered enough content to be considerably exposed to the content. Panelists then place bookmarks representing the midpoint between two items were the bookmark sits. The midpoint is averaged across panelists to determine the cut score. Dr. Fechter then presented an example of the process.

Dr. Fechter continued by presenting the construct definitions for the subtests under exploration.

- AI assesses knowledge of automobile technology. Items are designed to measure an examinee's basic knowledge, procedures, and principles, including automotive repair.
- EI assesses knowledge of electricity and electronics. Items are designed to assess an examinee's aptitude for understanding electrical currents, circuits, devices, and systems.
- SI assesses knowledge of tools and shop terminology and practices. Items are designed to measure an examinee's basic knowledge of general shop practices and building construction.
- MC assesses knowledge of mechanical and physical principles. Items are designed to assess an examinee's aptitude for principles of mechanical devices, structural support, and basic properties of certain materials. Problems cover the principles of gears, pulleys, levers, etc., as well as force and fluid dynamics. Items require general reasoning skills and the manipulation of spatial concepts.
- GS assesses knowledge of physical and biological sciences. Items cover three content areas: Life Science, Physical Science, and Earth/Space Science. The items are designed to measure the examinee's

ability to recognize, apply, and analyze scientific principles, including the facts, concepts, theories, and laws of science.

- The Cyber Test assesses knowledge of information and communication technology. Items are designed to assess examinees' aptitude for networking and telecommunications, computer operations, security and compliance, and software programming.

Dr. Fechter continued by providing an overview of the process to be followed. A panel of experts (two for each subtest) will be selected. These could be military training personnel, high school vocational teachers, post-secondary vocational/community college instructors, or members of the associated business community. Panelists should be considered SMEs in the content the subtest measures and should understand the variation of knowledge and skills of the youth population as it relates to the content area. The standard setting process is best conducted with all panelists gathered at the same location, typically a hotel conference center. The expected time commitment is three days. The process will begin with an orientation to standard setting and an administration of a sample test, which panelists will self-score. Two rounds of bookmarking will then be conducted. Feedback between rounds will include: the distribution of cut scores across panelists, impact data including the percent classified as significantly exposed, and National Assessment of Educational Progress High School Transcript Study longitudinal results. Cut scores will then be calculated and a definition of "significant exposure" will be developed for each test. The final step will be to evaluate the standard setting process.

Dr. Fechter concluded by identifying ways in which the DAC could contribute. These include offering technical advice on the proposed methodology (e.g., choice of the bookmark method, number of panelists, overall process), and offering technical advice on whether the outlined approach will sufficiently answer the questions of interest.

As Dr. Fechter described the requirement to norm-reference the ASVAB technical tests (slides 2-4), a committee member commented that tests administered adaptively present a unique set of challenges. Another committee member asked what Dr. Fechter meant by "significant exposure." Dr. Fechter replied that defining significant exposure would be part of the process, and possibilities may include exposure to course-related material as well as experiences outside the classroom. The committee member noted that it likely entails some level of mastery of the domain. Dr. Fechter said setting a cut score does imply mastery, at some level, but clarified that she was aiming at level of exposure to information, though she said high levels of exposure would not necessarily mean someone would perform well on the test. She said they were looking for a different way to frame the terms used for identification of assessment relevancy. Another committee member said the solution would likely have implications for validation, for example, making inferences about levels of exposure based on levels of performance. S/he said states use all sorts of levels to determine cut scores. Dr. Fechter explained that, at the end of her presentation, she would request recommendations on how to proceed. She added, however, that she may try to relate exposure to readiness for training, though the fact that Services use composite scores and not scores on single subtests would complicate the process of defining readiness. The committee member asked if the tests in the composites were weighted rationally or empirically. Dr. Segall said the Army uses optimally weighted subtests, but that other Services use integer weighted standard scores. Dr. Fechter said that each Service is responsible for validating its own composite scores.

Regarding the proposed bookmark method (slides 10-14), a committee member said that, when all examinees take a common form of a test, the form is used in assembling ordered item booklets for bookmark exercise participants. S/he asked how, in an adaptive context, all the items would be presented. Dr. Fechter explained that they would select items from the pool that

spanned the full range of difficulty. Another committee member asked what question would be presented to the SMEs. Dr. Fechter said they would be asked to find the place in the booklet that represents the level of knowledge required to be considered significantly exposed to the content of the subtest. She said the definition of "significantly exposed" would be developed by the panelists (i.e., SMEs) prior to the standard setting activities. She explained that panelists would be given an orientation to the material covered in training prior to the exercise. She added that prior research by American Institutes for Research (AIR), conducted in 1997, also explored whether high school students were significantly exposed to science and technical areas tested with the ASVAB subtests. In the AIR research, a questionnaire was developed and administered to students that solicited information about activities and coursework students engaged in that would help to define the level of exposure they had to science and technical areas before taking the ASVAB. The questionnaire and AIR findings could be used to help inform SMEs for developing definitions of "significant exposure" and in making judgments for where to place bookmarks during the standard setting activities.

A committee member then suggested that the process was mixing simple exposure with difficulty. Dr. Fechter agreed and acknowledged that a two-part inference was required. The committee member suggested that Dr. Fechter should target ability. Another committee member said s/he did not care about how people got exposed, only that they could perform at a level considered to be mastery. Another committee member suggested the goal should be to identify people with minimally acceptable performance on the test, which would be a clear way to express the requirement. She added that if the goal was to determine who could be successful in training, they should be able to express those requirements. S/he mentioned the Angoff method and reiterated the value of identifying the minimum level of knowledge required for mastery.

Another committee member asked if Dr. Fechter was planning to develop the definition of minimum performance prior to or during the workshop and encouraged her to do it ahead of time. The committee member said that would allow her to ensure the ordered item booklet includes content that matches the definition. S/h emphasized the need to do this given the large item pool and the importance of avoiding a disconnect between booklet content and the definition.

Dr. Fechter said that the process includes a step where panelists take a sample test so that they are familiar with the content. A committee member noted that the adaptive nature of the test would limit the effectiveness of this approach and asked for more information about the process. Dr. Fechter said all panelists would receive the same items. The committee member noted, however, that panelists would not be receiving the candidate's test. S/he noted also that each panelist would be performing the bookmarking exercise differently in accordance with their level of expertise, and that some would be better positioned than others. Another committee member asked if the tests were unidimensional and if content coverage would be a factor in item selection. Dr. Segall said they planned to balance content in the test booklets.

On hearing that panelists would be developing the performance level descriptors, a committee member remarked on the difficulty of that task and suggested providing a policy level starter version that the panelists could tweak. S/he emphasized the importance of not leaving that task solely to the panelists, but also said it should not be left to just one person. This prompted Dr.

Velgach to underscore the value of the panelists' experience in accomplishing the task. The committee member responded, however, that involving panelists could extend the length of the workshop by up to two days. Another committee member asked why it should take so long to define minimal acceptable performance. The first committee member explained that the panelists would have to derive connections between the test content and the prerequisite knowledge and skills. S/he said that, before using descriptors to set standards, the descriptors should be confirmed as sensible and understandable. Dr. Segall responded by saying the stakes were not that high in this case, because the purpose was only to decide whether there are enough people in the population to justify providing the test.

Another committee member said it was important to remember that the tests are only used in composite scores. Dr. Fechter also reiterated that the focus was on evaluating the test and not on setting standards for selection or classification decisions. A committee member pointed out, however, that once cut scores are set, people are going to want to use them. Dr. Velgach then clarified that the standard setting work would produce just one of many criteria for evaluating the tests. A committee member then suggested taking a step back to revisit the purpose of the effort. S/he said, if the purpose was to make decisions about the future utility of the tests, then maybe the standard setting approach was not the best method to use. S/he suggested that DPAC might, instead, conduct a survey to see what information accessed applicants possessed prior to taking the test and if that information was useful in their military occupational specialties (MOS). Dr. Velgach replied that the effort might have been mislabeled, and said the requirement was to understand the required level of knowledge specialization and how many people have it. She reiterated that this was an important point to understand.

To provide more background information, Dr. Segall said they had started by looking at historical data—score trends over time for the auto and shop tests—and had seen that scores had decreased over time. He said, however, they did not know what that meant in relation to exposure or mastery, which prompted the effort to determine a cut point. Dr. Velgach then said the this should have an impact on inferential validity; for example, if only 2-3% of the population had enough knowledge to score well on the test, there might be some other indicator that would have more utility. Dr. Pommerich suggested that the effort could be dialed back to a pseudo-standard setting process. A committee member said s/he still did not see how that would answer the primary question. Another committee member explained that focusing on a cut point would cause them to miss information about the distribution, which would be more informative in relation to the question. Another committee member said the point of interest is not how many pass the test, because that does not necessarily have anything to do with exposure.

After recognizing the ultimate need for a cut point, a committee member again stressed the value of the distribution, citing the possible case that the distribution might be shrinking, even though the mean might be the same. S/he also pointed out the value of looking at variances. Another committee member said DPAC could examine trends over time and make interpretations about whether the knowledge base has changed. The first committee member mentioned that DPAC would also want to make a claim about readiness. Another committee member said that "readiness" would have to be defined, at which point Dr. Segall said that a cut score would be needed to do that, which is why they are trying to produce one. Dr. Fechter acknowledged that

part of the ASVAB evaluation plan was to explore historical score trend data, including distributions, for the technical and science tests.

A committee member then suggested that DPAC could identify the point that leads to success in training empirically. Another committee member said using that approach would eliminate the requirement to define significant exposure. Dr. Segall replied that he liked that approach, and that they could use existing data to complete the task. After a short exchange about the potential need for different cut scores across jobs, a committee member restated that developing cut scores in the workshop did not appear to be the most straightforward way to meet the requirement. Another committee member suggested dropping the effort to determine exposure and using success in training to determine cut scores. Another committee member said training success would be a good reference point.

Moving forward, a committee member said that it would take a lot of time to determine what people needed to know to be considered qualified. S/he said, considering the purpose, the fundamentals of determining a cut score are predicated on a good, solid performance-level definition. Another committee member said that there are whole papers devoted to that alone.

A committee member reiterated his/her lack of confidence in letting a panel define standards and stated that a given panel might not include the right people for the task. Another committee member said that, for a bookmark process, the procedures looked good. S/he added, however, that the tough part would be defining the policy level descriptors and determining the focus of the standard setting task. Dr. Fechter said that was one of her questions: how would the descriptors be adequately defined?

Another committee member said it would help to look at the data before confirming the process, what will be inferred, and how it will be used. S/he said it would be important to determine if the data allow the types of inferences that need to be made. S/he continued to emphasize the value of the distribution, including the mean and variance. Another committee member suggested the possibility of defining several reference points instead of only one. The first committee member said non-group percentiles could be used for that purpose. Another committee member suggested looking at the percentage of people who pass training at various points on the scale. When a third committee member returned to the possibility of looking at what applicants would need to know, another committee member pointed out that the process would have to be performed for many jobs and restated that the overall purpose was to make an inference about the population.

A committee member suggested that it would make sense to administer some tests only to individuals who are interested in certain jobs. Dr. Segall replied that the Services do administer special tests for select occupations, and that the ASVAB has tried to standardize that process. He said, however, that it may be difficult to administer one of the technical tests as a special test. At that point, Dr. Velgach clarified that in many cases they want more, not fewer, people to take the technical tests. A committee member noted the vast amount of complexities that define the situation.

To conclude the discussion, a committee member summarized his recommendation to use existing data to define the cut points instead of asking a group of people to do it. S/he said using

an outcome, such as success in training, would be better than asking a group of people from different branches and specialties who may have varying opinions, with a median that might not be correct. Another committee member suggested rethinking what DPAC wanted to do considering the questions they are trying to answer. A third committee member restated the concern about the potential level of disagreement that may occur within a panel, even if they knew what was being taught at the various military and civilian schools. A fourth committee member said s/he could not even get people in the same company to agree. When a fifth committee member restated DPAC's interest in determining the level of mastery in the population, Dr. Fechter said the purpose of the test is to indicate that a person might be a good candidate for training within the technical area, not necessarily that he/she had mastered the content.

13. Future Topics (Tab P)

Dr. Dan Segall, DPAC, presented the briefing.

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

- ASVAB Resources
- ASVAB Development (pool development, evaluating/refining item and test development procedures, item writing guidelines and tools)
- Adverse Impact
- P*i*CAT/V$_{TEST}$ (Verification Test) Updates
- AFQT Prediction Test
- TAPAS Evaluation
- Test Security Compromise
- ASVAB Validity (improving the validation process and a review of Service validity studies, ASVAB validity framework, criterion domain/performance metrics)
- Career Exploration Updates (web site, expert panel recommendations, *i*CAT expansion)
- Adding New Cognitive Tests (Cyber Test, Working Memory, Abstract Reasoning including Adverse Impact)
- Adding New Non-Cognitive Measures (personality and interest measures, Adaptive Vocational Interest Diagnostic)
- Automatic Item Generation
- Web and Cloud efforts
- Device Evaluation and Expansion
- ASVAB Evaluation (standard setting study, other evaluation efforts)

Dr. Segall informed the committee that they had been briefed at some point on all the topics listed on the slide, and that DPAC was open to revisiting any of them. He also said he was open to new topics. He also asked the committee to let him know if they thought a topic should be briefed sooner rather than later.

A committee member recalled that adverse impact was on a two-year cycle. The committee member then asked about anticipated progress on the Cloud effort. Dr. Segall said he was hoping to be in the middle of the lift-and-shift effort. Dr. Pommerich suggested it could be addressed by milestones, and the committee member said that would work.

Another committee member asked if there would be any value in receiving an update on MCt. When Dr. Cyrus Foroughi (Naval Research Laboratory) said that the data collection would not be completed until summer, Dr. Velgach said the briefing could wait until the fall.

A committee member asked if there would be anything new on standard setting. Dr. Segall said it depends on whether they do it. Dr. Fechter said that the plan is to collect data next April through June, but that, if the revised plan is enacted, she would want to brief again in the Spring. Dr. Velgach asked about progress on the device study, and Dr. Fetcher said February 1 is the last scheduled data collection, and that they should have something to present by March. Dr. Velgach said that would work.

Ms. Miller asked Dr. Segall for his recommendations on what should be briefed. Dr. Segall said they would talk through the topics at the Military Accessions Policy Working Group, which would give him a better idea of what could be briefed to the DACMPT. Dr. Velgach said they might be able to brief the TAPAS evaluation project. A committee member then said s/he would like a P*i*CAT-Verification Test (VTEST) update. A committee member closed the discussion by giving approval for whatever came up between now and then.

Closing the meeting, Dr. Velgach thanked everyone for their participation and asked if there were any comments from the public, of which there were none. She then explained that the next DACMPT meeting would be held in Monterey, CA, instead of in Carmel, CA.

Tab A

LIST OF ATTENDEES

Defense Advisory Committee on Military Personnel Testing (DACMPT)
September 26-27, 2019, Sonesta Philadelphia, Philadelphia, Pennsylvania

Name	Position	Organization
Dr. Michael Rodriguez, Chair	Professor of Quantitative Methods	DACMPT, University of Minnesota
Dr. Neal Schmitt	Professor Emeritus	DACMPT, Michigan State University
Dr. Barbara S. Plake	Professor Emerita	DACMPT, University of Nebraska-Lincoln
Dr. Kevin Sweeney	Vice President, Psychometrics	The College Board
Dr. Nancy Tippins	Owner and Manager	Nancy Tippins Group, LLC
Dr. Sofiya Velgach	Designated Federal Officer (attendance req'd by FACA)	Accession Policy Directorate
Ms. Stephanie Miller	Director	Accession Policy Directorate
Mr. Christopher Graves	Senior Scientist	Human Resources Research Organization
Dr. Daniel Segall	Director	Defense Personnel Assessment Center
Dr. Mary Pommerich	Deputy Director	Defense Personnel Assessment Center
Dr. Shannon Salyer	Manager, Career Exploration Center	Defense Personnel Assessment Center
Dr. Tia Fechter	Personnel Research Psychologist	Defense Personnel Assessment Center
Ms. Mary (Ellie) Stone	Supervisory Survey Statistician	US Marine Corps, Manpower and Reserve Affairs
Dr. Donna Duellberg	Voluntary Education Program Manager	US Coast Guard
Dr. Cristina Kirkendall	Research Psychologist	US Army Research Institute
SGM Louis Johnson	Accession Policy Integrator	US Army, G1
Dr. Mark Rose	Chief, Market Research & Analysis	US Air Force Recruiting Service

Dr. Sophie Romay	Senior Personnel Research Psychologist	Air Force Personnel Center
Mr. Robert Tiegs	Testing Director	US Military Entrance Processing Command
Dr. Steve Watson	Director	US Navy Selection and Classification
CDR Henry Phillips	Operational Psychology Department Head	Naval Aerospace Medical Institute
Dr. Cyrus Foroughi	Engineering Research Psychologist	US Naval Research Laboratory
Ms. Latica Woods	Business Planning Manager	US Navy Testing Sciences
Mr. Tom Blanco	Vice President	S&T Consulting
Dr. Tim McGonigle	Program Manager	Human Resources Research Organization
Dr. Laura Ford	Program Manager	Human Resources Research Organization
Dr. Matthew Allen	Program Manager	Human Resources Research Organization
Dr. Art Thacker	Principal Scientist	Human Resources Research Organization
Dr. Furong Gao	Senior Staff Scientist	Human Resources Research Organization
Mr. Michael Crookeden	Engineer	Perspecta
Mr. Matthew Ellis	Deputy Program Manager	Northrop Grumman Systems Corporation

Tab B

DEFENSE ADVISORY COMMITTEE ON MILITARY PERSONNEL TESTING
AGENDA

September 26-27, 2019
Sonesta Philadelphia
Philadelphia, Pennsylvania

September 26, 2019

0800-0830	Committee Member Breakfast	
0830-0900	Executive Session	Dr. Michael Rodriguez, Chair
0900-0915	Welcome and Opening Remarks	Dr. Sofiya Velgach OASD (M&RA)/AP*
0915-0945	Accession Policy Update	Ms. Stephanie Miller Director, AP*
0945-1015	Milestones and Project Schedules	Dr. Mary Pommerich DPAC/OPA*
1015-1030	*Break*	
1030-1115	Abstract Reasoning Evaluation	Dr. Furong Gao HumRRO*
1115-1200	Cloud 101	Mr. Matthew Ellis Northrup Grumman Systems
1230-1300	Lunch	
1300-1330	Social Media Project Update	Dr. Tim McGonigle HumRRO*
1330-1430	Automatic Item Generation	Dr. Isaac Bejar ETS*
1430-1445	*Break*	
1445-1545	CEP* Update	Dr. Shannon Salyer DPAC/OPA*
1545-1645	Evaluation of the FYI*	Dr. Olga Fridman DPAC/OPA*
1645-1700	*Public Comments*	
1700-1730	Executive Session	Dr. Michael Rodriguez, Chair

September 27, 2019

0800-0830	Committee Member Breakfast	
0830-0900	Executive Session	Dr. Michael Rodriguez, Chair
0900-1000	ASVAB* Validity Framework	Dr. Art Thacker HumRRO*
1000-1100	Criterion Measurement	Dr. Laura Ford HumRRO*
1100-1115	*Break*	
1115-1145	Navy Validation Business Model	Dr. Stephen Watson Navy Selection and Classification
1145-1215	Standards Setting for ASVAB* Technical Tests	Dr. Tia Fechter DPAC/OPA*
1215-1230	Future Topics	Dr. Dan Segall DPAC/OPA*
1230-1245	*Public Comments*	
1245-1300	Closing Comments	Dr. Michael Rodriguez, Chair
1300-1500	Committee Working Lunch	

*** KEY:**

AP = Accessions Policy Directorate
ASVAB = Armed Services Vocational Aptitude Battery
CEP = Career Exploration Program, provided free to high schools nation-wide to help students develop career exploration skills and used by recruiters identify potential applicants for enlistment
DPAC/OPA = Defense Personnel Assessment Center/Office of People Analytics
ETS = Educational Testing Service
FYI = Find Your Interests
HumRRO = Human Resources Research Organization
NETC = Naval Education Training Command
OASD(M&RA)/AP = Office of the Assistant Secretary of Defense (Manpower & Reserve Affairs)/Accession Policy

Tab C

UNIVERSITY OF MINNESOTA

Twin Cities Campus

Quantitative Methods in Education
Department of Educational Psychology
College of Education and Human Development

170 Education Sciences
56 East River Road
Minneapolis, MN 55455

612-624-4324
mcrdz@umn.edu

October 6, 2019

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

Dear Ms. Miller:

The Defense Advisory Committee on Personnel Testing (DACMPT) is pleased to provide this committee report of our meeting of September 26-27, 2019, in Philadelphia, PA. Below, we provide summaries and recommendations from the DACMPT. The meeting was interesting and productive, and the DACMPT noted a number of times the advances being made on a number of projects.

The meeting began with opening remarks from Dr. Sofiya Velgach and Dr. Rodriguez (chair). Also, Drs. Barbara Plake, Neal Schmitt, Kevin Sweeney and Nancy Tippins were in attendance. In addition, staff and representatives from Defense Personnel Assessment Center (DPAC) and various military units were present.

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

Accession Policy Update
Ms. Stephanie Miller, the Director of Accession Policy, gave the DACMPT an overview of the Military Personnel Policy, reviewing the mission and the organizational structure of her unit. Although the focus of the meeting was testing issues, Ms. Miller mentioned other activities related to mental health, security clearances, medical standards, and draft registration. All services are on track to meet their recruiting mission except for the Navy Reserves. The shortfall in recruiting is attributed to the higher retention rates in the Navy, which limits the personnel who are moving to the Navy Reserves. Ms. Miller also shared information on Congressional Reports on the Armed Service Vocational Aptitude Battery (ASVAB) and recruiting for English Learners (non-native English speakers). The DACMPT concurred that the ASVAB is not an appropriate tool to use to evaluate the quality of education in counties across the U.S. and suggested a test like National Assessment of Educational Progress (NAEP) would be more useful. The DACMPT also discussed the implications of translating and adapting the ASVAB into Spanish, given the context of English-based training, orders, etc.

Major ASVAB R&D Efforts: Milestones and Project Schedules
Dr. Mary Pommerich of the DPAC, Office of People Analytics (OPA) provided an overview of the major R&D initiatives underway and updated the schedule for all the projects. This work included new Computerized Adaptive Testing (CAT)-ASVAB item pools; new CAT item pools for ASVAB Career Exploration Program (CEP); automating the generation of Arithmetic Reasoning (AR), Mathematical Knowledge (MK), and General Science (GS) items; the ASVAB technical bulletins; the ASVAB CEP; evaluation of new cognitive tests, including mental counters and the cyber test, and non-verbal reasoning tests; the addition of non-cognitive measures to selection and/or classification; the Air Force Compatibility Assessment (AFCA), the Defense Language Aptitude Battery (DLAB); and the web/cloud delivery of ASVAB and special tests. The DACMPT noted the impact of the freeze on pushing out new updates to various testing programs but understands the freeze is limited to software updates and does not include ongoing functional improvements.

Abstract Reasoning Test Evaluation
Dr. Gao presented a briefing on the evaluation of the Abstract Reasoning Test (ART), a non-verbal reasoning test developed by Embretson. Consideration of this test was in response to the ASVAB expert panel recommendation to examine a non-verbal reasoning test for possible inclusion in the ASVAB. Previous research with 728 Air Force recruits showed that ART loaded on quantitative reasoning factor (AR, MK, and Mechanical Comprehension (MC)) and had a strong positive relationship with Raven's Advanced Progressive Matrices (correlation of .78). The current study evaluated the ART with a sample of over 2000 military applicants who were interested in pursuing language training and who had already qualified for military service based on their Armed Forces Qualification Test (AFQT) scores. These applicants were a highly motivated applicants of high ability. Analyses based on this sample revealed that the ART had lower levels of adverse impact for gender and race/ethnicity than is the case for ASVAB subtests. Further analyses of ART in relation to ASVAB subtests and Mental Counters indicated that ART had a stronger loading on fluid intelligence reasoning than did the Mental Counters or ASVAB subtests. An additional analysis considered the fit of a 3PL Item Response Theory (IRT) model to the data; 25 of the 30 items showed adequate fit to the model. Based on that model, test-retest reliability was estimated at .77, which is probably an underestimate due to the restriction in ability due to the high ability sample. A response time analysis suggested that the time limit for administration should be increased by 5 minutes to ensure a 99% completion rate.

Conclusions and Recommendations
The DACMPT found the study on the utility of ART as a non-verbal assessment to be very promising due to the lower indicators of adverse impact for gender and race/ethnicity. However, the interpretation of the results should be cautioned, as recognized by the researchers, due to the non-representative nature of the sample. Based on results from a broader ability sample, it may be needed to develop harder items to better measure the construct of non-verbal complex reasoning. In addition, due to the fairly large number of response options (6-8 options), 1 or 2 parameter IRT models may provide a better fit to the data.

Cloud 101
A DPAC briefing updating the progress on the movement of computing and test delivery platforms to the cloud was provided by Matthew Ellis. Cloud computing service models were

described as a broad introduction to the mechanics and operations of systems that will bring DPAC computing operations and resources into the future, as well as provide needed stability to systems security. The DACMPT asked a few clarifying questions and appreciated the overview and update. The DACMPT looks forward to future updates on progress.

Social Media Project

Dr. Tim McGonigle of HumRRO updated the DACMPT on his work on the use of social media in military recruitment and selection. He explained that although social media has some promise for attraction and recruitment, its use in selection is limited. At best, social media might be used for prioritization of candidates who are predisposed to military careers and identification of those who are likely to succeed in training. Members of the DACMPT expressed their concerns about the fairness of using social media when some individuals have no social media footprint. There appear to be two essential questions:

1. Can you use social media to predict outcomes relevant to military attraction, recruitment, and selection?
2. Should you use social media for these purposes?

Automatic Item Generation

The DACMPT received a briefing from Dr. Bejar, Educational Testing Service, with technical updates on the ongoing effort to develop automatic item generation models and procedures. Currently, three areas have been addressed, including Work Knowledge (WK), MK, AR, and GS, focusing on components of the AFQT. GS is the newest addition to these efforts and was selected because it is primarily fact based and highly verbal (useful in the ASVAB testing programs to predict cheating). The WK models and process have been delivered. The DACMPT received information about the performance of the Math Knowledge models. The GS test has generated items that will be ready for field testing in January 2020.

Comments and Recommendations

The DACMPT acknowledges the positive direction of these efforts. The ability to automatically generate large numbers of items will increase test development efficiency. During the discussion of the MK item model performance, the DACMPT recommended to attend to distractor functioning when making judgments about item performance. Ideally, all distractors will function (be selected) at the lower levels of ability; each distractor should be selected at a relatively uniform rate, and that selection rate should decrease as ability increases. The DACMPT also acknowledged the challenges inherent in the GS item generation models, as the models require a biology knowledge base (ontology), and look forward to seeing future progress.

CEP Update

Dr. Shannon Salyer provided the briefing on the ASVAB CEP. She reviewed usage metrics, noting the increase in numbers of schools and students tested in 2019. Unfortunately, the website utilization data suggest that, although traffic is increasing, participants are not fully employing the online exploration tools. Dr. Salyer reviewed recommendations from the ASVAB CEP Expert Panel. A series of examples were provided regarding the use of ASVAB CEP in states, in some cases written into state legislation and Every Student Succeeds Act state plans. The troubling part of this trend includes potential inappropriate use of the ASVAB and the lack of

post-test interpretation (PTI) support for students. On the other hand, the PTI training currently under way is very promising, increasing the capacity of schools to meet PTI volume.

Comments and Recommendations
The DACMPT raised questions about the tracking of ASVAB CEP participants who retake the ASVAB multiple times. The DACMPT recommended the creating of a system to track retakes for the purpose of test security and evaluating the potential for practice effects. One possibility, as discussed at the meeting, is to alternate two ASVAB CEP forms across grades, as it appears most students may repeat the test across grades.

In order to address some of the promising recommendations from the Expert Panel, the DACMPT recommended that the CEP explore the opportunity to hire a graduate student intern for the summer. This would be a great opportunity for a counseling student. The DACMP supports the goal of increasing the visibility of the ASVAB CEP in academic journals.

The DACMPT encourages the program staff to continue communicating with state leaders, including the Council of Chief State School Officers, the National Association of Assessment Directors, and other state leaders, to minimize inappropriate test use.

Finally, the DACMPT briefly discussed the question of rebranding the ASVAB-CEP program by dropping the ASVAB component of the title and program materials. The DACMPT unanimously agreed that ASVAB is a core component of the program and should not be dropped from the title nor the marketing and program materials.

Evaluation of Find Your Interests Inventory
Dr. Salyer briefed the DACMPT, on behalf of Dr. Fridman, who was unable to be present at the meeting. Using Multidimensional Scaling (MDA) and Factor Analysis, this study attempted to replicate, using data from over 300,000 records, the RIASEC (Holland) hexagon model of interest dimensions. Although the results left some room for improvement, there was evidence of a 5-6 dimension representation from the data analyses. With regard to difference in results from males and females, again there was room for improvement in the comparability of the results, particularly with regard to the Realistic dimension, where it appeared that there was higher endorsement of the statements by males than for females, and less so for the Artistic dimension where females tended to show somewhat higher endorsement for the statements than did males. This raised two considerations: (1) should separate norm by gender be reported and (2) should these statements be revised to provide contexts that are more gender neutral.

Comments and Recommendations
The DACMPT agreed that the recovery of the RIASEC hexagon model, although not perfect, seemed fairly well represented in the data. Regarding reporting of gender-based norms, the committee was non-committal, discussing arguments of both pros and cons). The Committee supported the interpretive language that indicates that the interest results should be considered as guides and not prescriptions of job interests. Further they supported the current practice of reporting gender and combined norms for interpretation. Revisions of statements in the inventory should be done cautiously as it would disrupt the database of relationships between jobs and interests currently supported by the Find Your Interests Inventory.

ASVAB Validity Framework
Dr. Art Thacker of HumRRO provided a briefing on the validity argument approach to the validation of the AFQT and ASVAB. Beginning with the inferences that the military hopes to make with use of these measures in selection, he presented a theory of action for the AFQT and ASVAB, describing the logical steps or links in the inferential process along with the types of evidence that might support each of these links. He then provided a summary of the existing evidence for these links along with references to the reports that describe this evidence. Although there is a mountain of existing evidence related to each of these claims, the challenge is to organize this evidence from multiple sources and users with variation in the available evidence into a coherent evidence centered design.

Comments
The DACMPT believes that this effort represents a model effort in validation research and documentation. We look forward to continued work and the production of a technical manual.

Criterion Measurement
Dr. Laura Ford briefed the DACMPT on DOD-wide first-term enlisted service member criterion measurement. This addresses the effort to identify and develop a unified set of criterion instruments or metrics that could be used by all services. The Criterion Measures Advisory Panel supported the development of a taxonomy and documented existed criterion measures, leading to recommendations regarding evaluation criteria for validation research efforts. These activities have been carefully documented and resulted in the identification of 74 current criterion instruments. Dr. Ford mentioned that the existing measures did not cover a number of important areas, including psycho-social well-being, counterproductive work behaviors, and physical endurance. This work has identified a series of activities that are moving the goals in a productive direction. This includes deeper attention to existing administrative data, development of attitude measures and performance measures, and alignment of outcome criteria across services. The DACMPT sees these efforts as essential to the evidence base for criterion-related validation and looks forward to future briefings on the progress being made.

Navy Validation Business Model.
The Navy approach to continuous validation of various composites used to evaluate applicant potential was described by Dr. Steve Watson. The criterion in their studies is training success (defined as successful completion without a setback or recycle). These studies are conducted automatically and annually for various alternate composites of ASVAB tests. Since training success and test scores are readily available, validations can be conducted automatically. Qualification rates, adverse impact levels, and setback or attrition rates can be computed for various cut scores and alternative test or composite use. Impact sheets describing the alternatives and risk/benefits associated with each are readily produced and provided to decision makers. Corrected validity coefficients in the thirties and forties (correlations of .3 and .4) were reported with some change in adverse impact across test composites. These impact sheets have proven to be especially helpful to decision makers as they consider alternative selection strategies.

Comments
The DACMPT believes this is a very useful way to present the results of validation studies and applauds this approach to continuous validation of test use. We also encourage the plan to consider alternate outcomes other than the dichotomous training success criterion now used.

Standard Setting for ASVAB Technical Tests
Dr. Tia Fletcher provided a briefing on a proposed standard setting study for the ASVAB technical tests as a way to consider the utility of ASVAB technical tests. Using the criterion of "significantly exposed", the intent of the standard setting methodology was to set a performance standard on each of the technical tests using a Bookmark standard setting procedure.

Comments and Recommendations
The DACMPT had several concerns about this approach, first wondering if the criterion of interest was "significant exposure" to the content of the technical tests, or whether mastery of the domain of knowledge represented by the technical test would be more relevant to the intended purpose of the study. Further, though, the Committee wondered if a standard setting methodology was the most appropriate approach to examining the utility of these technical tests. The Committee encouraged the researchers to consider more empirical ways to examine this question looking at trends in performance data on these subtests over time.

Future Topics
Dr. Segall facilitated a discussion about future topics and priorities for the DACMPT. Although the DACMPT is pleased to continue receiving updates on most topics and projects underway, there was particular interest in hearing about Pending internet Computerized Adaptive Test (PiCAT)/Verification Test (VTEST), Tailored Adaptive Personality Assessment System (TAPAS) evaluation, device evaluation and expansion, ASVAB evaluation including the plan to address criterion referenced standard for ASVAB tests, and updates on the efforts to move to cloud computing and test administration.

The DACMPT was pleased to hear about the progress being made on numerous fronts and attributes much of the impressive progress to full funding that the program has received. We continue to encourage the federal government to maintain full funding for the ASVAB program and associated projects, to maintain a high level of security and quality, which meets the DoD goals for accession, training, and force readiness. Overall, the meeting was informative and useful. The DACMPT appreciates the high quality efforts of Accession Policy and DPAC staff, and the research staff of each of the services. Their frank interactions with the committee continue to be helpful and appreciated. We look forward to our next meeting.

Sincerely,

Michael C. Rodriguez, Ph.D.
Professor and Campbell Leadership Chair in Education & Human Development
Chair, Defense Advisory Committee on Military Personnel Testing

Tab D

OASD Manpower & Reserve Affairs

Military Personnel Policy
(Accession Policy)

MANPOWER & RESERVE AFFAIRS

Our Mission

Develop, review, and analyze policies, resources, and plans for Services' enlisted recruiting and officer commissioning programs

"Stewards of the All-volunteer Force"

Total Force Accession Policy

Director
Ms. Stephanie Miller

Military Entrance Processing Command (MEPCOM)
Commander: Col Richard Brady, USMC
65 Military Entrance Processing Stations
Personnel: 2,884 authorized (Military-568 and Civilian-2,256)

Joint Advertising Market Research and Studies "JAMRS" (OPA)
Dr. Katherine Helland

Personnel Testing Center (OPA)
Dr. Eric Lang

Joint Recruiting Facilities Program
Mr. Kevin Quinette

Accessions Systems
Mr. Chris Arendt

Defense Personnel Assessment Center "DPAC" (OPA)
Dr. Daniel Segall

Testing Standards
Dr. Sofiya Velgach

GI Bill Programs
Ms. Patricia Leopard

Enlisted Recruiting and Marketing Programs
Mr. Dennis Drogo

Recruiting Resources, Research and Analysis
Ms. Evelyn Dyer

USMEPCOM Liaison Officer
MAJ Maria Sanchez, USA

Personnel Security & Military Naturalization
COL Mike Mayes, USAR

Officer Commissioning Programs
Lt Col Naomi Henigin, USAF

Medical Accession Standards
LTC Peggy Urbano, USAR

Reserve Programs and Incentives
LTC Steve King, USAR

IMA Positions

Team Director - CAPT John Bellissimo

Recruiting Analyst – Lt Col Kelli Beaty

Officer Programs Analyst - CDR Suzy Tovar
- PCS Luis Dall'Orso

Enlisted Programs Analyst - COL Ginger Norris

Critical Items

Testing

- PiCAT: Internet-administered version of the enlistment test
- APT: Internet-administered screening test
- ASVAB Career Exploration Program
- Transition to the Cloud
- Device Evaluation
- Character Assessment Working Group

Mental Health and Department of Defense Military Family Readiness Council

Raised concerns regarding "alleged" unfairness of access to dependent medical records

- Informed Council that future medical screening processes moving toward verifiable medical data for all applicants; supported by following initiatives: Joint Legacy Viewer (JLV); Prescription Medication Reporting System (PMRS); MHS Genesis

Security Clearance Process

Enhanced Screening Protocol (ESP): Uniform standards and a centralized process for the screening and vetting of individuals seeking access to DoD systems, facilities, and information with foreign influence and foreign preference concerns

Medical

- Recent deaths at boot camp raised concerns from Lionheart Foundation
  - Championing for EKGs during MEPs physical

National Commission on Service

- Requested USD P&R testify on mobilization requirements; Raised concerns that Department does not have a current for plan for MEPCOM to execute in case of national emergency
  - MEPCOM currently assessing level of effort/timeframe to update plan
- NCoS Interim Report released January 2019; final report March 2020

Fiscal Year 2019 Mission

Service	Goal
Army – Active, Guard, and Reserve	122,600
Navy – Active and Reserve	47,162
Marine Corps – Active and Reserve	40,155
Air Force - Active, Guard, and Reserve	47,132
DoD Total	257,049

Source: Services

The Department of Defense is also projected to gain approximately 29,000 officers in 2019

Mission Attainment-August 2019

- Fiscal Year 2019 -	Active Recruiting/Accession Data				
	Annual Goal	FYTD Goal	FYTD Accessions	FYTD Percent of Goal	
Army	68,000	59,040	58,060	98.34	Y
Navy	39,000	35,792	35,783	99.97	Y
Marine Corps	31,767	28,467	28,528	100.21	G
Air Force	32,300	30,049	30,153	100.35	G
Total	171,067	153,348	152,524	99.46	

KEY: *100 percent of goal or above*; *90-99 percent of goal*; *below 90 percent of goal*

- Fiscal Year 2019 -	Reserve Recruiting/Gains Data				
	Annual Goal	FYTD Goal	FYTD Gains	FYTD Percent of Goal	
Army National Guard	39,000	35,834	35,953	100.33	G
Army Reserve	15,600	14,110	13,848	98.14	G
Navy Reserve	8,162	7,823	6,787	86.76	R
Marine Corps Reserve	8,388	8,137	8,390	103.11	G
Air National Guard	9,422	8,415	9,110	108.26	G
Air Force Reserve	5,410	5,392	6,780	125.74	G
Total	85,982	79,711	80,868	101.45	

KEY: *100 percent of goal or above*; *90-99 percent of goal*; *below 90 percent of goal*

New Recruit Quality
All Components

	*HSDG		**AFQT Cat I-IIIA		***AFQT Cat IV	
Active Components						
Army	93.8	G	60.7	G	1.90	G
Navy	97.6	G	71.8	G	0	G
Marine Corps	99.5	G	69.3	G	0	G
Air Force	98.4	G	81.9	G	0	G
Reserve Components						
Army National Guard	97.3	G	63.6	G	4.10	Y
Army Reserve	97.1	G	65.4	G	0.94	G
Navy Reserve	97.2	G	73.6	G	0.0	G
Marine Corps Reserve	99.3	G	75.6	G	0.0	G
Air National Guard	100.0	G	76.7	G	0.04	G
Air Force Reserve	98.5	G	74.8	G	0.0	G

Quality Key: ***100 percent or above benchmark; 90-99 percent benchmark; below 90 percent benchmark***

***HSDG:** Percent High School Diploma Graduates; *Department of Defense Benchmark ≥ 90 percent*
** **AFQT Cat I-IIIA:** Percent scoring at / above 50th Percentile on the Armed Forces Qualification Test; *Department of Defense Benchmark ≥ 60 percent*
*** **AFQT Cat IV:** Percent scoring at / below 30th Percentile on Armed Forces Qualification Test; *Department of Defense Benchmark ≤ 4 percent*

Congressional Reports

- Report on the Armed Services Vocational Aptitude Battery
  - 10 years of applicant data
  - Demographic information (age, gender, age, state, education level, etc.)
  - Number in Category V (AFQT < 10)
  - Counties scoring in lowest 5 percent
  - Sharing information with DoEd
- Report on Recruiting for English Learners (Non-native English Speakers)
  - Enlistment testing practices (ASVAB)
    - Mental ability
    - Academic potential
    - Academic achievement
  - Marketing efforts
  - Recruiting interactions
  - Enlistment rate

Questions?

Tab E

Major ASVAB R&D Efforts

Milestones and Project Schedules

Mary Pommerich

Briefing presented to the DAC

Philadelphia, PA

September 2019

Projects

- **ASVAB Development**
  - New CAT-ASVAB Item Pools
  - Developing New CAT Item Pool for CEP
  - Automating Generation of AR and MK Items/GS Items*
  - ASVAB Technical Bulletins
- **Career Exploration Program***
- **ASVAB and ETP Revision**
  - Evaluating New Cognitive Tests for ASVAB
    - Nonverbal Reasoning Tests*
    - Mental Counters
    - Cyber Test
  - Adding Non-cognitive Measures to Selection and/or Classification
  - Expanding Test Availability
    - Web Delivery of Special Tests
    - Moving to the Cloud*
- **Air Force Compatibility Assessment**
- **Defense Language Aptitude Battery**

*Will be presented/discussed at this meeting.

NOTE: Dates given in this document are subject to change depending on available resources, unexpected issues that arise, and other factors that may be beyond our control. Any changes will be communicated as soon as possible.

New CAT-ASVAB Item Pools

- **Objective**
  - Develop CAT-ASVAB item pools (designated as Pools 11–15) from new items
- **Projected Completion**
  - New item pool implementation: Sep 2020
- **Subtasks**
  - Write items ✓
  - Pretest items (Summer 2018) ✓
  - Calibrate and scale items (Summer 2018) ✓
  - Conduct item screenings (May 2019) ✓
  - Identify item enemies (June 2019) ✓
  - Complete preliminary form assembly (July 2019) ✓

New CAT-ASVAB Item Pools (continued)

- **Subtasks** (continued)
  - Complete final form assembly (Aug 2019–Sept 2019)
  - Modify, test, and deliver CAT-ASVAB software and item pools to MEPCOM (Oct 2019–Nov 2019)†
  - Collect and analyze IOT&E data (Dec 2019–Jul 2020)††
  - Implement operationally in WinCAT and iCAT (Aug 2020– Sept 2020)

- **Predecessors**
  - ASVAB Item Development

- **Successors**
  - Operational administration of new CAT-ASVAB item pools
  - Final development of next set of item pools
  - Use of retired item pools in CEP, AFCT, P*i*CAT, APT

† Actual completion date tied to the release of WinCAT package 3.0 to MEPCOM
† Actual start/completion dates dependent upon MEPCOM's QA/deployment schedule

Developing New CAT Item Pool for CEP

- **Objective**
  - Build a CAT item pool from P&P Forms 20B, 21 A & B, and 22 A & B. The new CAT pool is for use in the implementation of CEP *i*CAT
- **Projected Completion**
  - Fall 2019
- **Subtasks**
  - CAT Pool
    - Compute preliminary score information functions for CAT pool (Aug 2010) ✓
    - Review content for obsolescence, accuracy, sensitivity (Aug–Oct 2010) ✓
    - Compute final score information functions and evaluate (Nov 2010) ✓

Developing New CAT Item Pool for CEP (continued)

- **Subtasks** (continued)
  - CAT Pool
    - Reformat items for electronic delivery (Dec 2010–Oct 2011) ✓
    - Load items into database and review (May 2012–Oct 2013) ✓
    - Modify software to incorporate Pools 4 and 10 for equating (May 2017) † ✓
    - Administer in MEPS to obtain final equating algorithms (Mar 2018) †† ✓
    - Conduct final equating analyses (Aug 2018) †† ✓
    - Implement in CEP *i*CAT (Fall 2019) †††
- **Successors**
  - Implementation of new CAT pool for CEP *i*CAT

† Dates impacted by DMDC Cyber Hardening Initiative
†† Dates dependent upon MEPCOM's QA and deployment schedule
††† Date dependent upon development/QA/deployment schedules for iCAT releases 16.0 and 16.1

Automating Generation of AR and MK Items*

- **Objective**
  - Develop procedures for automating Arithmetic (AR) and Mathematics Knowledge (MK) item generation so that AR and MK item pools can be replaced on a frequent basis
- **Projected Completion**
  - Mar 2020
- **Subtasks**
  - Review literature relevant to mathematics (Jan 2018) ✓
  - Model MK and AR items from existing items (May 2018) ✓
  - Construct item generation software (Jul 2018) ✓
  - Generate MK pilot items (Jun 2018) ✓
  - Generate AR pilot items (Aug 2019) ✓
  - Conduct MK data collection (Mar 2019–June 2019)
  - Assess MK item quality and parameter accuracy (Jul 2019–Aug 2019)
  - Conduct AR data collection (Aug 2019 – Nov 2019)
  - Assess AR item quality and parameter accuracy (Dec 2019–Jan 2020)
  - Provide final generator, interface, and documentation (Mar 2020)

Automating Generation of GS Items*

- **Objective**
  - Develop procedures for automating General Science (GS) item generation so that GS item pools can be replaced on a frequent basis
- **Projected Completion**
  - Sep 2020
- **Subtasks**
  - Review literature relevant to general science (Jan 2019)
  - Model GS items characteristics from existing items (May 2019)
  - Construct item generation software (Sep 2019)
  - Generate GS pilot items (Jan 2020)
  - Conduct GS data collection (Feb 2020–May 2020)
  - Assess GS item quality and parameter accuracy (Jun 2020–Jul 2020)
  - Provide final generator, interface, and documentation (Sep 2020)

ASVAB Technical Bulletins

- **Objective**
  - Develop a series of electronic ASVAB technical bulletins to meet APA standards
- **Projected Completion**
  - Ongoing
- **Subtasks**
  - CAT-ASVAB Pools 5–9 (Dec 2008) ✓
  - APT (Fall 2019)
  - CAT-ASVAB Pool 10 for CEP *i*CAT (Fall 2019)
  - CAT-ASVAB Pools 11–15 (Fall 2020)
  - Other ASVAB Studies (as required)
- **Predecessors**
  - New item pool development
  - New test development

Career Exploration Program*

- **Objective**
  - Revise/maintain all CEP materials (websites & print materials), conduct program evaluation studies, and conduct research studies, as needed
- **Projected Completion**
  - Ongoing
- **Subtasks**
  - Update and develop new military occupational profiles (May 2016) ✓
  - Revise printed materials for websites (Sep 2016) ✓
  - Implement revised CEP Website (Sep 2016) ✓
  - Develop CEP program briefings and materials for external sources, as needed (ongoing)
  - Develop CEP Research and Evaluation Plans (in progress)
  - Develop plans for implementing CEP iCAT in schools and assessing impact of eliminating paper-and-pencil ASVAB (ongoing)

Career Exploration Program*
(Continued)

- **Subtasks** (continued)
  - Redesign Careers in the Military Website (FY 2017) ✓
  - Enhance functionality of websites (ongoing)
  - Automate score hosting on websites (Dec 2018) ✓
  - Develop an application for the collection of Service Occupational data (UNIform) (in progress)
  - Cross-walk civilian and military occupations for inclusion in the OCCU-Find (in progress)
  - Conduct Needs Analysis for computerized testing (Dec 2018) ✓
  - Develop and conduct post-test interpretation training (Aug 2019) ✓
  - Revise program materials as suggested by Expert panel and evaluation efforts (ongoing)
  - Monitor State usage of ASVAB and ASVAB CEP as related to legislative changes (ongoing)

Evaluating New Cognitive Tests: Mental Counters

- **Objective**
  - Conduct a validity study that will evaluate the benefits of adding Mental Counters (MCt) to the ASVAB and will provide the data to establish operational composites that include MCt and operational cut scores for the new composites
  - Navy is lead on this project
- **Projected Completion**
  - TBD
- **Subtasks**
  - Modify Software (Apr–Oct 2011) ✓
  - MEPCOM QA & deployment (Oct 2012–May 2013) ✓
  - Conduct item analyses and possible revision of test (Sep–Dec 2013) ✓
  - Revise, if necessary, and conduct new item analyses (Apr–Jul 2015) ✓

Evaluating New Cognitive Tests: Mental Counters (continued)

- **Subtasks** (continued)
  - Conduct predictor and criterion data collection (Jun 2013–Nov 2015) ✓
  - Investigate psychometric properties (in progress)
  - Evaluate/refine instructions and practice items (in progress)
    - Updated MCt software for evaluation (Summer 19) ✓
    - Data collection (Fall 19–Spring 20)
    - Final Report (Summer 20)
  - Conduct predictor and criterion data analyses (TBD)
  - Examine projected impact of operational use of MCt scores for selected jobs (TBD)

- **Successors**
  - Possible revisions to ASVAB content (TBD)

Evaluating New Cognitive Tests: Cyber Test

- **Objectives**
  - Develop and evaluate the Cyber Test (CT), formerly known as the Information Communication Technology Literacy (ICTL) test
  - Air Force is lead on this project
- **Projected Completion**
  - Ongoing
- **Successors**
  - Possible revisions to ASVAB content (TBD)
- **Subtasks**
  - Phase I: Initial Development/Pilot Test (Feb–Sep 2008) ✓
  - Phase II: Predictive Validation Study (USAF & Navy) (Jan–Sep 2009) ✓

Evaluating New Cognitive Tests: Cyber Test (continued)

- **Subtasks** (continued)
  - Phase III: MEPS Data Collection I – Norms, Construct Validity, Subgroup Differences, New Form Development (2010–2014) ✓
    - Use as special test; seed new items to develop follow-on forms (Aug 2013) ✓
    - Operational implementation: Air Force (May 2014), Army (June 2014), Navy (Oct 2016), USMC (Oct 2018) ✓
  - Phase IV: MEPS Data Collection II: Operational Support/Adv. Development
    - Integrate CT scores into classification process (Oct 2015) ✓
    - Develop scoring and reporting procedures/responsibilities (in progress)
    - Analyze existing items and develop new items (Nov 2018) ✓

Evaluating New Cognitive Tests: Cyber Test (continued)

- **Subtasks** (continued)
  - Phase IV: MEPS Data Collection II: Operational Support/Adv. Development Continued
    - Develop CAT item pools (Dec 2018) ✓
    - Evaluate feasibility of CAT-Cyber Test (Feb 2019) ✓
    - Develop CAT-Cyber Test (TBD)
    - Conduct additional validation studies (TBD)
    - Program versions of the AF Electronic Data Processing Test and selected Cyber Aptitude and Talent Assessment (CATA) tests, to evaluate psychometric properties and incremental validity (AF) (in progress)
      - Complete programming (Feb 2018) ✓
      - Conduct initial data collection using basic military trainees (Aug 2018) ✓
      - Develop web-based versions (Oct 2019)
      - Evaluate psychometric properties (TBD)
      - Design predictive validation study to evaluate EDPT and CATA against training grades (in progress)

Evaluating New Cognitive Tests: Cyber Test (continued)

- **Subtasks** (continued)
  - Phase IV: MEPS Data Collection II: Operational Support/Adv. Development Continued
    - Administer CT for CTN training and collect data for analysis purposes (Navy) (TBD)
    - Conduct predictor and criterion data analyses (Summer 2019)†
    - Examine project impact of operational use of CT scores for selected jobs (Summer 2019)†
  - Develop in-Service version of CT (Army project) (in progress)
    - Phase 1: Develop item pool ✓
    - Phase 2: Pilot test new items ✓
    - Phase 3: Analyze pilot items and develop two parallel forms ✓
    - Phase 4: Implement the new forms for in-service testing (TBD)
    - Phase 5: Develop new administration platform (TBD)

†Assuming transmission of requisite data on Navy applicants from DPAC to Navy no later than Apr 2019

Evaluating New Cognitive Tests: Cyber Test (continued)

- **Subtasks** (continued)
  - Explore utility of a serious gaming approach to assess cyber aptitude (AF) (in progress)
    - Phase I: Literature review ✓
      - Review archival materials regarding aptitudes & traits needed for success in cyber career fields (FY 19) ✓
      - Document critical aptitudes for cyber jobs (FY 19) ✓
      - Summarize literature & recommendations on how serious gaming could be used to enhance assessment of cyber aptitude (FY 19) ✓
    - Phase II: Cyber game development
      - Initial development (Feb 2019–May 2020)
      - Validation (TBD)

Evaluating New Cognitive Tests: Cyber Test (continued)

- **Subtasks** (continued)
  - Develop game-based assessment of Systems Thinking Ability (STA) (Army project) (in progress)
    - Phase I: Develop validate component measures (2016) ✓
    - Phase 2: Incorporate measures into shell (in progress)
    - Phase 3: Conduct validation of STA (Q2 FY20)
    - Phase 4: Validate to cyber populations (TBD)
  - Develop test of capabilities not covered by established measures that predicts success in cyber- the Common Cyber Capabilities (C^3) Test (Army project) (in progress)
    - Phase I: Literature review (2016) ✓
    - Phase 2: SME meetings to identify capabilities (2016-18) ✓
    - Phase 3: Develop measure of selected capabilities (2018) ✓
    - Phase 4: Conduct validation of items and scales (2019) ✓
    - Phase 5: Combine developed and validated measures into one cohesive computer-administered self-scoring test (in progress)
    - Phase 6: Validate to cyber populations (Q1-Q2 FY20)

Evaluating New Cognitive Tests: Nonverbal Reasoning Tests*

- **Objective**
  - Address the ASVAB Expert Panel's recommendation to investigate including a test of fluid intelligence, such as a nonverbal reasoning test
  - Plan and conduct construct validation studies
- **Projected Completion**
  - TBD
- **Subtasks**
  - Evaluate nonverbal reasoning tests
    - Design research (Mar–Sep 2008) ✓
    - Modify Software (Sep–Nov 2011) ✓
    - Software Quality Assurance (Jan 2013–Jan 2015) ✓

Evaluating New Cognitive Tests: Nonverbal Reasoning Tests (continued)*

- **Subtasks** (continued)
  - Evaluate nonverbal reasoning tests continued
    - MEPCOM QA & deployment (Feb–Mar 2015) ✓
    - Collect data for DLAB bridge study (Sep 2015–Aug 2017) ✓
    - Analyze linking data & report results (Dec 2018) ✓
    - Evaluate Abstract Reasoning Test data (in progress)
    - Plan additional validation studies (TBD)

- **Successors**
  - Possible revisions to ASVAB content (TBD)

Adding Non-cognitive Measures to Selection and/or Classification

- **Objective**
  - Address the ASVAB Expert Panel's recommendation to evaluate the use of non-cognitive measures in the military selection and classification process
  - Army is lead on this project (excluding AF-WIN and JOIN efforts)
- **Projected Completion**
  - Ongoing
- **Successors**
  - Possible revisions to the ASVAB or addition of new special tests (TBD)
- **Subtasks**
  - Empirically evaluate Army measures of work interests (Work Preferences Assessment, formerly PE-Fit) using Army applicants
    - Program WPA for ASVAB Platform (Jan–Oct 2010) ✓
    - MEPCOM QA & Deployment (Oct 2012–July 2013) ✓
    - Begin data collection (June 2017) ✓

Adding Non-cognitive Measures to Selection and/or Classification (continued)

- **Subtasks** (continued)
  - Evaluate NCAPS and SDI items/scales, for possible use in TAPAS
    - Compile/review existing materials & psychometric data (Jan 2019) ✓
    - Administer TAPAS/NCAPS/SDI tests to Basic Recruits to examine construct validity (in progress) (Oct 2018) ✓
    - Examine psychometric evidence (FY19) ✓
  - Empirically evaluate the Tailored Adaptive Personality Assessment System (TAPAS)
    - Begin initial TAPAS testing on the ASVAB platform (May 2009) ✓
    - TAPAS use by Army for applicant screening (Jan 2010–ongoing)
    - TAPAS use by Air Force for classification and to evaluate for person-job matching (June 2014–ongoing)
    - Air Force analyses and presentation on score inflation, reliability, validity, and utility to date (June 2017) ✓
    - Air Force Testing Modernization effort:
      - Develop/Integrate new scales (e.g., Responsibility, Situational Awareness) into AF TAPAS (July 2018) ✓
      - Evaluate alternative item formats (e.g., unidimensional pairwise preference) (FY19)
      - Develop Dark Tetrad facet items (FY19)

Adding Non-cognitive Measures to Selection and/or Classification (continued)

- **Subtasks** (continued)
  - Empirically evaluate the Tailored Adaptive Personality Assessment System (TAPAS) continued
    - TAPAS testing of Navy applicants on ASVAB platform (Apr 2011–Mar 2013) ✓
      - Conduct analyses and evaluate impact for Navy applicants (Sep 2015–TBD)
    - USMC evaluation effort:
      - Begin initial research TAPAS testing on the ASVAB platform (Sept 2015) ✓
      - TAPAS use by USMC as an applicant requirement (Aug 2018 – ongoing)
      - Develop proof of concept predictor for Recruit Training success based on TAPAS and other accession-related features (April 2019) ✓
      - Follow-on analysis for MOS Selection and First Term of Enlistment Success (TBD)
      - Efforts to resolve data system limitations and procedural challenges (real-time scoring, algorithm updates, MEPS visibility on scoring, etc.) (Ongoing)
  - Develop and evaluate an Army interest inventory (AVID)
    - Identify basic interests ✓
    - Develop items, pretest items, and conduct preliminary analysis ✓
    - Develop computer adaptive software (Fall 2017) ✓
    - Conduct initial validation study (Summer 2018) ✓
    - Expand concurrent validation evidence (Fall 2020)

Adding Non-cognitive Measures to Selection and/or Classification (continued)

- **Subtasks** (continued)
  - Develop, evaluate, and implement an Air Force interest inventory (AF-WIN)
    - Update job profile markers for 65 career fields (Aug 2017) ✓
    - Complete validation analyses (Sep 2017) ✓
    - Implement AF-WIN on AirForce.com (CY 2018) ✓
    - Update all job profile markers through 30 Apr 19 AFECD (Jul 2019) ✓
    - Gather ordinal SME marker data on all enlisted AFSCs (July 2019) ✓
    - Develop processes to streamline/automate AF-WIN processes (Oct 2019)
  - Develop the Job Opportunities in the Navy (JOIN) personalized career interest assessment
    - Develop recruiting job/rating structure mode ✓
    - Develop for pre-service use (2017 Start; 2018 IOC)
      - Pilot version available for NRC use (Q3, 2017) ✓
      - Implement JOIN within recruiting process (08 Sep 2018) ✓
    - Develop new items and validate DNA (Q4, 2019)
    - Proof of Concept for gaming environment vice self report format (Ongoing through Q3, 2020)

Air Force Compatibility Assessment (AFCA)

- **Objective**
  - Program the Air Force Compatibility Assessment for WinCAT administration
- **Projected Completion**
  - TBD††
- **Subtasks**
  - Receive test specifications and instructions from Air Force (Nov 2016) ✓
  - Develop software (Dec 2016–Dec 2017)† ✓
  - Conduct software QA (Jan 2018–Jun 2018) ✓
  - Conduct psychometric scoring QC (Jun 2018–Aug 2018)
  - Release WinCAT package to MEPCOM (July 2019) ✓
  - Deploy in production environment (TBD)††
  - Program AFCA for web delivery (pending approvals, TBD)

† Dates have been impacted by the Cyber Hardening Initiative

†† Dates are dependent upon (1) Air Force approvals and (2) MEPCOM's QA and deployment schedule

Defense Language Aptitude Battery

- **Objective**
  - Transition to all computer-based testing and improve the predictive validity of the Defense Language Aptitude Battery
- **Subtasks**
  - Develop a computer-based DLAB that will run on the WinCAT platform in MEPS (Jan 2007–Jul 2008) ✓
  - Develop a web-based DLAB (Jan 2008–Jan 2009) ✓
  - Conduct an ASVAB/DLAB comparison (Sep 2009–Dec 2011) ✓
  - Develop a new generation of the DLAB (DLAB2) (Dec 2018) ✓
    - Collect data for an equating study (Sep 2015–Dec 2017) ✓
    - Perform DLAB equating analysis (Jan 2018–Dec 2018) ✓
  - Move Win-DLAB to the iCAT platform (Jan 2021-Dec 2021)†

† WinCAT is slated to be decommissioned at a date TBD.

Expanding Test Availability: Web/Cloud Delivery of ASVAB and Special Tests

- **Objective**
  - Transition delivery of special tests from Windows-based platform to web-based and/or cloud platform
- **Projected Completion**
  - Dec 2021
- **Predecessors**
  - Cyber hardening and code modernization (ongoing)
  - Develop cloud infrastructure (ongoing)
- **Subtasks**
  - Identify requirements and design transition (Jan 2018–Sep 2018) ✓
  - Migrate Test 1 to DMDC web-based platform (Oct 2018–Jul 2019)†
  - Modify iCAT software to accommodate special tests (Oct 2018–Jul 2019)
  - Modify iCAT-A&R software to accommodate special tests (Oct 2019–Oct 2019)

† Test 1 is the Cyber Test.

Expanding Test Availability: Web/Cloud Delivery of ASVAB and Special Tests (continued)

- **Subtasks** (continued)
  - Develop web service for transferring scores to MEPCOM (Apr 2019–Sep 2019)†
  - QA Test 1 on DMDC web platform (Aug 2019–Nov 2019)
  - Deploy Test 1 to Production on DMDC web platform (Dec 2019–Dec 2019)
  - Migrate Tests 2 and 3 to DMDC web platform (Aug 2019–Dec 2019)††
  - QA Tests 2 and 3 on DMDC web platform (Jan 2020–Feb 2020)
  - Deploy Tests 2 and 3 to Production on DMDC web platform (Mar 2020–Mar 2020)

† Ability to complete is impacted by MEPCOM's move to the cloud. Interim approaches TBD.

†† Tests 2 and 3 are tentatively slated to be Coding Speed and Mental Counters.

Expanding Test Availability: Web/Cloud Delivery of ASVAB and Special Tests (continued)

- **Subtasks** (continued)
  - SOFTWARE FREEZE (Feb 2020–Jul 2020)†
  - Migrate iCAT and iCAT-A&R to the cloud, including Tests 1-3, and QA (Feb 2020–Jul 2020)
  - Deploy iCAT & iCAT-A&R to Production in the cloud (Aug 2020–Aug 2020)
  - Migrate TAPAS to the cloud platform (Feb 2020–Aug 2020)††
  - Deploy TAPAS to Production in the cloud (Aug 2020–Aug 2020)
  - Decommission WinCAT (Aug 2020–Nov 2020)
  - Migrate Tests 4 and 5 to the cloud platform & conduct QA (Aug 2020–Jan 2021)†††
  - Complete deployment of Special Tests 1-5 to Production in the cloud (Jan 2021–Jan 2021)
  - Transition DLAB2 from WDLPT to special test environment (Jan 2021–Dec 2021)

† Software changes will *not* be allowed while iCAT & iCAT-A&R are transitioned to the cloud.

†† TAPAS will go straight to the cloud because the language it is programmed in is incompatible with the DMDC web. The transition start and end dates are dependent upon the development of the cloud infrastructure and could shift.

††† Tests 4 and 5 are tentatively slated to be AFCA and Abstract Reasoning.

Appendix A
List of Acronyms

List of Acronyms

AF	Air Force
AFCA	Air Force Compatibility Assessment
AFCT	Armed Forces Classification Test
AFQT	Air Force Compatibility Assessment
AF-Win	Air Force Work Interest Navigator
AIM	Assessment of Individual Motivation
AO	Assembling Objects
APT	AFQT Predictor Test
ASVAB	Armed Services Vocational Aptitude Battery
ATO	Authority to Operate
AVID	Adaptive Vocational Interest Diagnostic
CAT-ASVAB	Computerized Adaptive Testing version of the ASVAB
C^3	Common Cyber Capabilities
CEP	Career Exploration Program
CS	Coding Speed
DHRA	Defense Human Resources Agency
DIF	Differential Item Functioning

List of Acronyms (continued)

DLAB	Defense Language Aptitude Battery
DLPT	Defense Language Proficiency Test
DMDC	Defense Manpower Data Center
ECL	English Comprehension Level Test
ETP	Enlistment Testing Program
IATT	Interim Authority to Test
*i*CAT	Internet-based CAT-ASVAB
iCAT-A&R	iCAT Authorization and Registration
ICTL	Information Communications Technology (CyberTest)
IOT&E	Initial Operational Test and Evaluation
IRB	Institutional Review Board
JOIN	Job Opportunities in the Navy
MCt	Mental Counters
MEPCOM	Military Entrance Processing Command
MET sites	Military Entrance Testing sites
MEPS	Military Entrance Processing Stations
NCAPS	Navy Computer Adaptive Personality Scales
NRC	Navy Recruiting Command

List of Acronyms (continued)

OCCU-Find	Occupational Finder
P&P	Paper and Pencil
Pay97	Profile of American Youth, 1997
PC	Paragraph Comprehension
P-E Fit	Person-Environment Fit
P*i*CAT	Prescreen (CAT) ASVAB
QA	Quality Assurance
QC	Quality Control
R&D	Research and Development
STA	Systems Thinking Ability
STP	Student Testing Program
TAPAS	Tailored Adaptive Personality Assessment System
TBD	To Be Determined
USMC	United States Marine Corps
WinCAT	Windows-based CAT-ASVAB
WPA	Work Preferences Assessment

Tab F

OPA

OFFICE OF PEOPLE ANALYTICS

Abstract Reasoning Test Evaluation

Presented to the DACMPT

Furong Gao, HumRRO
Ping Yin, HumRRO
Mary Pommerich, DPAC
Dan Segall, DPAC

September 26, 2019 | Philadelphia, PA

OVERVIEW

- **Background**
- **Previous research results**
- **New data analyses**
  - Item difficulty distribution; test reliability
  - Test score distributions by different demographic groups
  - Correlation with ASVAB subtests
  - Factor structures (constructs measured)
    - Evaluate together with the ASVAB data
    - CHC (Cattel-Horn-Carroll) theory of cognitive abilities
  - Test response time
- **Conclusion, discussion, and recommendations**

BACKGROUND

- **A test of non-verbal reasoning:***

Project Goals

- Respond to ASVAB Review Panel recommendation:
  - One or more non-verbal reasoning tests should be developed and evaluated for inclusion in the ASVAB. The evaluation process should consider the several aspects of validity. The efficacy of coaching and item familiarity, as well as the feasibility of creating multiple forms should be examined in conjunction with test development.
- Assess the quality of some experimental NVR measures by examining their relationships with a well-researched NVR marker test and ASVAB subtests (construct validation).

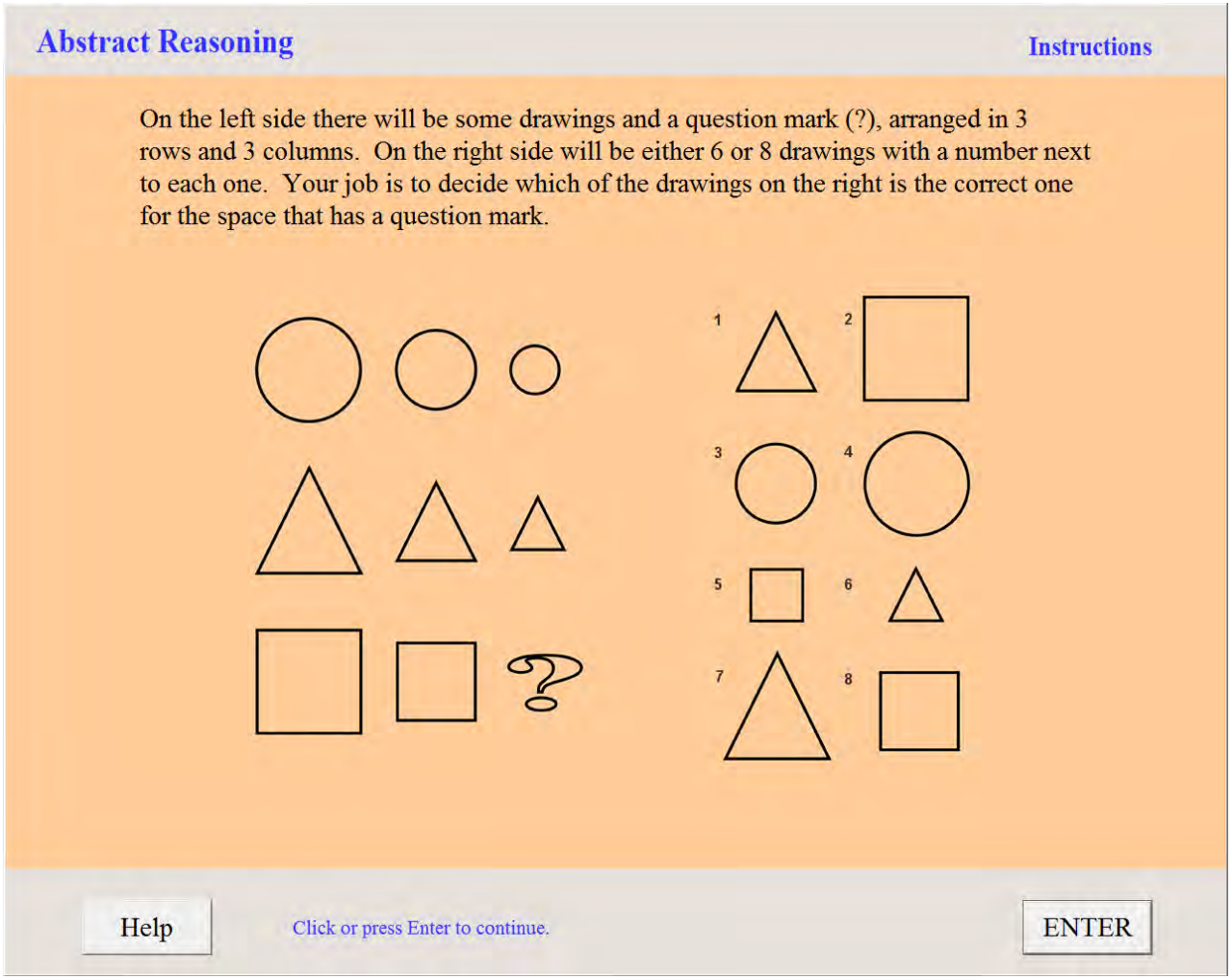
Abstract Reasoning Test

- Developed by Embretson (1998, 1999).
  - Uses a cognitive design system to automatically generate items.
- Has demonstrated a strong relationship with the Raven's Advanced Progressive Matrices (r = 0.784) in a sample of military recruits.
- Has demonstrated a similar pattern of relationships with the ASVAB subtests as the Raven's.
- Was found to load on the ASVAB quantitative reasoning factor in studies with military recruits.

OPA
OFFICE OF PEOPLE ANALYTICS

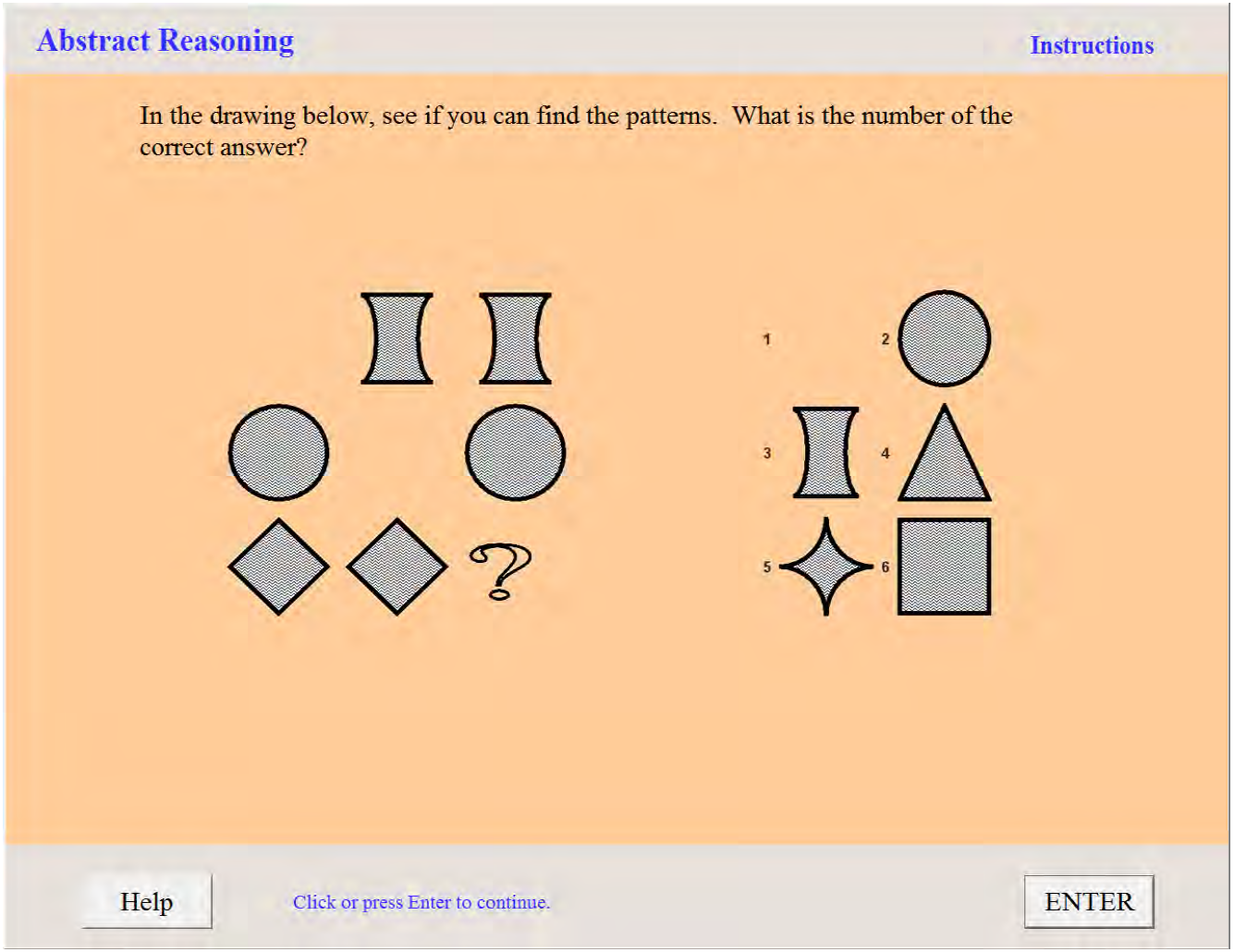
*From Pommerich (2008) DAC briefing

SAMPLE ABSTRACT REASONING TEST (ART) ITEM



OPA
OFFICE OF PEOPLE ANALYTICS

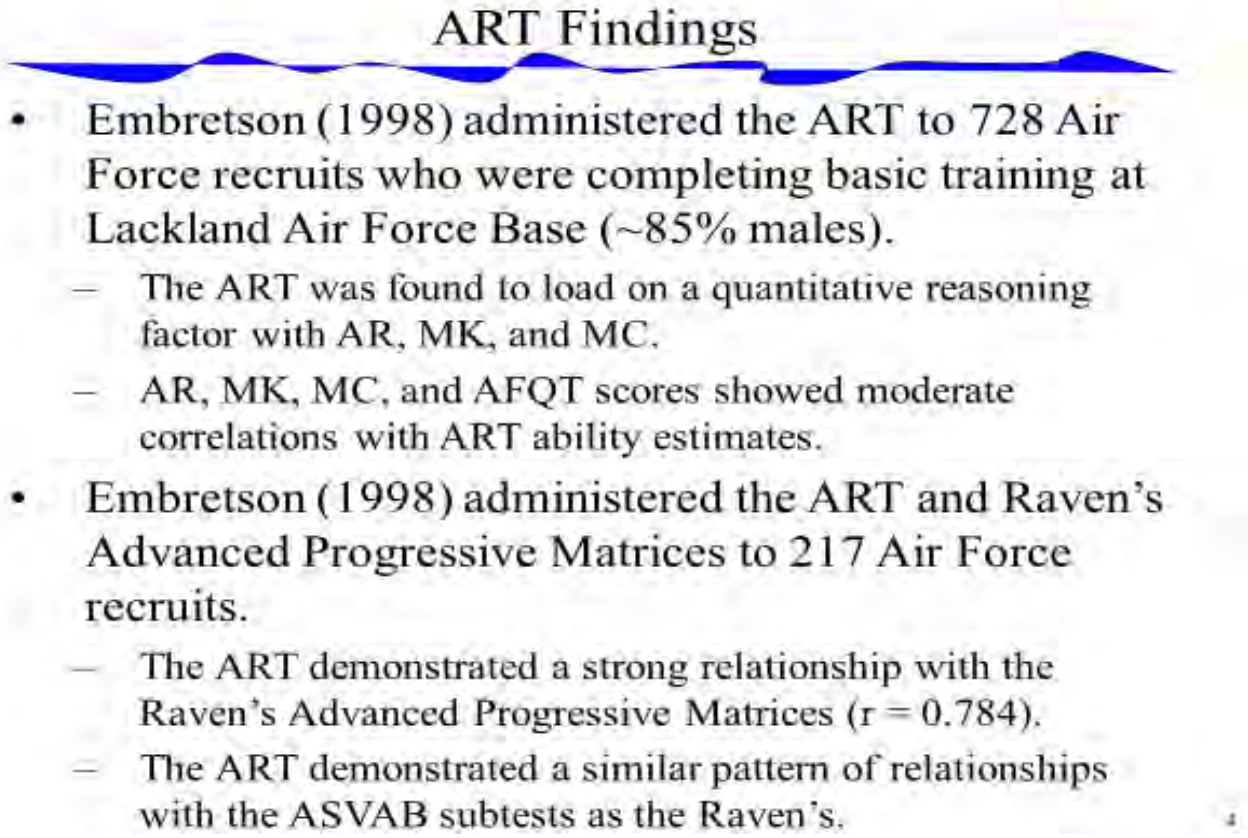
SAMPLE ART ITEM



OPA
OFFICE OF PEOPLE ANALYTICS

PREVIOUS RESEARCH RESULTS

- **Finding summary:***



OPA
OFFICE OF PEOPLE ANALYTICS

*From Pommerich (2014) MAPWG briefing

PREVIOUS RESEARCH RESULTS

ART Findings

- Descriptive Statistics and Factor Loadings for Generated ART Ability and ASVAB (N=728), from Embretson (1998).*

	Descriptive statistics			Factor loadings			
Test	*M*	*SD*	*r*	Verb	Quant	Tech	Spee
ART Ability	896	982	1.000		.635		
ASVAB							
GenSci	54.934	6.430	.268	.579		.377	
ArithRes	55.337	6.172	.487		.709		
WordKnow	54.253	4.342	.181	.733			
ParaComp	54.905	4.835	.118	.531			
NumOps	55.109	6.051	.109				.79
CodeSped	54.820	6.725	.069				.66
AutoShop	53.155	8.159	.125			.739	
MathKnow	56.922	6.544	.428		.723		.10
MechComp	56.125	7.206	.362		.329	.570	
ElecInfo	52.890	7.488	.201	.218		.671	
AFQT	221.486	15.711	.464				

Note. ART = Abstract Reasoning Test (Embretson, 1995b); ASVAB = Armed Services Vocational Aptitude Battery; Verb = verbal ability; Quant = quantitative reasoning; Tech = technical ability; GenSci = general science; ArithRes = arithmetic reasoning; WordKnow = word knowledge; ParaComp = paragraph comprehension; NumOps = numerical operations; CodeSped = coding speed; AutoShop = automobile shop information; MathKnow = mathematical knowledge; MechComp = mechanical comprehension; ElecInfo = electrical information; AFQT = Air Force qualification test.

*Correlations were not corrected for range restriction on the AFQT and therefore underestimate the correlation between the ASVAB subtests and the ART in the applicant population.

ART Findings

- Descriptive Statistics and Factor Loadings for ART, Raven's, and ASVAB (N=217), from Embretson (1998).*

			Correlations		Factor loadings			
Test	*M*	*SD*	ART	RAVEN	Verb	Quant	Tech	Speed
ART	23.327	6.562	1.000	.784		.656		
RAVEN	29.447	7.043	.784	1.000		.628		
ASVAB								
GenSci	54.903	6.578	.385	.409	.699		.384	
ArithRes	55.046	6.789	.467	.459		.758		
WordKnow	54.189	4.375	.335	.301	.710			
ParaComp	54.350	4.606	.250	.286	.601			
NumOps	55.037	6.174	.105	.124				.736
CodeSped	54.115	6.343	.217	.192				.692
AutoShop	53.134	7.783	.088	.112			.787	
MathKnow	56.000	6.839	.484	.504		.618		.281
MechComp	56.853	7.338	.362	.331		.372	.598	
ElecInfo	53.659	6.871	.153	.167	.258		.674	
AFQT	220.014	17.069	.545	.542				

Note. ART = Abstract Reasoning Test (Embretson, 1995b); RAVEN = Raven's Advanced Progressive Matrices Test (Raven, Court, & Raven, 1992); ASVAB = Armed Services Vocational Aptitude Battery; Verb = verbal ability; Quant = quantitative reasoning; Tech = technical ability; GenSci = general science; ArithRes = arithmetic reasoning; WordKnow = word knowledge; ParaComp = paragraph comprehension; NumOps = numerical operations; CodeSped = coding speed; AutoShop = automobile shop information; MathKnow = mathematical knowledge; MechComp = mechanical comprehension; ElecInfo = electrical information; AFQT = Air Force qualification test.

*Correlation between the ART and the Raven's is expected to be higher if corrected for range restriction on the AFQT.

OPA
OFFICE OF PEOPLE ANALYTICS

NEW DATA ANALYSES

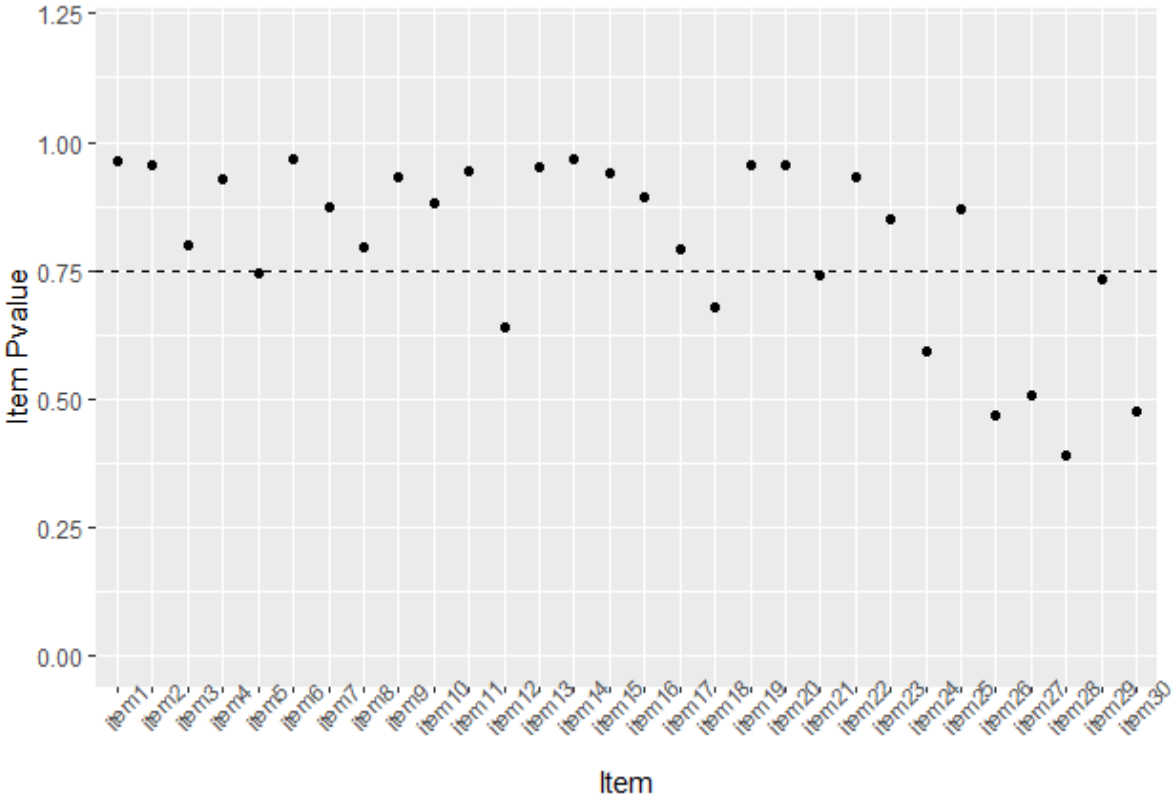
- **ART:**
  - 30 items, items scored right (1) or wrong (0)
  - 25 minutes testing time limit
- **2,162 test takers**
  - Military applicants who want to take language trainings
  - Already qualified based on their AFQT scores
  - Highly motivated, high abilities
- **Administered March 2017–September 2017**
- **Test-takers' other available test scores:**
  - ASVAB
    - From their enlistment profiles; taken 1, 2, or 3 years before
  - Mental Counters (MCt)

ART Item and Test Descriptive Statistics

ITEM DIFFICULTY, RAW SCORE DISTRIBUTION

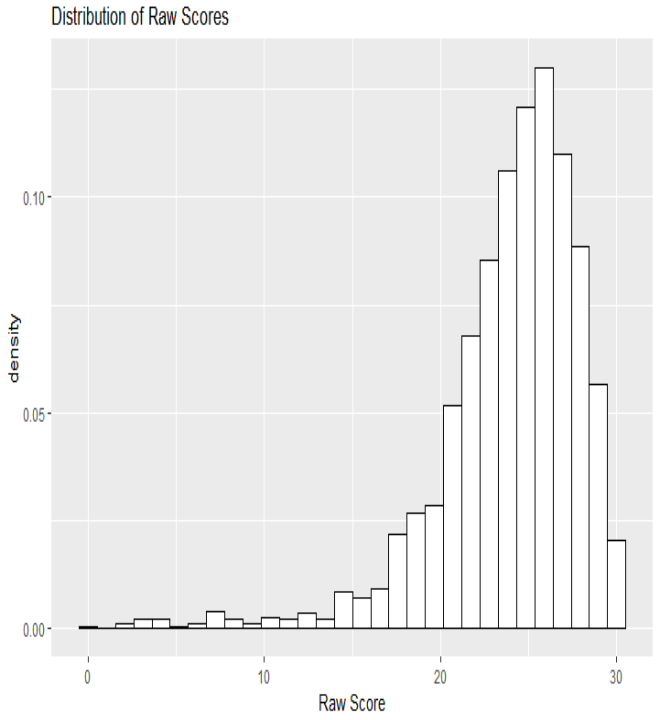
- **Item difficulty**

  80% of the item p-values >= 0.75



- **Raw score**

Min	Median	Mean	SD	Max	Reliability
0	25	24.1	4.2	30	0.803



RAW SCORE DISTRIBUTION—BY GENDER

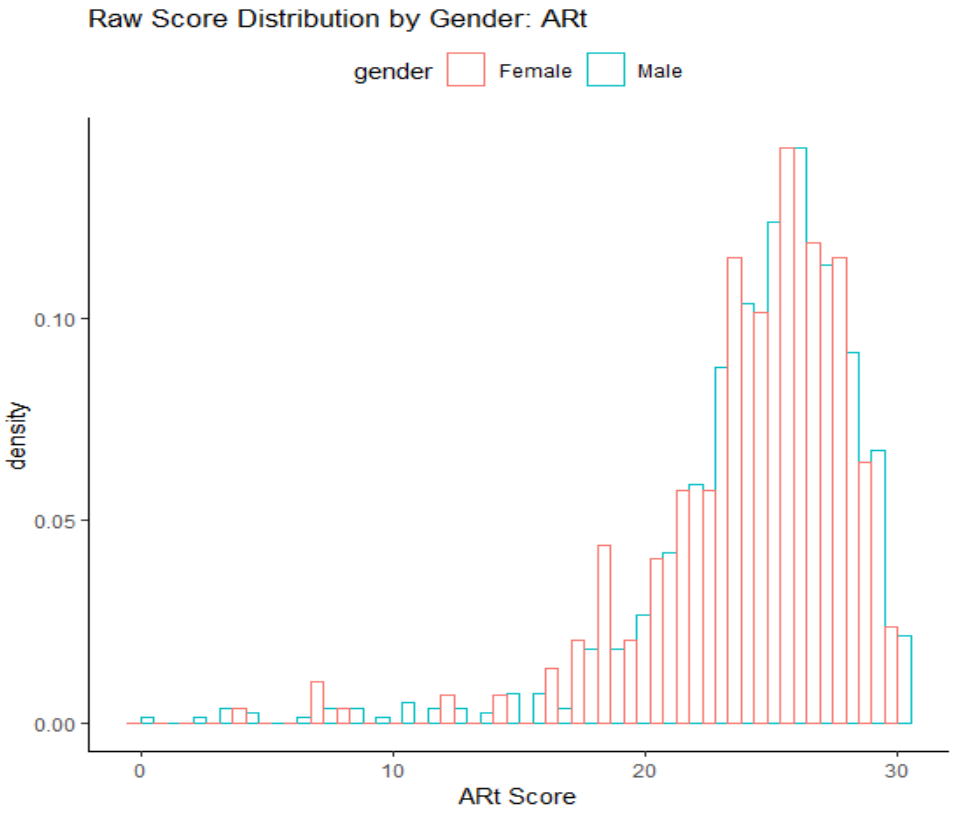
Gender	N	Min.	Median	Mean	SD	Max.	Reliability	Effect Size
Female	288	4	25	24.42	4.08	30	0.80	-0.03
Male[1]	804	0	25	24.28	4.30	30	0.82	

[1] Reference group

- **Effect size (ES):**

$$ES = \frac{u_R - u_F}{\sigma_p}$$

  - Where u_R and u_F are the means of the reference and focal groups
  - σ_p is the pooled standard deviation across the two groups
- **Category of ES:**
  - Small: >= 0.2 but < 0.5
  - Moderate: >= 0.5 but < 0.8
  - Large: >= 0.8

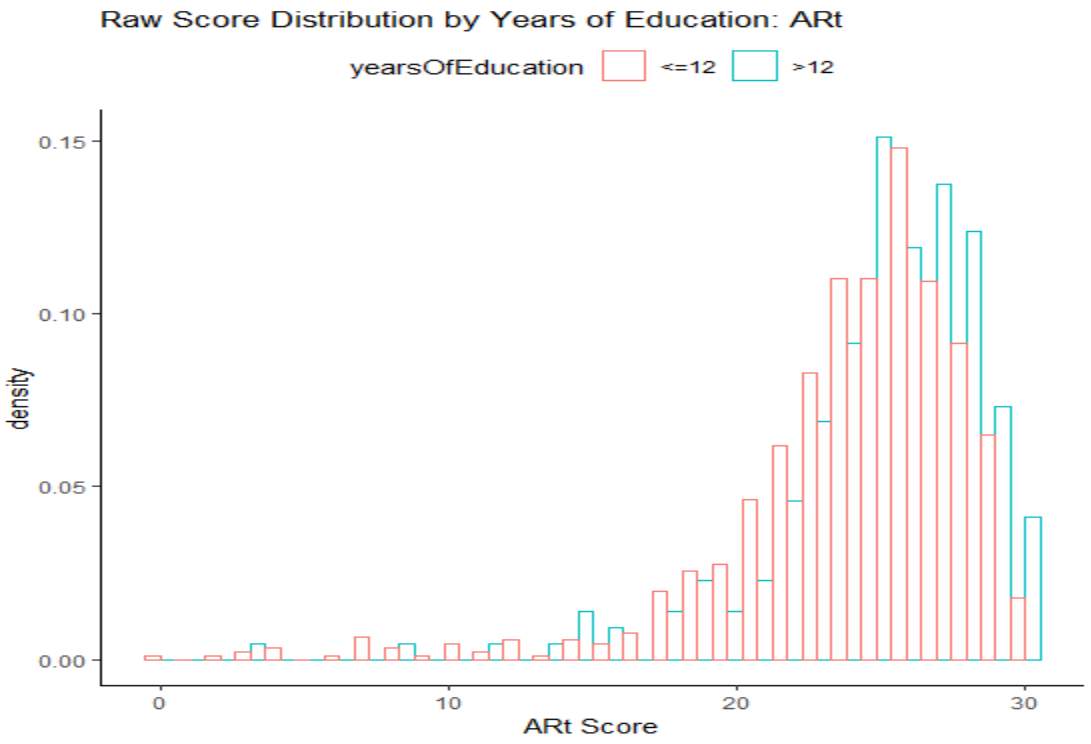


OPA
OFFICE OF PEOPLE ANALYTICS

RAW SCORE DISTRIBUTION—BY YEARS OF EDUCATION

Years of Education	N	Min.	Median	Mean	SD	Max.	Reliability	Effect Size
<= 12	881	0	25	24.16	4.32	30	0.79	0.19
> 12[1]	211	3	26	24.98	3.87	30	0.85	

[1] Reference group

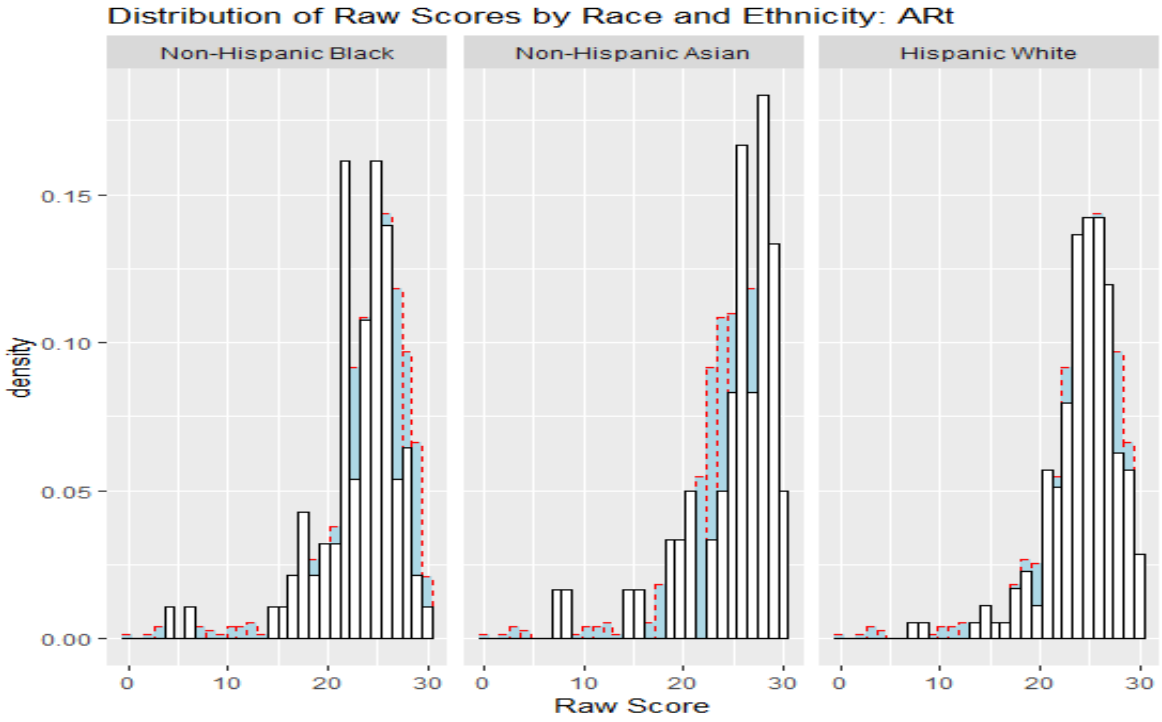


OPA
OFFICE OF PEOPLE ANALYTICS

SCORE DISTRIBUTION—BY RACE AND ETHNICITY: ART

Race and Ethnicity	N	Min.	Median	Mean	SD	Max.	Reliability	Effect Size
Non-Hispanic Black	90	4	24	23.30	4.17	30	0.79	0.23
Non-Hispanic Asian	58	7	26	25.07	4.83	30	0.88	-0.18
Hispanic White	170	7	25	24.45	3.63	30	0.75	-0.04
Non-Hispanic White[1]	687	0	25	24.29	4.35	30	0.83	

[1] Reference group

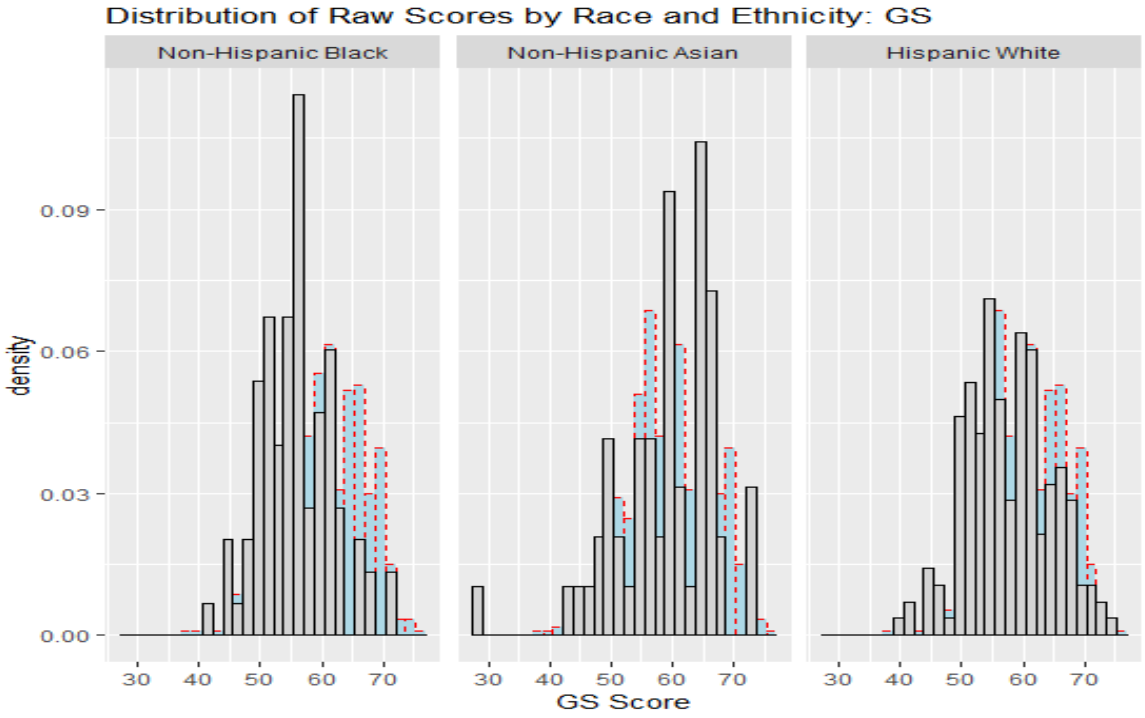


OPA
OFFICE OF PEOPLE ANALYTICS

SCORE DISTRIBUTION—BY RACE AND ETHNICITY: ASVAB GS

Race and Ethnicity	N	Min.	Median	Mean	SD	Max.	Effect Size
Non-Hispanic Black	90	41	56	55.97	6.04	72	0.63
Non-Hispanic Asian	58	28	60	59.09	8.22	73	0.15
Hispanic White	170	40	57.5	57.77	6.95	74	0.35
Non-Hispanic White[1]	687	38	60	60.08	6.59	76	

[1] Reference group

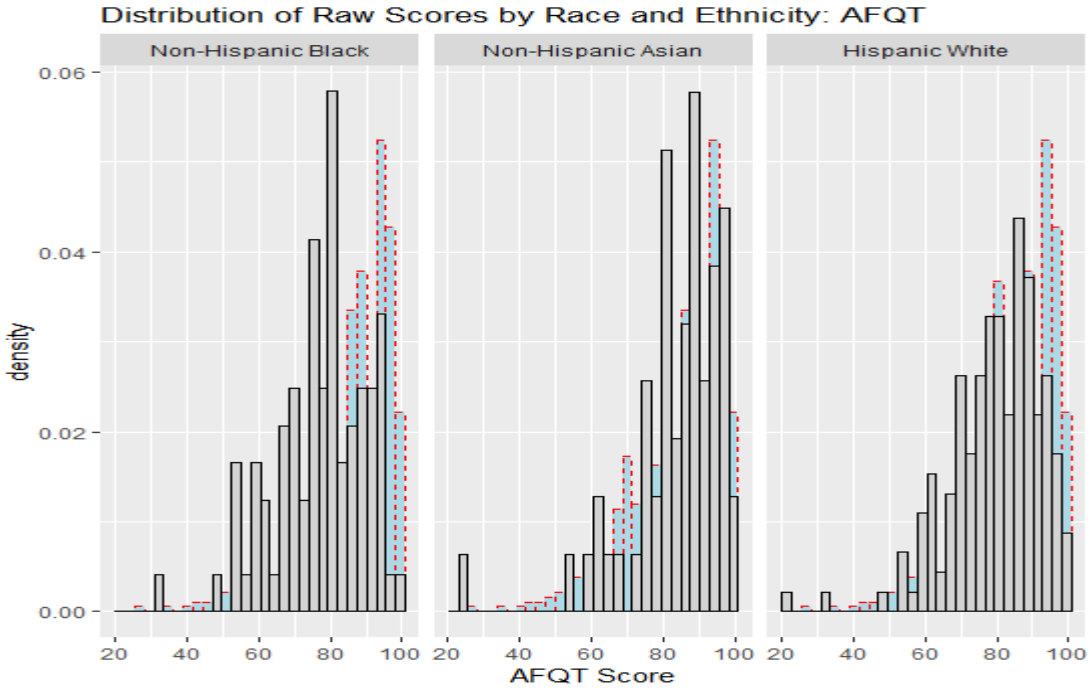


OPA
OFFICE OF PEOPLE ANALYTICS

SCORE DISTRIBUTION—BY RACE AND ETHNICITY: AFQT

Race and Ethnicity	N	Min.	Median	Mean	SD	Max.	Effect Size
Non-Hispanic Black	90	33	79.5	77.32	12.66	99	0.55
Non-Hispanic Asian	58	24	86.0	83.78	13.27	99	0.02
Hispanic White	170	21	82.0	79.85	12.59	99	0.34
Non-Hispanic White[1]	687	28	86.0	84.04	12.23	99	

[1] Reference group



OPA
OFFICE OF PEOPLE ANALYTICS

EFFECT SIZE OF MEAN DIFFERENCE: COMPARISON

- With 95% confidence interval:

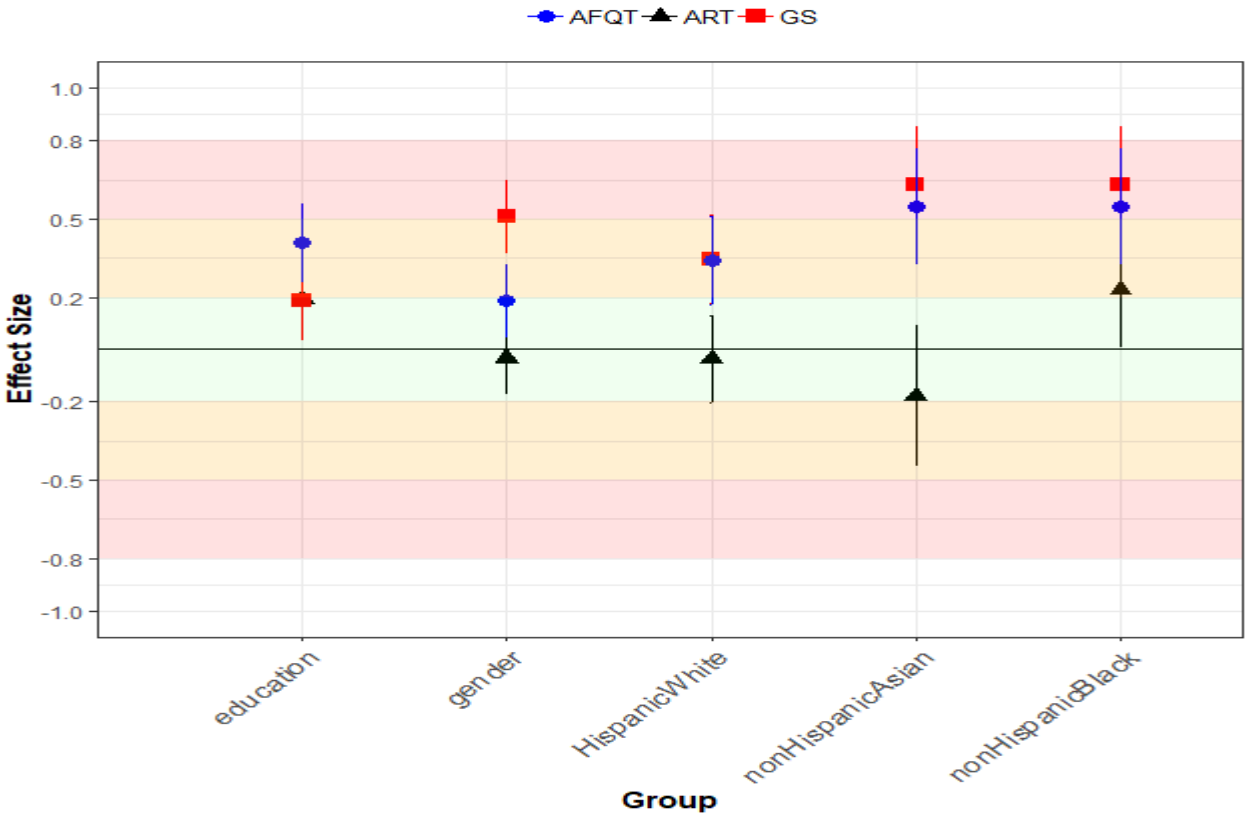
$$ES \pm 1.96 * \widehat{\sigma(ES)}$$

where

$$\widehat{\sigma(ES)}$$

$$= \sqrt{\frac{n_R+n_F}{n_R*n_F} + \frac{ES^2}{2*(n_R+n_F)}}$$

- AFQT, GS: similar pattern as seen in the adverse impact analysis with the general ASVAB population



OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation Along with ASVAB and MCt

Confirmatory Factor Analysis

OPA
OFFICE OF PEOPLE ANALYTICS

SUMMARY STATISTICS

- **ART**
  - Total case count: 2,162
  - Matched cases with both MCt and ASVAB: 1,724

	Test	N	Min.	Median	Mean	Max.	SD	Reliability[1]	Disattenuated Corr. w/ ART	Disattenuated Corr. w/ MCt
	ART	2162	0	25	24.09	30	4.16	0.80	--	--
	ART	1724	0	25	24.15	30	4.17			
	MCt	1724	0	21	19.99	32	6.38	0.86	0.52	
ASVAB Subtest	**GS**	1724	25	59	58.81	76	6.85	0.87	0.29	0.26
	AR	1724	37	60	59.41	72	6.06	0.91	0.43	0.51
	WK	1724	27	59	58.79	76	6.21	0.92	0.25	0.23
	PC	1724	33	59	59.00	69	5.36	0.86	0.31	0.28
	AS	1724	25	48	49.23	82	7.49	0.81	0.09	0.15
	MK	1724	40	60	59.97	73	5.60	0.91	0.37	0.35
	MC	1724	26	57	57.54	81	7.40	0.85	0.36	0.41
	EI	1724	26	55	55.88	83	7.92	0.88	0.20	0.24
	AO	1724	26	60.5	59.68	68	6.81	0.89	0.42	0.37
	AFQT	1724	21	85	82.76	99	12.61	0.97	0.46	0.45

[1] ART and MCt, Cronbach's α; ASVAB, from http://www.officialasvab.com/reliability_res.htm

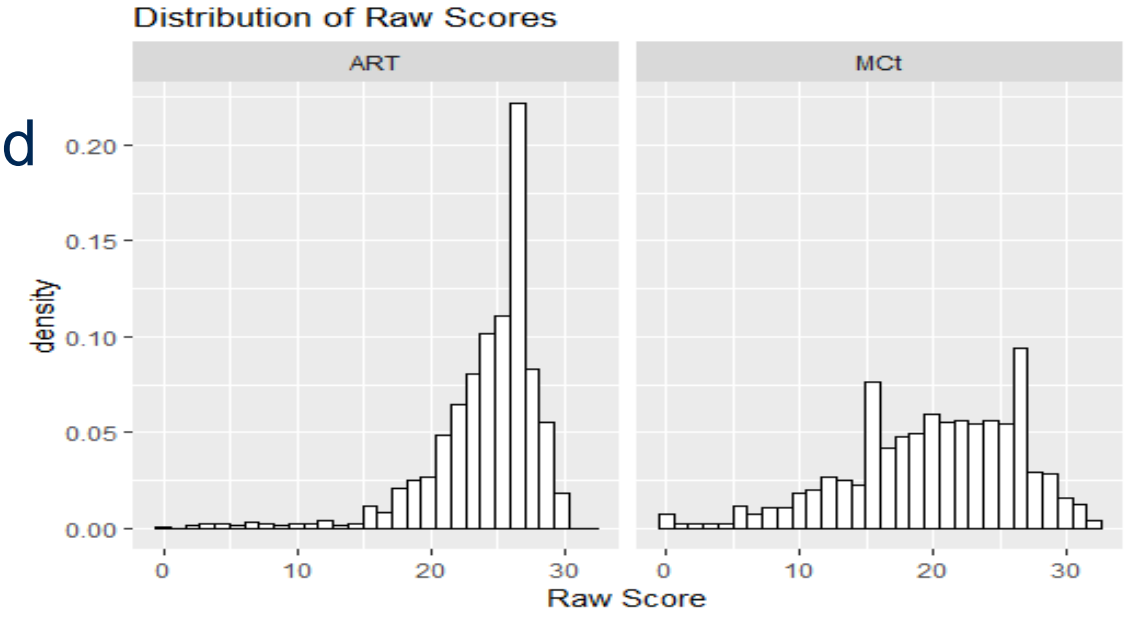
OPA
OFFICE OF PEOPLE ANALYTICS

SUMMARY STATISTICS—CONT'D

- Samples are a high ability group

	Group	N	Min	Mean	Max	SD
AFQT	**ART sample**	1,724	21	82.76	99	12.61
AFQT (2017 ETP)	**Female**	53,252	1	51.65	99	23.04
	Male	183,972	1	58.75	99	23.59

- Raw score distribution compared to that of MCt



OPA
OFFICE OF PEOPLE ANALYTICS

FACTOR ANALYSIS (FA)—CONFIRMATORY

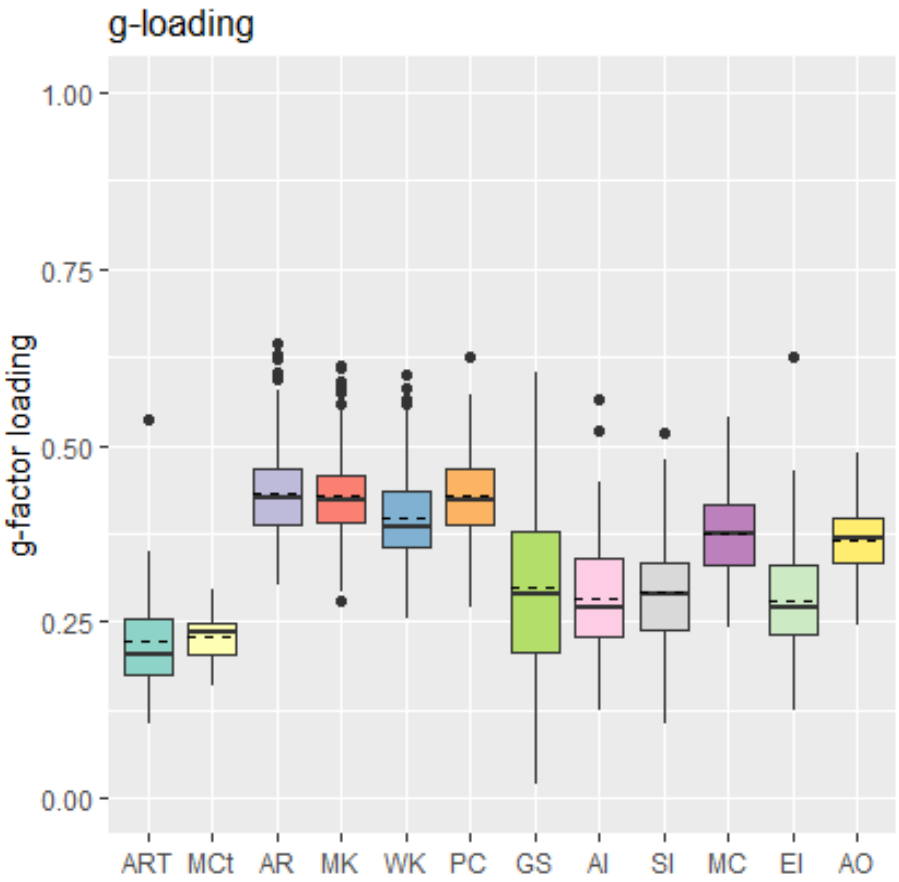
- Factor structure superimposed onto a subset of Broad (Stratum II) CHC ability definitions

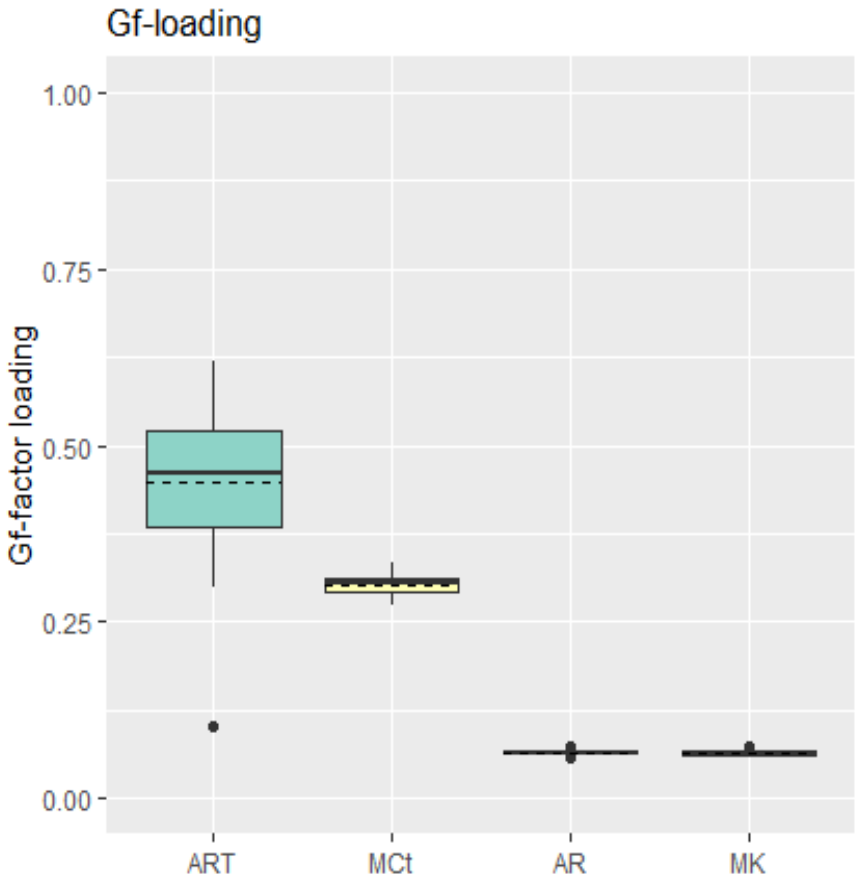
	Stratum II							Stratum I (Subtest Dimensions)									
	g	Gf	Gc	Gkn	Gv	Gq	Grw	GS	AR/MK	WK	PC	AS	MC	EI	AO	ART	MCt
GS	x		x	x				x									
AR/MK	x	x				x			x								
WK	x		x				x			x							
PC	x		x				x				x						
AS	x			x								x					
MC	x			x	x								x				
EI	x			x										x			
AO	x				x										x		
ART	x	x														X	
MCt	x	x															x

- Gf: Fluid intelligence/reasoning—the ability to solve new problems, use logic in new situations, and identify patterns
- Gc: Crystallized intelligence/knowledge—the ability to use learned knowledge and experience
- Gkn: General (domain-specific) knowledge
- Gv: Visual-spatial abilities
- Gq: Quantitative knowledge
- Grw: Reading/Writing

OPA
OFFICE OF PEOPLE ANALYTICS

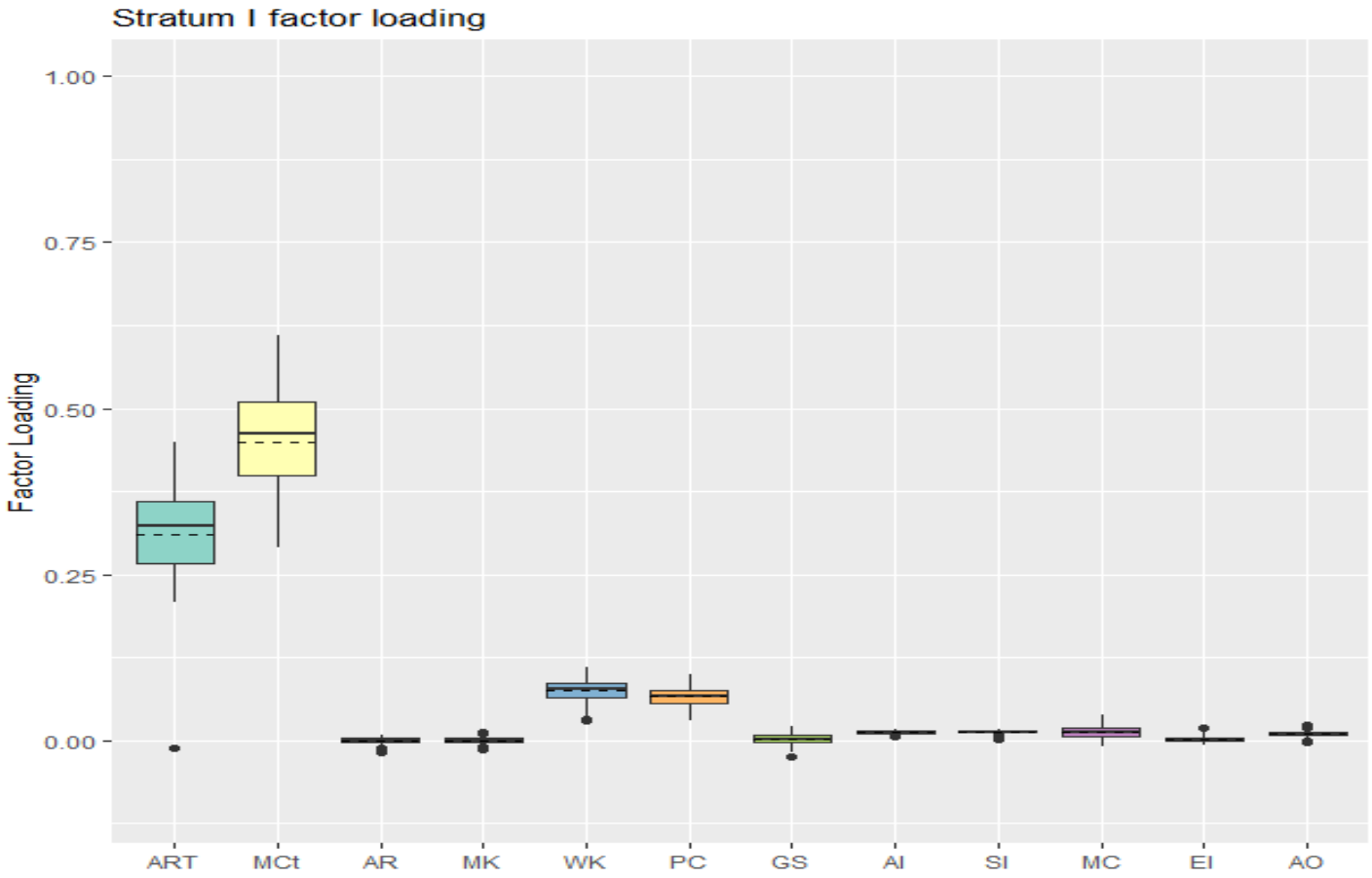
CONFIRMATORY FA: FACTOR LOADING





OPA
OFFICE OF PEOPLE ANALYTICS

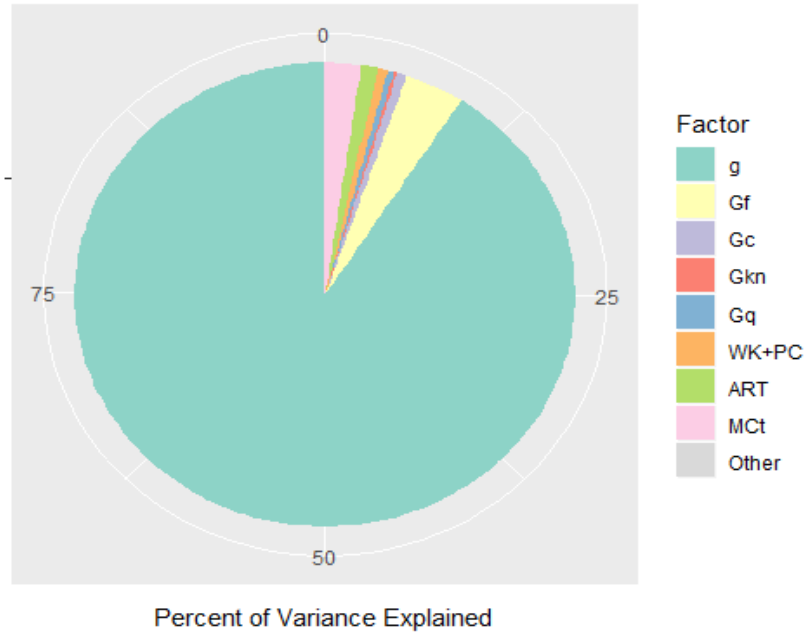
FACTOR LOADING ON STRATUM I ABILITIES



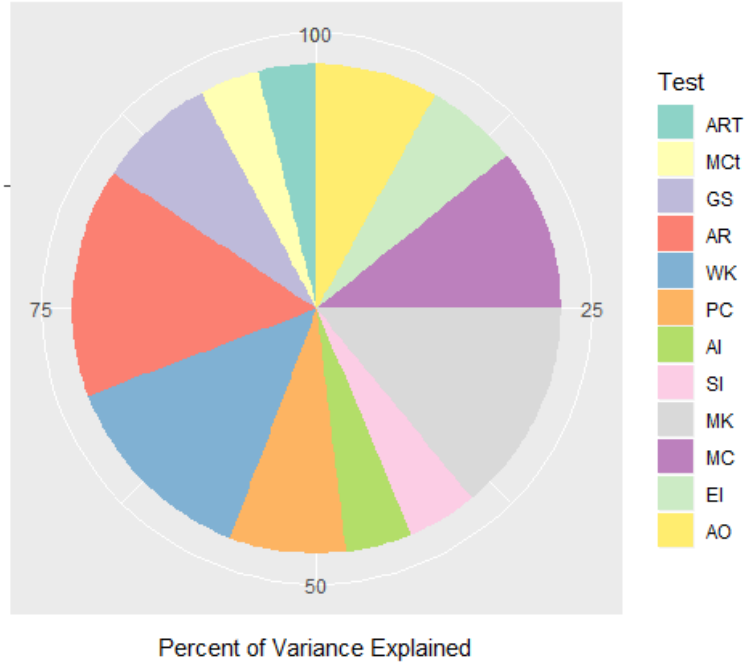
OPA
OFFICE OF PEOPLE ANALYTICS

VARIANCE EXPLAINED

- **By factor**
  - Factors with loading values < 0.1 are grouped into “Other”



- **By test**



OPA
OFFICE OF PEOPLE ANALYTICS

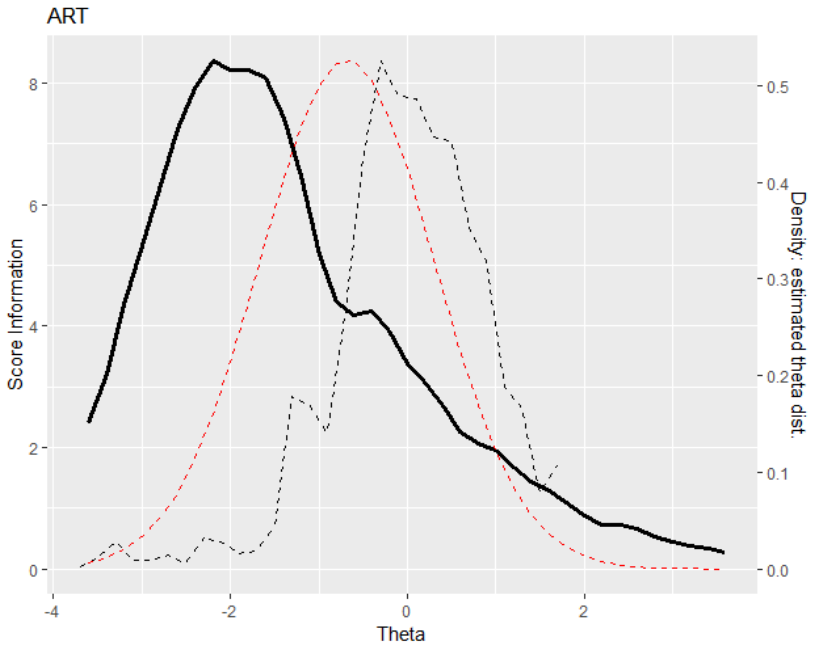
IRT Analysis

IRT ANALYSIS

- **3PL model fitted using BILOG-MG**
  - 25(83%) of the items showed adequate model fit (at a significant level 0.01)

Parameter	Min.	Median	Mean	Max.	SD
a	0.38	0.73	0.75	1.27	0.21
b	-3.52	-1.80	-1.53	1.19	1.17
c	0.06	0.10	0.11	0.19	0.03

- **Estimated test-retest reliability = 0.77**
  - Lower bound due to restricted sample
- **Score information function**
  - Dark dashed line: θ distribution of restricted sample (from Bilog-MG θ estimates)
  - Red dashed line: estimated un-restricted θ distribution[1]



OPA
OFFICE OF PEOPLE ANALYTICS

[1]Lord & Novick (1968). *Statistical Theories of Mental Test Scores.* used estimated AR θ values in the estimation.

Response Time

OPA
OFFICE OF PEOPLE ANALYTICS

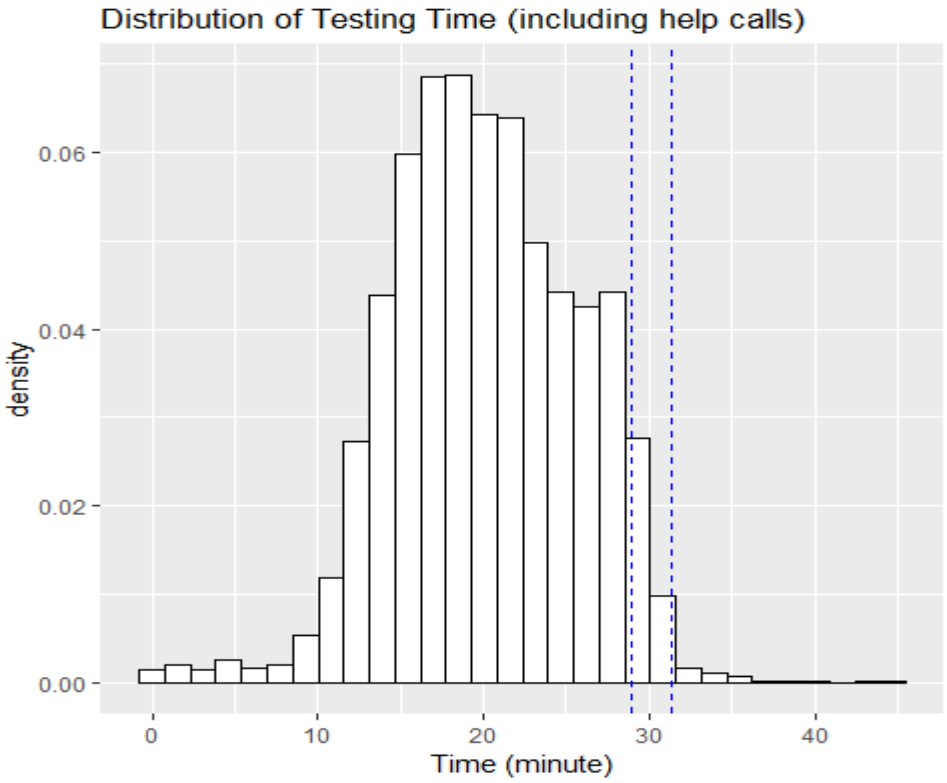
RESPONSE TIME ANALYSIS

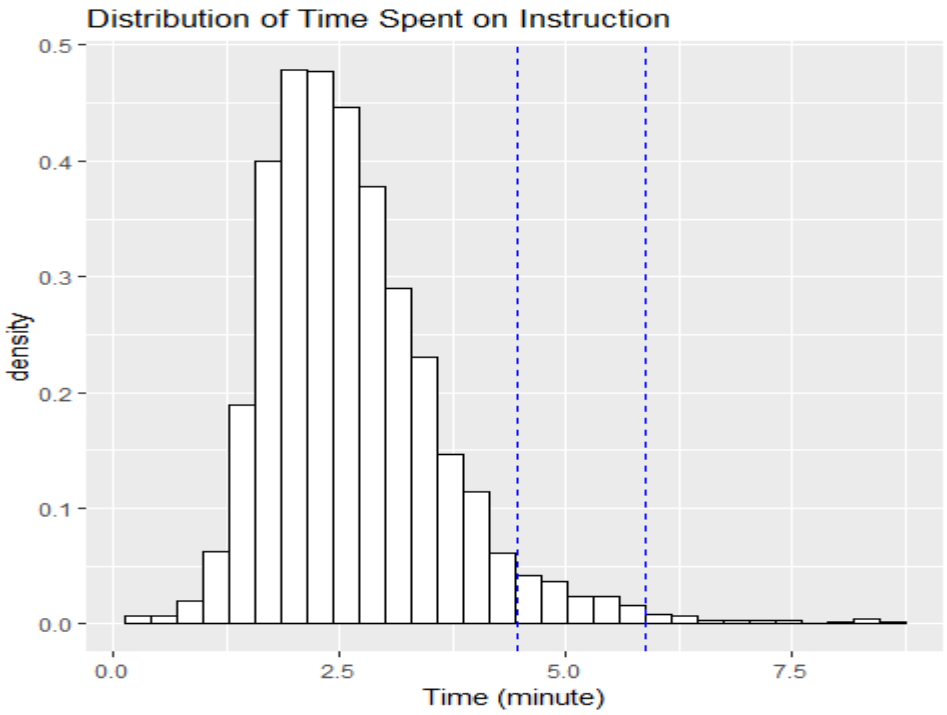
- **Item completion rate:**
  - Last item: 98.2%; second to last item: 99.3%
- **Total testing time (sum of item response times)**
  - Dashed blue lines: the 95th and 99th percentiles of the fitted (green) distribution
  - Q99 = 28

OPA
OFFICE OF PEOPLE ANALYTICS

RESPONSE TIME ANALYSIS

- Testing time—including help calls, time spent on instruction





OPA
OFFICE OF PEOPLE ANALYTICS

CONCLUSION AND DISCUSSION

- Evaluation was done on limited and restricted data
  - Test-takers are a high-ability group
- Items appear easy to the samples administered
- Reliability findings likely reflect lower bound due to restricted sample
- The test appears to measure a unique domain not currently represented in the ASVAB
- Test results appear to have small impact across demographic groups
- The 25-minute test time limit may not be adequate
- Results are promising, but requires further study with more representative sample

RECOMMENDATIONS

- Further evaluation with a bigger and representative, unrestricted sample
- Increase the test time limit to 30 minutes; will evaluate again with bigger and more representative sample
- Develop a similar test (complex reasoning)
  - Develop harder items if evaluation of ART with an unrestricted sample confirms the test is too easy
- Investigate the feasibility of using AIG to develop the items
- Develop a CAT version of the test

Thank You!

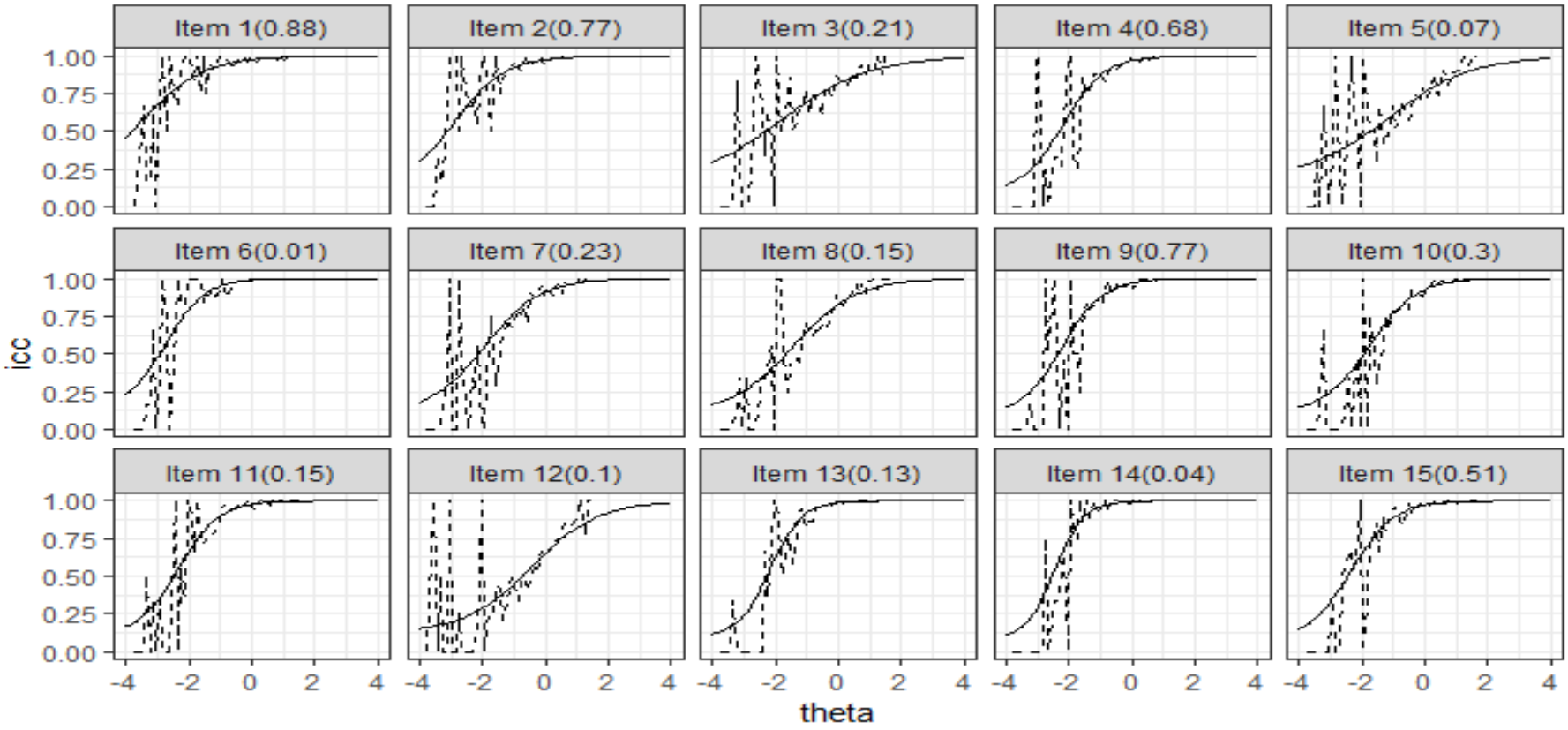
OPA
OFFICE OF PEOPLE ANALYTICS

OPA

OFFICE OF PEOPLE ANALYTICS

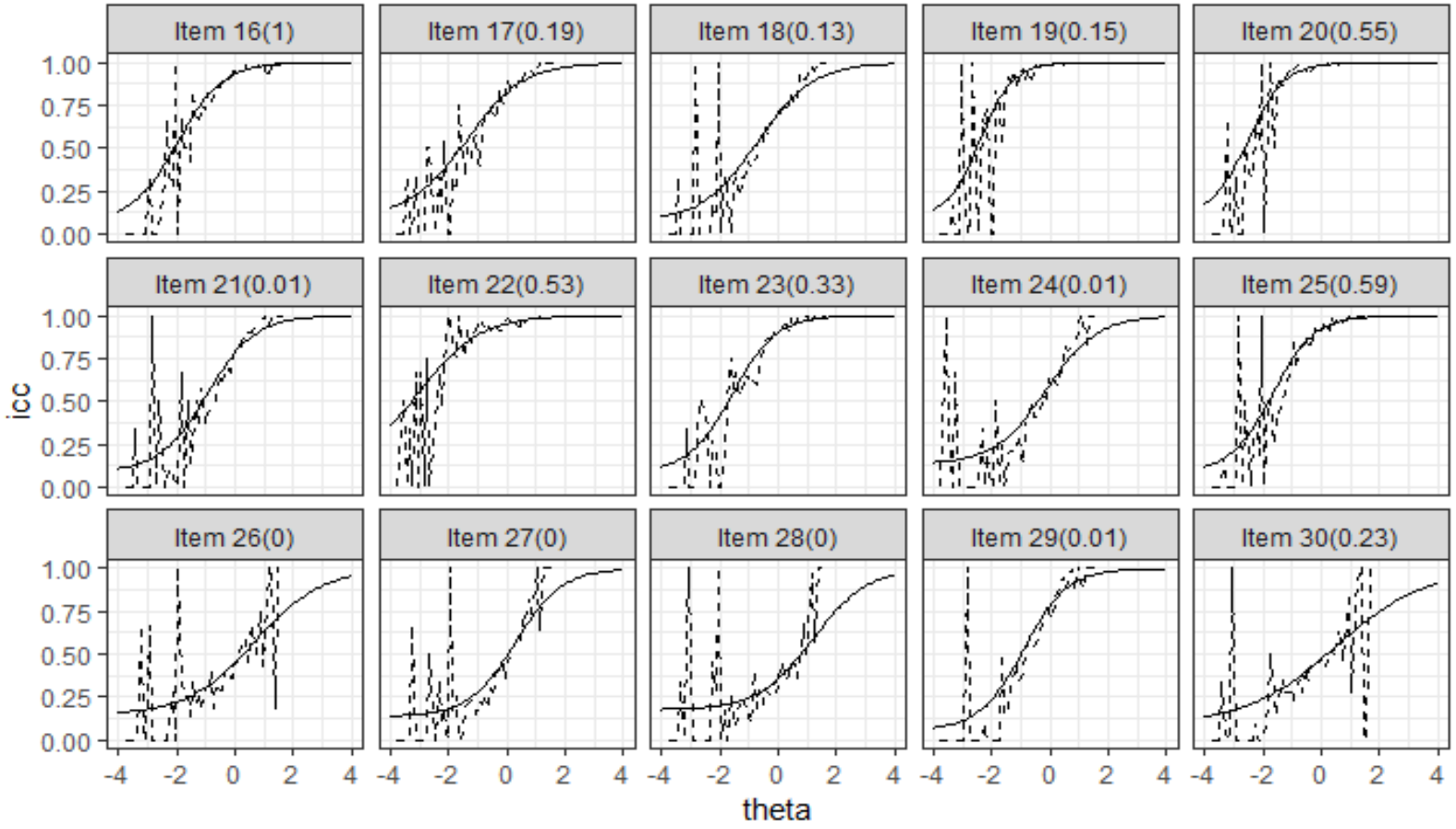
Backup Slides

IRT: FITTED AND OBSERVED ICC



OPA
OFFICE OF PEOPLE ANALYTICS

IRT: FITTED AND OBSERVED ICC



OPA
OFFICE OF PEOPLE ANALYTICS

Tab G

OPA

OFFICE OF PEOPLE ANALYTICS

Defense Personnel Assessment Center (DPAC) Cloud 101

Presented by Matthew Ellis (Northrop Grumman / HumRRO)

Overview

- Purpose
- Definition of Cloud Computing
- Department of Defense (DoD) Strategic Objectives
- Cloud Security
- DPAC Cloud Initiatives:
  - The current work being performed
  - The planned work for the future
- Risks, Challenges & Opportunities

Cloud Overview

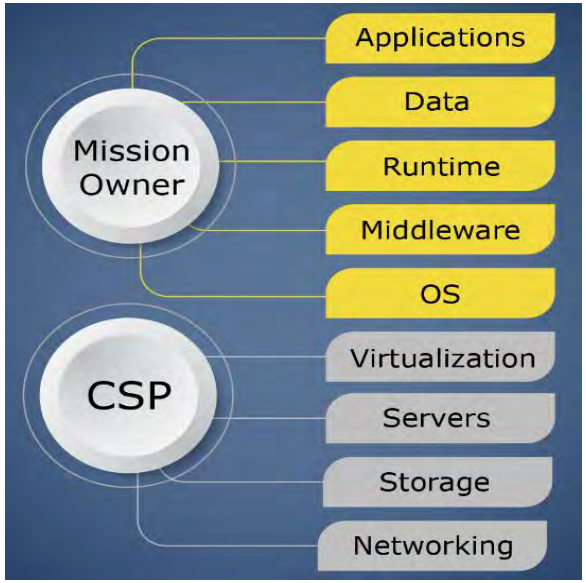
What is cloud computing?

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics and defines three service models and four deployment models.

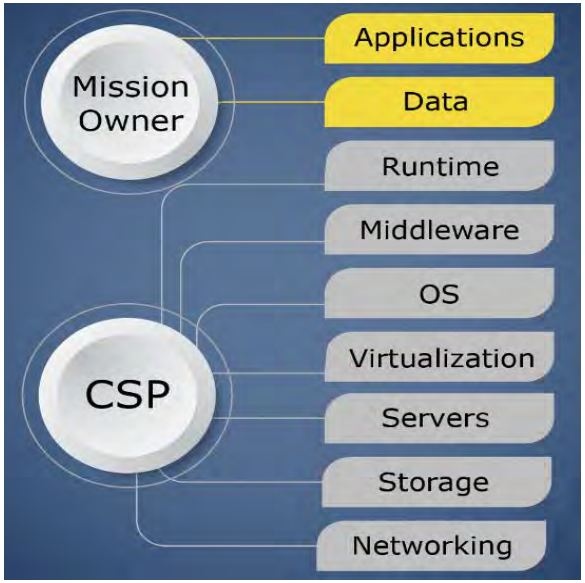
Essential characteristics

CLOUD COMPUTING SERVICE MODELS

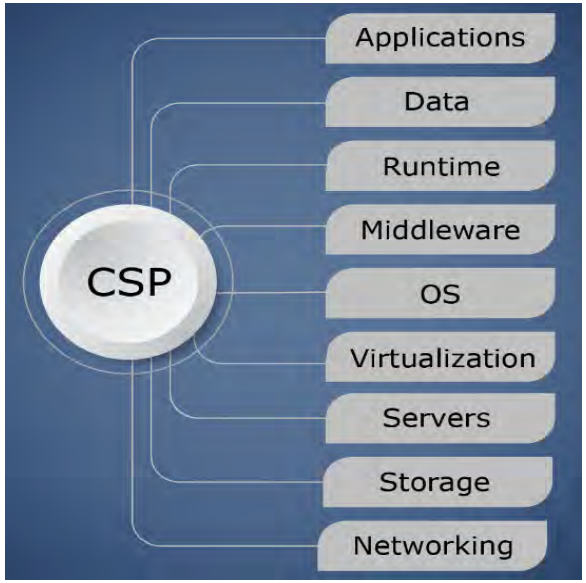
Infrastructure as a Service (IAAS)



Platform as a Service (PAAS)



Software as a Service (SAAS)



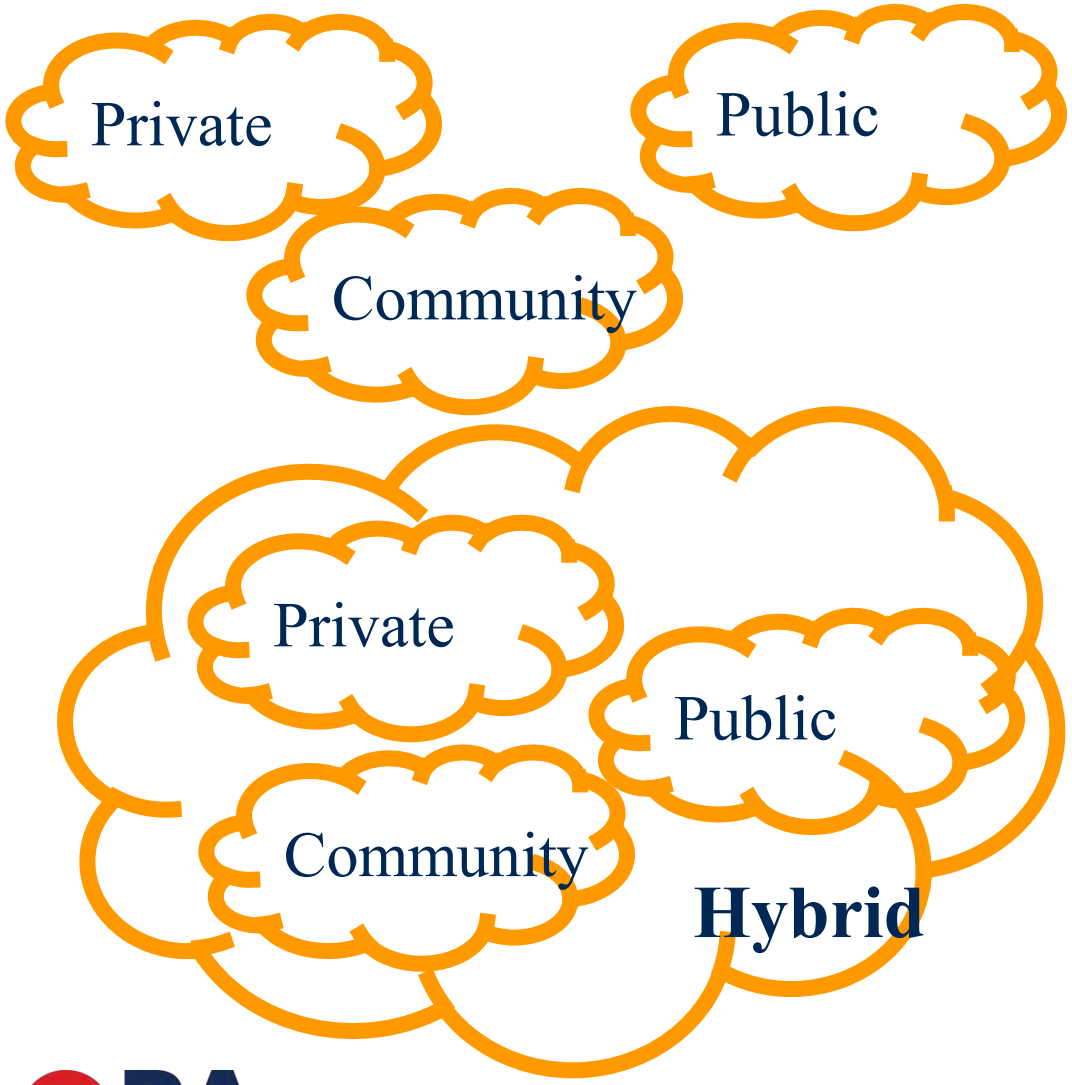
Ex: DPAC

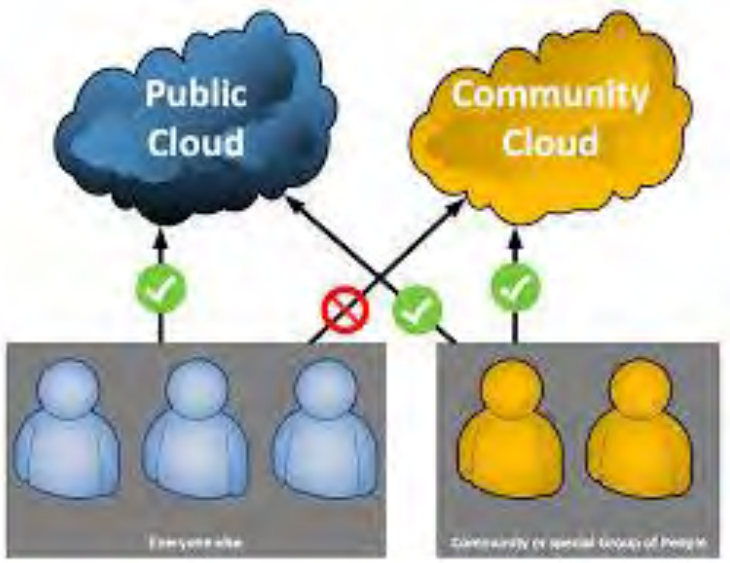
Mission Owner = DPAC

Cloud Service Provider (CSP) = Amazon Web Services (AWS) GovCloud

OPA
OFFICE OF PEOPLE ANALYTICS

Cloud Deployment Models



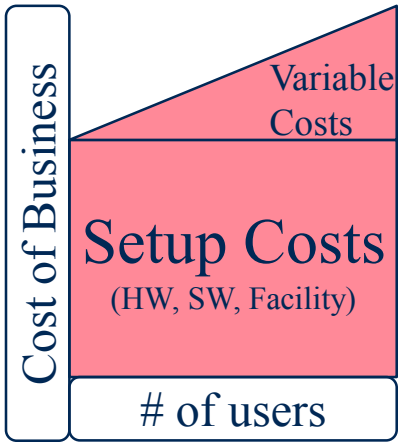


OPA
OFFICE OF PEOPLE ANALYTICS

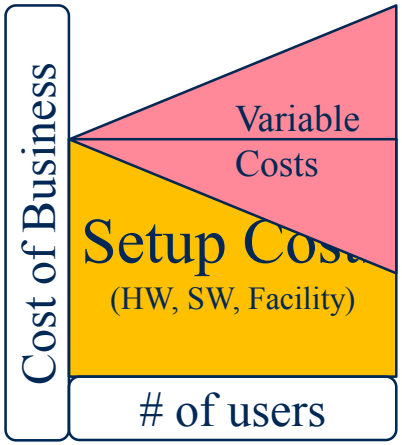
Cloud Benefits

EFFICIENCY	
Cloud Benefits	**Current Environment**
• Improved asset utilization (server utilization > 60-70%) • Aggregated demand and accelerated system consolidation (e.g., Federal Data Center Consolidation Initiative) • Improved productivity in application development, application management, network, and end-user	• Low asset utilization (server utilization < 30% typical) • Fragmented demand and duplicative systems • Difficult-to-manage systems
AGILITY	
Cloud Benefits	**Current Environment**
• Purchase "as-a-service" from trusted cloud providers • Near-instantaneous increases and reductions in capacity • More responsive to urgent agency needs	• Years required to build data centers for new services • Months required to increase capacity of existing services
INNOVATION	
Cloud Benefits	**Current Environment**
• Shift focus from asset ownership to service management • Tap into private sector innovation • Encourages entrepreneurial culture • Better linked to emerging technologies (e.g., devices)	• Burdened by asset management • De-coupled from private sector innovation engines • Risk-adverse culture

Datacenter Compute Costs



Cloud Compute Costs



OPA
OFFICE OF PEOPLE ANALYTICS

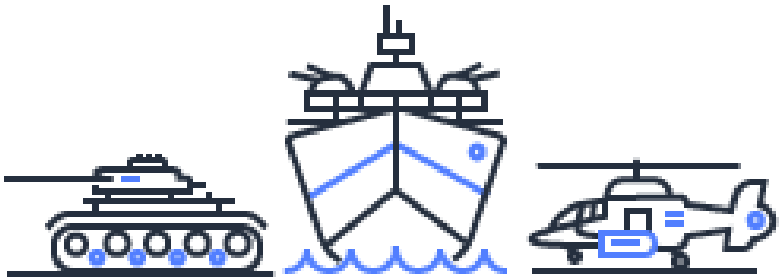
Ref: Cloud benefits: Efficiency, Agility, Innovation –Federal Cloud Computing Strategy

DoD IT Systems & Cloud

DoD Cloud Strategy

"If we fail to adapt ... at the speed of relevance, then our military forces ... will lose the very technical and tactical advantages we've enjoyed since World War II " – (former) Secretary of Defense James N Mattis

DoD requires an extensible and secure cloud environment that spans the homeland to the global tactical edge, as well as the ability to rapidly access computing and storage capacity to address warfighting challenges at the speed of relevance. **DoD Cloud Strategy**



Guiding Principals

- Mission First
- Cloud Smart-Data Smart
- Leveraging Commercial Industry Best Practices
- Creating a culture better suited for Modern Technology Evolution

OPA
OFFICE OF PEOPLE ANALYTICS

INFORMATION TECHNOLOGY SYSTEMS IN DOD

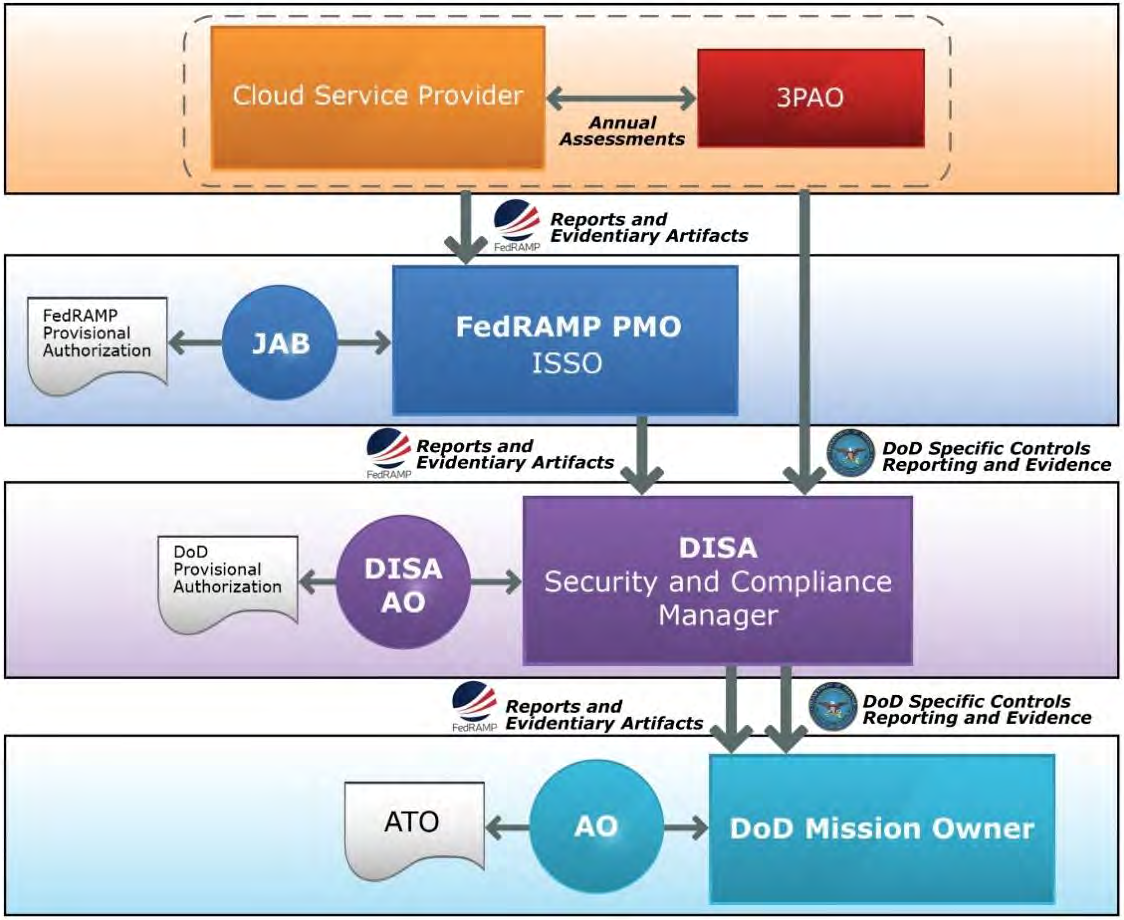
- **Every IT System exposes the DoD to some element of risk**
  - Information loss, Illicit Entry to DoD Networks, PII/PHI breach
- **IT Systems require approval to operate**
  - Agency's Authorizing Official (AO) signs a memo - Authority to Operate (ATO)
- **Use guidance from the Risk Management Framework (RMF) from National Institute of Standards & Technology (NIST) to assess**

AO must be cognizant of the risks introduced by IT Systems

- **Cloud introduces additional risks to the IT system that must be managed**
- **Federal Risk and Authorization Management Program (FedRAMP) and Defense Information Security Agency (DISA) have established standards for the Cloud computing in the DoD**

OPA
OFFICE OF PEOPLE ANALYTICS

CLOUD SOLUTION PROVIDER (CSP) SELECTION



Approvals Required:

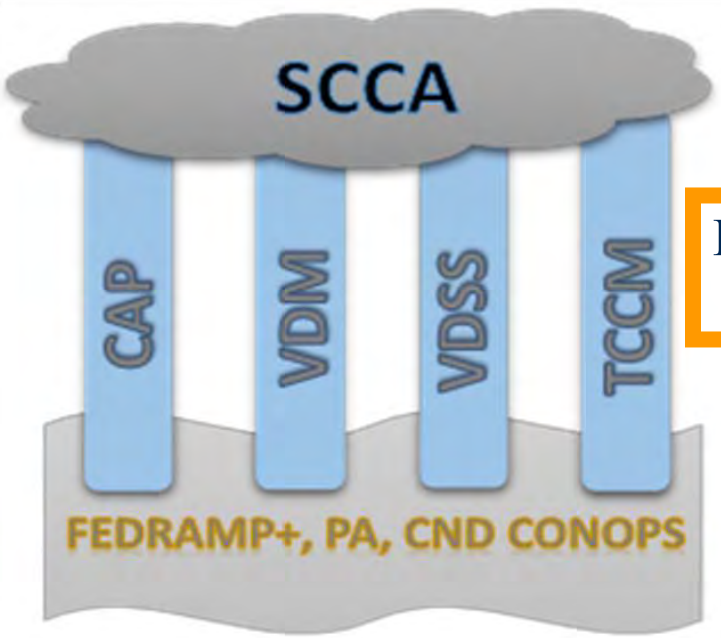
aws

FedRAMP

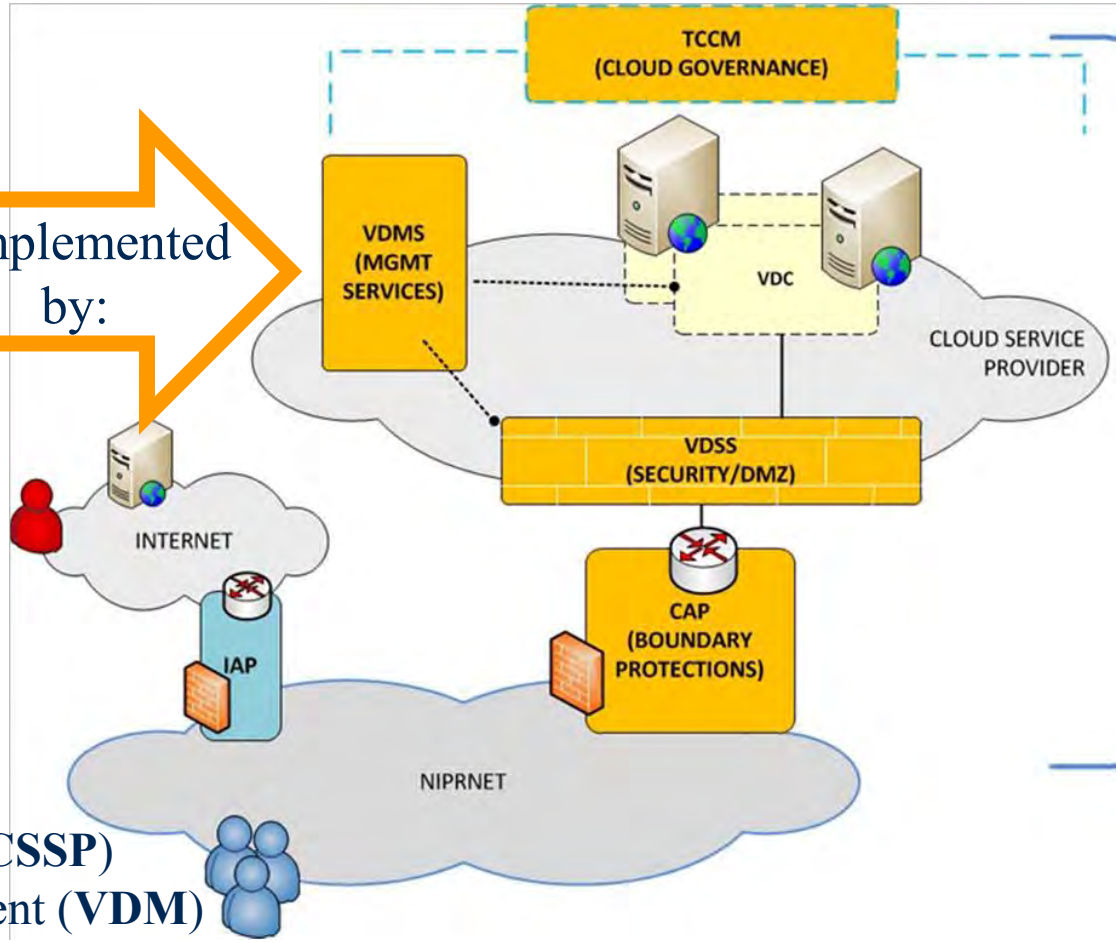
DISA

OPA
OFFICE OF PEOPLE ANALYTICS

SECURE CLOUD COMPUTING ARCHITECTURE (SCCA)



Implemented by:



Pillars of Security

- Cloud Access Point (**CAP**)
- Cyber Security Services Provider (**CSSP**)
  - Virtual Data Center Management (**VDM**)
  - Virtual Data Center Security Stack (**VDSS**)
- Trusted Cloud Credential Manager (**TCCM**)

OPA
OFFICE OF PEOPLE ANALYTICS

DPAC Cloud

DPAC - BUSINESS CASE ANALYSIS

- **Established the "why" – improve reliability & availability / increase scalability**
- **Established scope – ASVAB and Language suites of applications**
- **Established Impact Level / Information security requirements – IL4**
- **Trade study of Cloud Service Providers**

Overall Comparison of Alternatives and "As Is"	Financial				Non-Financial			Best Option
	Economic Viability (Strong, Mod, Weak, Not Viable)	Cost (FY17-22) Millions	Unfunded (FY17-22)	Savings (FY17-22)	Requirements (Exceeds, Meets, Not acceptable)	Operational Benefits (Significant, Moderate, Low, None)	Mitigated Risk (Low, Med, High, Catastrophic)	
"As- Is"	N/A			N/A			N/A	
AWS GovCloud	Strong				Exceeds	Significant	Low	←
Microsoft Azure	Strong				Meets	Moderate	Low	
IBM SoftLayer	Strong				Meets	Moderate	High	
DISA MilCloud	Not Viable				Meets	Low	High	

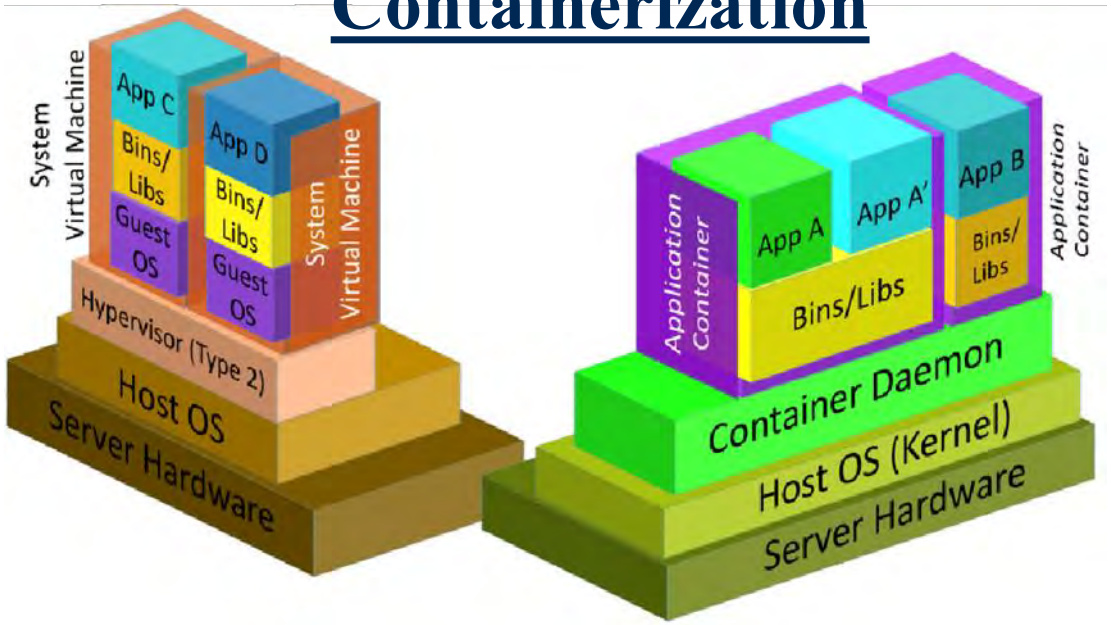
- **Key Leadership Support**
  - Concurrence with migration of a pilot set of applications to Amazon Web Services.
  - Concurrence that IL 4 Sensitive data protection requirements can be successfully satisfied through Amazon's ATO, DISA's SCCA and DPAC's shared IA responsibility.

DPAC APPLICATION MIGRATION STRATEGY

Lift & Shift

Database

Containerization

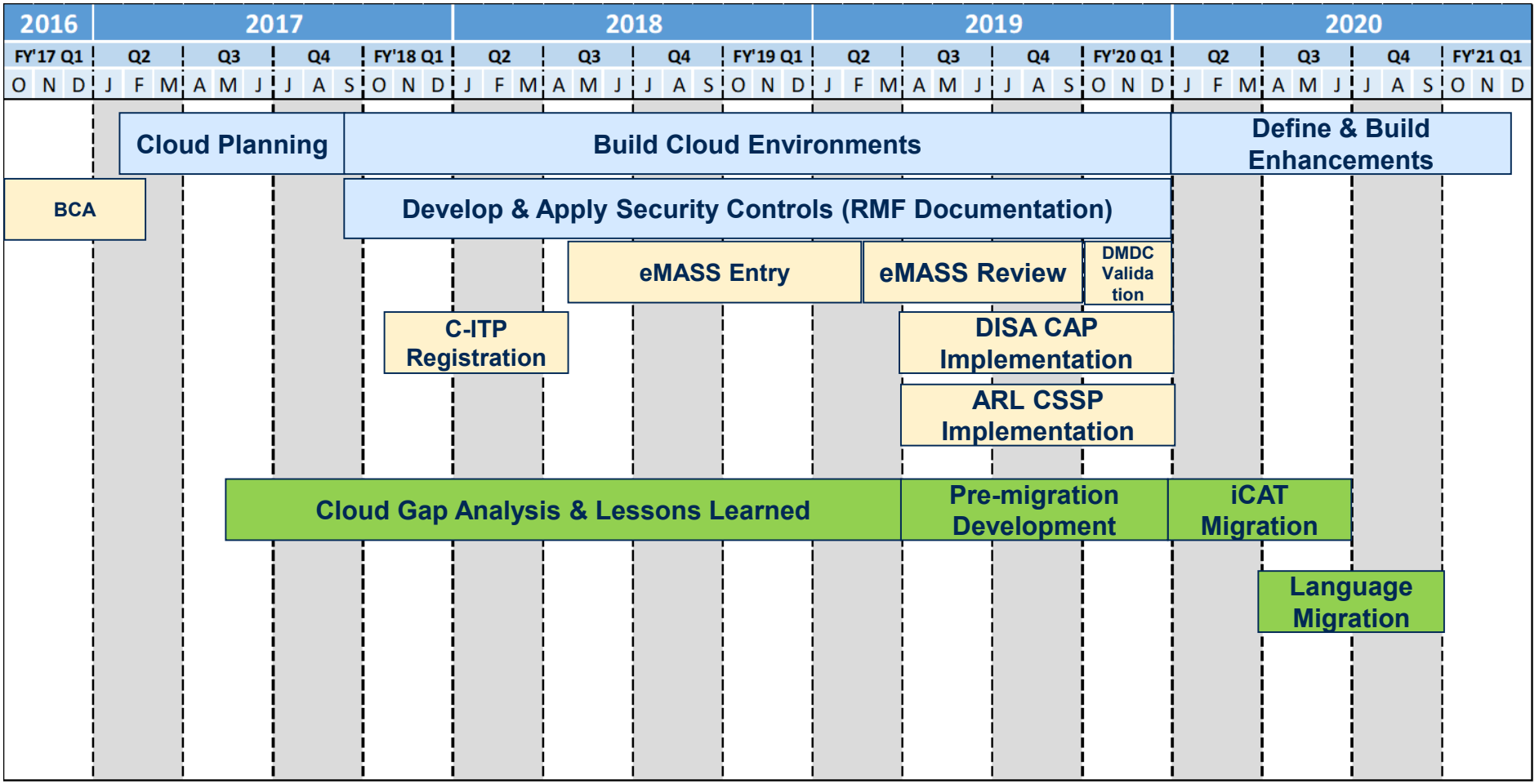


Goals

- Expedite Cloud Migration
- Reduce Licensing Costs
- Reduce System Administration
- Improve Reliability / Delivery

OPA
OFFICE OF PEOPLE ANALYTICS

DPAC CLOUD ROADMAP



OPA
OFFICE OF PEOPLE ANALYTICS

CHALLENGES & OPPORTUNITIES

Challenges:

- **iCAT and Language application migration to the cloud "freezes" new feature development for a period of time**
- **Changes to Help Desk via DMDC Support Center (DSC) and DPAC with cloud-hosted applications**

Opportunities:

- **Improve System Availability & Reliability – reducing the MEPS Travel impacts of system outages**

Questions?

Backup

ACRONYMS

3PAO	Third Party Assessment Organization	IAP	Internet Access Point
AO	Authorizing Official	IL	Impact Level
ARL	Army Research Labs	ISSO	Information System Security Officer
ATO	Authority to Operate	IT	Information Technology
AWS	Amazon Web Services	JAB	Joint Advisory Board
BCA	Business Case Analysis	NIST	National Institute of Standards & Technolo
CAP	Cloud Access Point	OS	Operating System
C-ITP	Cloud Information Technology Project	PA	Provisional Authority
CND	Computer Network Defense	PaaS	Platform as a Service
CSP	Cloud Service Provider	PHI	Protected Health Information
CSSP	Cybersecurity Services Provider	PII	Personably Identifiable Information
DISA	Defense Information Security Agency	RMF	Risk Management Framework
DMDC	Defense Manpower Datacenter	SaaS	Software as a Service
DoD	Department of Defense	SCCA	Secure Cloud Computing Architecture
DPAC	Defense Personnel Assessment Center	TCCM	Trusted Cloud Credential Manager
eMASS	Enterprise Mission Assurnace Support Services	VDC	Virtual Datacenter
FedRAMP	Federal Risk and Authorization Management Program	VDM	Virtual Datacenter Management
IaaS	Infrastructure as a Service	VDSS	Virtual Datacenter Security Stack

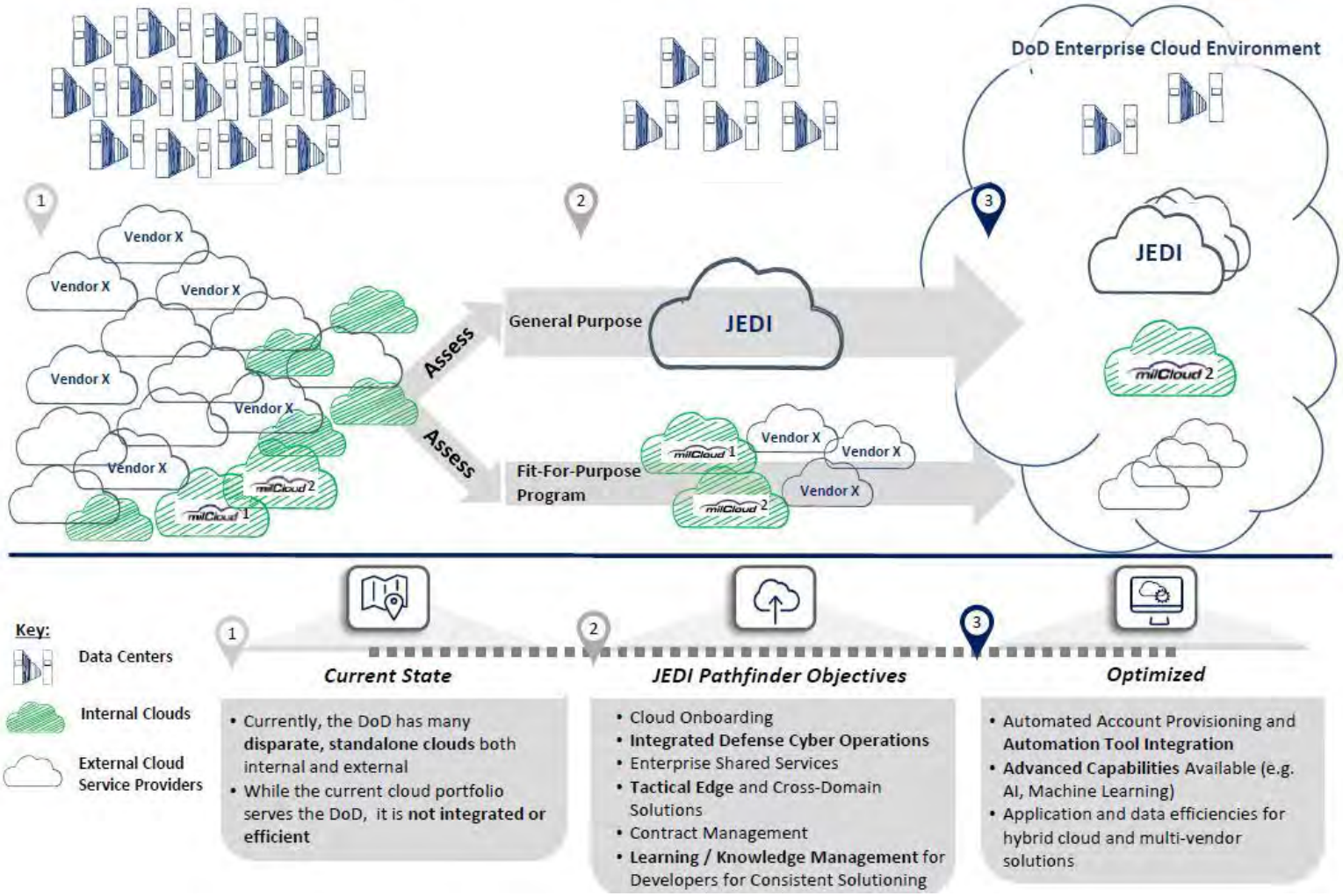
DOD CLOUD ACQUISITION GUIDANCE

- ❑ Conduct Information Technology (IT) Business Case Analysis
- ❑ Apply DoD Cloud Computing Security Requirements Guide
- ❑ Use Commercial Cloud Services that have a DoD Provisional Authorization (PA) and Obtain a Component Authority to Operate (ATO)
- ❑ Use an Approved DoD Boundary Cloud Access Point (BCAP) and Cybersecurity Service Provider (CSSP) to Protect Sensitive Data
- ❑ Apply the Defense Federal Acquisition Regulation Supplement Rule to Commercial Cloud Contracts
- ❑ Apply DoD Secure Cloud Computing Architecture (SCCA) and DISA's Secure Cloud Computing guidance

Unclassified

Department of Defense Cloud Strategy

DoD Pathfinder to hybrid cloud environments and multiple vendors



CYBERSECURITY SERVICE PROVIDER CAPABILITY

DISA

Available CSSP Services Summary

CSSP Offerings	Traditional CSSP	milCloud	milCloud+	Commercial Cloud (Initial)	Commercial Cloud (Basic, IaaS only)	Commercial Cloud (SCCA, IaaS only)
Availability:	Now	June, 2017	June, 2017	Now	FY 18	TBD
CSSP Subscription Services						
Malware Notification Protection (MNP)	☑	☑	☑		☑	☑
Support and Training (S&T)	☑	☑	☑	☑	☑	☑
INFOCON/CPCON	☑	☑	☑	☑	☑	☑
Information Assurance Vulnerability Management (IAVM)	☑	☑	☑	☑	☑	☑
Attack Sensing and Warning (ASW)	☑	☑	☑		☑	☑
Warning Intelligence (WI)	☑	☑	☑		☑	☑
Incident Reporting (IR)	☑	☑	☑	☑	☑	☑
Incident Handling Response (IHR)	☑	☑	☑			☑
Forensic Media Analysis (FMA)	☑	☑	☑			☑
Reverse Engineering/Malware Analysis (RE/MA)	☑	☑	☑	☑	☑	☑
Volatile Data Analysis (VDA)	☑	☑	☑			☑
Network Security Monitoring (NSM) Service	☑	☑	☑	CAP Only	☑	☑
Vulnerability Analysis & Assessment Support Services						
External Vulnerability Scans (EVS)	☑	☑	☑ (Optional)		☑	☑
Web Vulnerability Scans (WVS)	☑	☑	☑	☑ (Optional)	☑	☑
Penetration Testing (Pen Test)	☑ (Optional)	☑ (Optional)	☑ (Optional)		☑ (Optional)	☑ (Optional)
Red Team Operations (RTO)	☑ (Optional)	☑ (Optional)	☑ (Optional)		☑ (Optional)	☑ (Optional)
Intrusion Assessment	☑ (Optional)	☑ (Optional)	☑ (Optional)		☑ (Optional)	☑ (Optional)
Sensor Sustainment Services						
Sensor Sustainment & Configuration Management	☑	☑ (one-time fee)	☑ (one-time fee)	☑ (one-time fee)	☑	☑

IMPACT LEVEL ASSESSMENT

IMPACT LEVEL	INFORMATION SENSITIVITY	SECURITY CONTROLS	LOCATION	OFF-PREMISES CONNECTIVITY	SEPARATION	PERSONNEL REQUIREMENTS
2	PUBLIC or Non-critical Mission Information	FedRAMP v2 Moderate	US / US outlying areas or DoD on-premises	Internet	Virtual / Logical PUBLIC COMMUNITY	National Agency Check and Inquiries (NACI)
4	CUI or Non-CUI Non-Critical Mission Information Non-National Security Systems	Level 2 + CUI-Specific Tailored Set	US / US outlying areas or DoD on-premises	NIPRNet via CAP	Virtual / Logical Limited "Public" Community Strong Virtual Separation Between Tenant Systems & Information	US Persons ADP-1 Single Scope Background Investigation (SSBI)
5	Higher Sensitivity CUI Mission Critical Information National Security Systems	Level 4 + NSS & CUI-Specific Tailored Set	US / US outlying areas or DoD on-premises	NIPRNet via CAP	Virtual / Logical FEDERAL GOV. COMMUNITY Dedicated Multi-Tenant Infrastructure Physically Separate from Non-Federal Systems Strong Virtual Separation Between Tenant Systems & Information	ADP-2 National Agency Check with Law and Credit (NACLC) Non-Disclosure Agreement (NDA)
6	Classified SECRET National Security Systems	Level 5 + Classified Overlay	US / US outlying areas or DoD on-premises CLEARED / CLASSIFIED FACILITIES	SIPRNET DIRECT With DoD SIPRNet Enclave Connection Approval	Virtual / Logical FEDERAL GOV. COMMUNITY Dedicated Multi-Tenant Infrastructure Physically Separate from Non-Federal and Unclassified Systems Strong Virtual Separation Between Tenant Systems & Information	US Citizens w/ Favorably Adjudicated SSBI & SECRET Clearance NDA

Tab H

Use of Social Media in Military Recruitment and Selection

Status Update

Presented to: Defense Advisory Committee on Military Personnel Testing (DACMPT)

September 26, 2019

Presenters: Tim McGonigle, HumRRO

Headquarters: 66 Canal Center Plaza, Suite 700, Alexandria, VA 22314-1578 | Phone: 703.549.3611 | www.humrro.org

Agenda

- Overview and Goals of Project
- Project Team
- Summary and Highlights of First Meeting
- Upcoming Activities

Overview and Goals of Project

- For private industry, social media (SM) has changed the way that recruiters find and attract applicants; social media is used in all phases of the hiring process
  - 70% of employers screen candidates using social media data
  - Private industry (and military) recruiters may currently use SM data idiosyncratically
- Military recruiting faces significant challenges
  - Historically low unemployment
  - Low propensity to join
  - Low eligibility
  - Low knowledge of military service
- DPAC requested the establishment of a project team to consider how social media can help with military recruiting and selection. The evaluators will consider pre-employment steps:
  - Attraction – dissemination of positive organizational image
  - Selective recruitment – generate high quality lead lists, reduce recruiting costs, evaluate SM footprints to identify applicants with favorable characteristics
  - Selection – make decisions about applicant qualifications, where appropriate (many potential technical, psychometric, and legal issues)
- Project team has expertise in industrial-organizational psychology; psychometrics; data science; and law and ethics
  - Meet four times
  - Address questions about the use of SM data in military recruiting and selection
  - Advise on a research and development agenda

Introduction of the Project Team

TEP Member	Title	Affiliation	Primary Area(s) of Expertise			
			Measurement	Technology	Ethics	Legal
Richard Landers	John P. Campbell Distinguished Professorship of Industrial-Organizational Psychology	University of Minnesota	✓	✓		
Ann Marie Ryan	Professor of Psychology	Michigan State University	✓		✓	
H. Andrew (Andy) Schwartz	Assistant Professor of Computer Science	Stony Brook University		✓		
Jalal Mahmud	Manager and R&D Tech Lead, Watson	IBM		✓		
Jeff Stanton	Professor, School of Information Studies	Syracuse University	✓	✓	✓	
Tom Serrano	*Attorney*	*Defense Human Resources Activity*			✓	✓
Dan Putka	*Principal Scientist*	*HumRRO*	✓	✓		

Richard Landers

John P. Campbell Distinguished Professor of Industrial-Organizational Psychology, University of Minnesota

- A leading researcher on the application of HR-relevant technologies, Dr. Landers' specific research interests focus on the influence of social media, gamification, and related technologies on selection and training
- Recently edited *Social Media in Employee Selection and Recruitment: Theory, Practice and Current Challenges* and *Cambridge Handbook of Technology and Employee Behavior*
- His recent research has included big data, game-based learning, game-based assessment, gamification, unproctored Internet-based testing, mobile devices, virtual reality, and online social media
- His work has been published in *Journal of Applied Psychology*, *Industrial and Organizational Psychology Perspectives*, *Computers in Human Behavior*, *Simulation & Gaming*, *Social Science Computer Review*, and *Psychological Methods,* and his research and writing have been featured in *Forbes, Business Insider, Science News, Popular Science, Maclean's*, and *The Chronicle of Higher Education*
- He currently serves as Associate Editor of *Simulation & Gaming* and the *International Journal of Gaming and Computer-Mediated Simulations*, and he is also part of the steering committee of the Coalition for Technology in Behavioral Science

Ann Marie Ryan

Professor of Psychology, Michigan State University

- Dr. Ryan has a 30-year record of research focused on improving the quality and fairness of employee selection methods and topics related to diversity and justice in the workplace
- Her particular research focus has been on practical concerns in implementing fair and accurate selection and survey programs in organizations
- A former President of the Society for Industrial and Organizational Psychology (SIOP) and past editor of *Personnel Psychology*, Dr. Ryan brings expertise in legal and psychometric considerations related to assessment and selection
- She has been active in setting professional standards for tests and assessments, such as serving on the Ad Hoc Committee for the 2018 revision of the *Principles for the Validation and Use of Personnel Selection Procedures*, the American Psychological Association Committee on Psychological Testing and Assessment, and the Defense Advisory Committee on Military Personnel Testing

H. Andrew (Andy) Schwartz

Assistant Professor of Computer Science, Stony Brook University

- Dr. Schwartz has pioneered the open vocabulary method for predicting personality from social media text
- Using natural language processing and machine learning techniques, Dr. Schwartz's research focuses on large-scale language analysis for health and social sciences
- His current projects include predicting and characterizing mental and physical health from one's language in social media, automatic lexicon refinement, measuring human temporal orientation and optimism, passive crowd-sourcing through apps, and algorithms for data-driven discovery of human insights
- Principal Investigator, World Well-Being Project, a project from the University of Pennsylvania Positive Psychology Center and Stony Brook University's Human Language Analysis Lab, to develop scientific techniques for measuring psychological well-being and physical health based on the analysis of language in social media

Jalal Mahmud

Research Staff Member: User Systems and Experience Research (USER) Group, IBM Research Almaden

- Dr. Mahmud manages the Personality Analytics research group under IBM Watson Innovation where he has conducted research on social media analysis and engagement, user modeling from social media, social collaboration tools, web task models and intelligent browsing, web automaton and testing
- He previously led several IBM Watson research teams, including personality insights, tone analyzer, and emotion modeling
- His technical accomplishments include:
  - Developing and implementing algorithms for task-based web information retrieval and web task models (using Java) from browser log analysis
  - Developing and implementing an algorithm to infer location of a Twitter user as part of IBM's First-of-a-Kind (FOAK) program and IBM's customer engagement
  - Developing models of users' likelihood to respond and information spreading from social media; delivered research result to DARPA-funded project and IBM's customer engagements
  - Developed models of user attitude from social media and delivered research result to DARPA-funded project

Jeff Stanton

Professor, School of Information Studies, Syracuse University

- Dr. Stanton is an I/O psychologist who brings combined expertise in psychometrics, data mining, and machine learning. His recent research in applied data science focuses on the management, analysis, and visualization of large data sets.
- He has published extensively in both psychology and computer science journals on these topics as well as data privacy. This experience allows him to provide an independent, but informed, perspective on the use of of social media data in recruitment and selection.
- He has conducted numerous research projects that have applied the principles of behavioral science and organizational research toward understanding the interactions of people and technology in institutional contexts.
- He is the author of:
  - *Reasoning with Data: An Introduction to Traditional and Bayesian Statistics with R*
  - *An Introduction to Data Science* (with Jeffrey Saltz)
  - *Information Nation: Education and Careers in the Emerging Information Professions* (with Indira Guzman and Kathryn Stam)
  - *The Visible Employee: Using Workplace Monitoring and Surveillance to Protect Information Assets Without Compromising Employee Privacy or Trust* (with Kathryn Stam).

Summary of First Meeting (May 14, 2019)

- Attendees from
  - DPAC, AP, JAMRS, PERSEREC, ARNG, USAREC, AFRS, MCRC, CNRC, ARI, AFPC, UK DSTL

- Agenda focused on (a) project goals and background and (b) current use of social media data in military recruiting
  - **Dr. Dan Segall** provided background information on the project and introduced a vision for the project
  - **ARNG, USAREC, AFRS, MCRC, and CNRC** presented on their current use of social media in recruiting
  - Facilitated group discussion of project vision and current use of SM data

- Provided detailed minutes to project team

Highlights from Presentations

- Dr. Segall presented on background and vision for the project:
  - Discussed use of SM in private sector recruiting and the challenges in military recruiting
  - Described the purpose of the project as to provide input on pre-employment uses of SM, specifically selective recruitment and selection, in a military context
    - How to use SM and what characteristics are predictable using SM
    - Legal, ethical, public relations, and technical considerations for SM use
    - How to supplement lead lists, prioritize leads, NOT disqualify candidates
  - Expressed interested in obtaining feedback on questions for the project team and research that should be conducted SM data are used
- Army National Guard (Ms. Julie Yorkshire)
  - Discussed use of several SM platforms to provide realistic previews and engage potential candidates
  - Answered questions about managing friend requests (they become fans of the page) and connecting candidates to recruiters (call center transfers information to the Army Lead Processing System)
- Army Recruiting Command (Mr. David Grimm)
  - Virtual Recruiting Center for national leads and Virtual Recruiting Station for every battalion - used to generate new leads, to try to dispel myths about the military and Army, and to refine leads by gathering SM data
  - Currently gathering SM data manually at the Virtual Recruiting Stations – look for information to start recruiting conversations

Highlights from Presentations

- Air Force Recruiting Service (Maj. James Kramer)
  - Do target paid marketing and human efforts based on SM profiles (of those who have interacted with Air Force content only)
  - Measure sentiment, value, and scale of posts, and use this information to determine which topics generate the most interest
  - Interested in learning about handling PII, direct messaging techniques, and employment law
- Marine Corps Recruiting Command (Mr. Brian Kornelius)
  - Concerned with identifying candidates who will not drop out of the selection and training process
  - SM posts are image-based and meant to garner interest in service
  - Create lists of users who like, comment, or reblog Marine Corps SM posts to supplement other lead generation methods
  - Interested in information on how to increase the number of leads generated or how to reduce the time spent on recruiting each candidate
- Navy Recruiting Command (Chief Grant Khanbalinov)
  - Local recruiters use SM only after they receive leads from the national level or after they meet people in person
  - Conducted live demonstration of SM use on Facebook, Instagram, Snapchat, and Reddit
    - Instagram is currently best for number of leads and interviews generated
  - Noted that recruiters are given feedback about the quality of their SM activities; supervisors monitor benchmarks
  - Clarified that he contacts all leads regardless of whether he believes they are qualified, but is more aggressive if he believes that someone would be a good candidate

Open Discussion

- Sample of technical and measurement topics
  - Predictive models built on one SM platform will have around 80% validity when applied to other platforms
  - How much information, and of what quality, is required to predict personality using SM data?
  - Many of the correlations between SM data and traditional measures are significant but very small – SM not always the better predictor
  - Test-retest reliability of SM models is about .70 for a six-month interval and about .60 after two years (similar to traditional measures)
- Sample of legal and ethical topics
  - Effects of gendered language use on text analysis
  - Not legal to consider protected demographic characteristics when making selection decisions or to use within-group norming
  - Consider how SM information could be used in an effective way - law and political environment do not allow for SM models to make selection decisions; can't collect data from third party
  - How would the Services access to SM data for use in these models – most platforms limit data scraping
  - Potential recruits may be minors – does that change the legal considerations
- Sample of "vision" topics
  - Focus on "screening in" rather than "screening out" candidates
  - Avoid mental health or security risk assessments via SM data
  - Focus on predicting propensity to join the military and scores on selection measures; not using SM as a replacement for current measures

Upcoming Activities

- Second Meeting
  - October 22, 2019
  - Alexandria, VA
  - Agenda items (technical and psychometric issues)
    - JAMRS – discuss youth survey/youth use of social media and techniques for micro-targeting using social media
    - PERSEREC – how they access social media data to conduct security clearance screening and the associated operational challenges
    - ARI - research on the use of social networks (e.g., to generate leads or to inform priors for measures)
    - Presentations from panelists:
      - Their personal research
      - Psychometric and legal considerations
- Third Meeting (Winter 2019/Spring 2020)
  - Legal and ethical issues
- Fourth Meeting (Summer 2020)
  - Recommendations; R&D roadmap
- Final Report (Fall 2020)

Questions?

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

Tab I

ASVAB AIG (WK, AR, MK, and GS)

MAPWG/DAC Update

September 26, 2019
Philadelphia, PA

Presented by
Isaac Bejar
ETS

Staffing

WK

NLP: Michael Flor

Linguistics: Paul Deane

Psychometrics:
Dan McCaffrey, Jonathan Weeks

Project Management:
James Bruno

Data Analysis: Steven Holtzman

Test Development: Adam Banta, Serguei Denissov

MK and AR

Mathematicians:
Mary Morley
James Fife
Aurora Graf

Psychometrics:
Jonathan Weeks

Data Analyst: Steven Holtzman

Project Management:
James Bruno

GS

NLP: Michael Flor

Linguistics: Paul Deane

Content specialists: Janet Koster-van Groos, Katherine Heavers

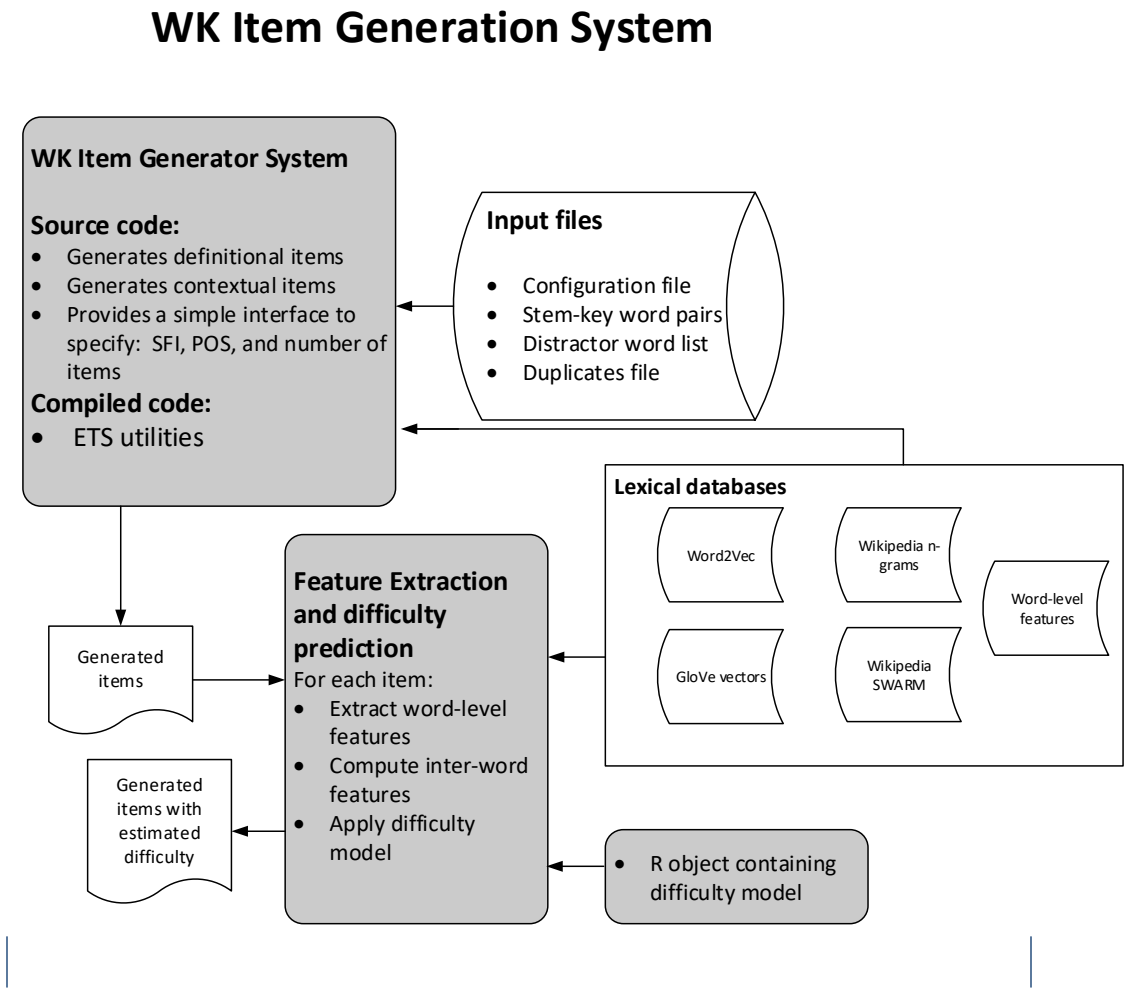
Psychometrics:
Jonathan Weeks

Data Analysis: Steven Holtzman

Intern: Denisse Garcia

WK

WK (Delivered)



MK and AR

Tasks

Task	MK	AR
4.3.1 Review of literature relevant to mathematics	☺	☺
4.3.2 Model MK and AR items from existing items	☺	☺
4.3.3 Construct item generation software	☺	☺
4.3.4 Generate pilot items	☺	☺
4.3.5 Assess quality of parameter accuracy	☺	Field test to be completed 11/30

Approach: Item Modeling

- An item model is a template that, *together with the appropriate software,* produces items intended to be on the *same* difficulty.
- Importantly, an item model includes *a set of constraints* that limits precisely the items produced by an item model.
- Typically, an item model is based on a *existing and calibrated item* that serves as the basis for authoring the item model.
- When properly authored, the items generated by an item model have similar difficulty and discrimination parameters, such that all the items produced from a given model can be pre-calibrated as if it were a single item.

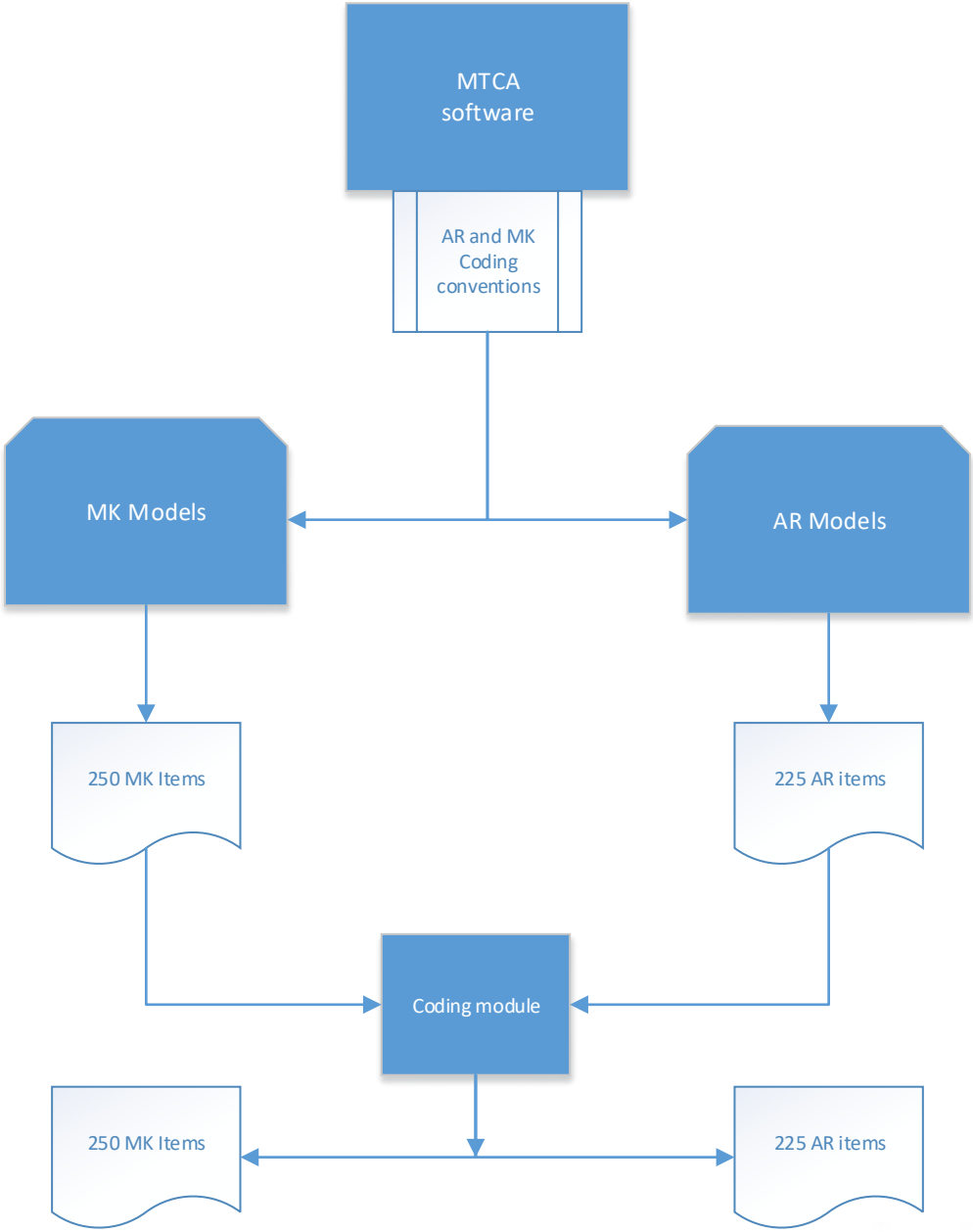
Workflow

Authoring

Review

Production

Deployment



Sample MK Item and Item Model

ITEM

Question 1. $\sqrt{\frac{27}{3}} =$

- A. $\sqrt{3}$
- B. 3
- C. 9
- D. 12

ITEM MODEL

$$\sqrt{\frac{N3}{N1}} =$$

N1 is an integer 2, 3, 4, or 5

N2 is an integer 2, 3, or 4

N3 = N1 * N2^2

Key = N2

Distractor 1 = $\sqrt{N2}$

Distractor 2 = N2 * N2

Distractor 3= N1 * (N2 +1)

Distractor 4 = N1

Distractor 5 = N1^2

MK Field Test Design

50 item models

5 items per models with different keys

Or,

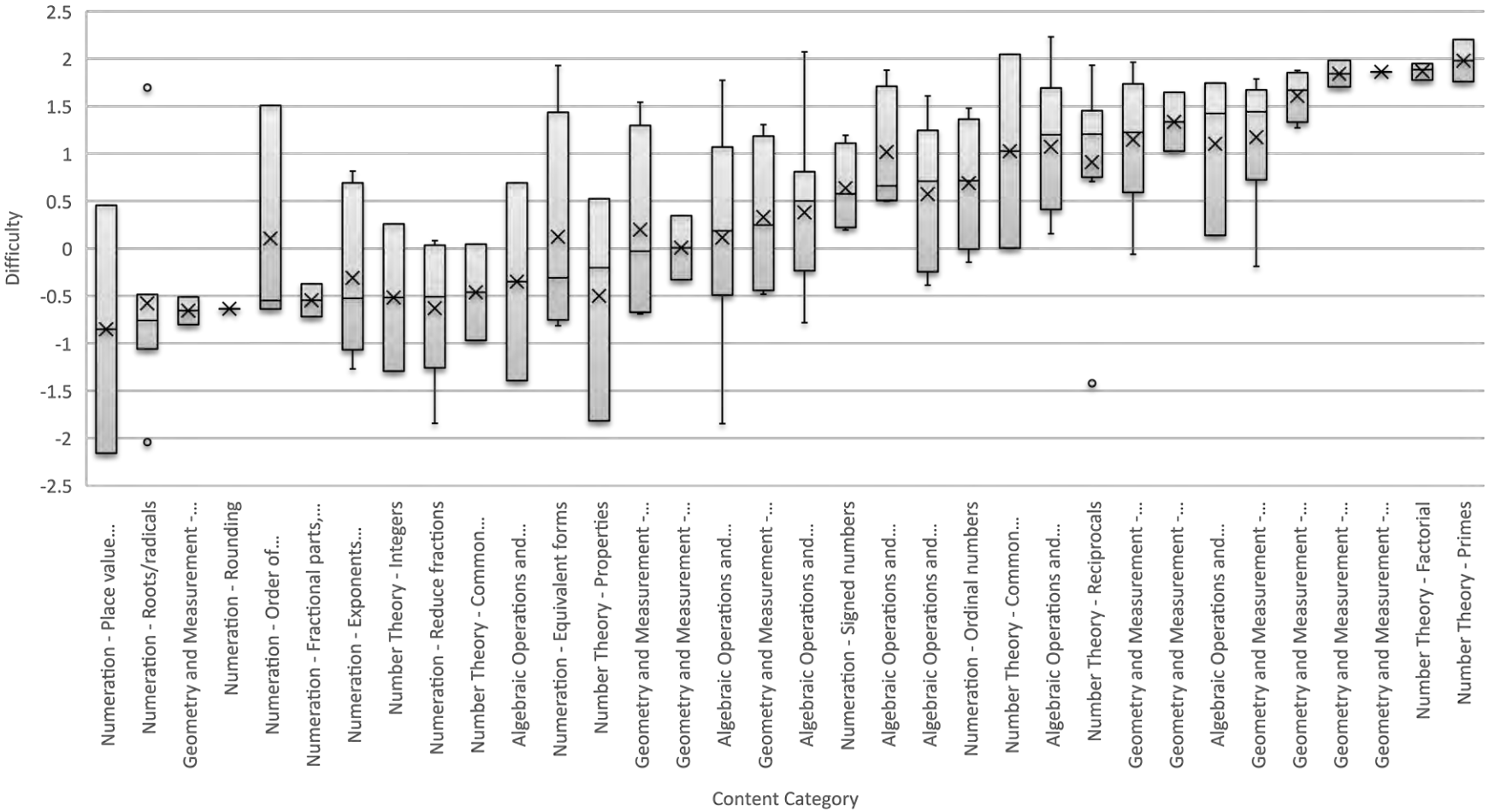
250 items

Item Models with Graphics

- Four models had graphics
- The graphics were made part of the model
- What was the fate of graphic items and models?

MK Results

Bird's Eye View of MK



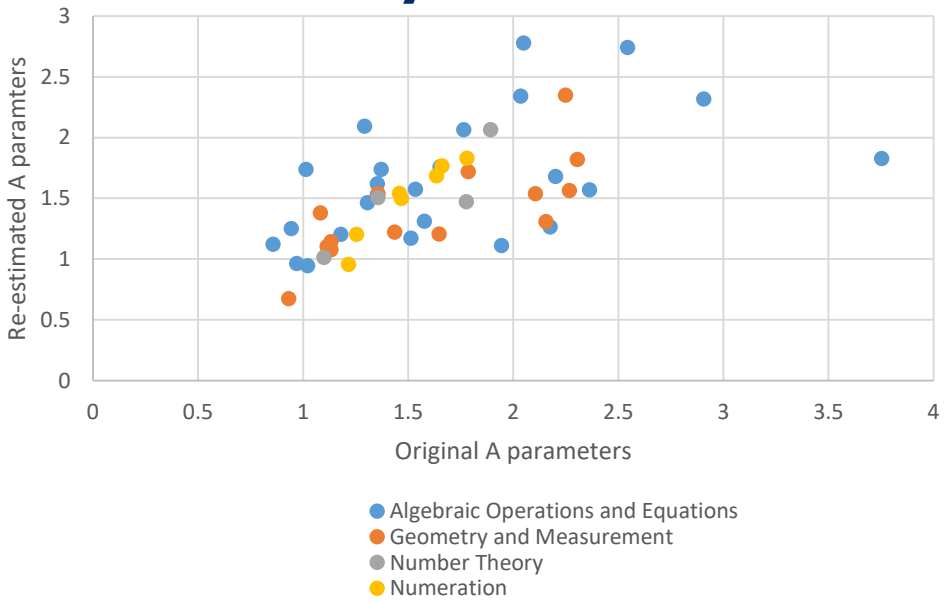
Selection of Items to Model

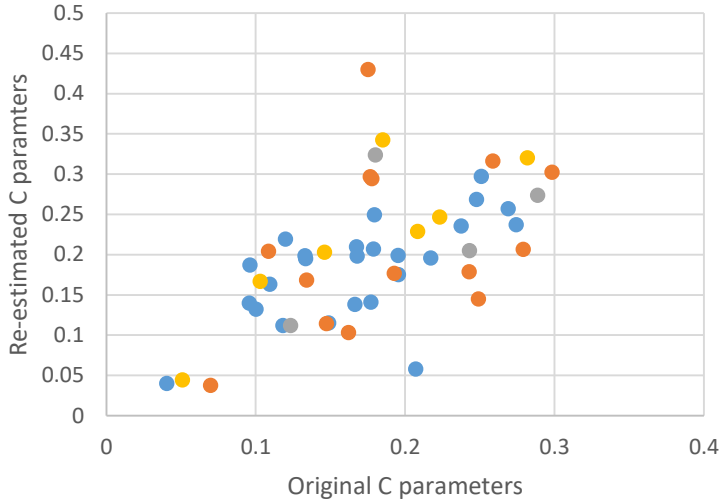
Classification	Pool 3	Pool 7	Total
Algebraic Operations and Equations - Determine equation	1		1
Algebraic Operations and Equations - Factoring		2	2
Algebraic Operations and Equations - Inequalities	1	1	2
Algebraic Operations and Equations - Simplify algebraic	1	4	5
Algebraic Operations and Equations - Solve for unknown	3	5	8
Algebraic Operations and Equations - Substitute given values	3	4	7
Geometry and Measurement - Area	1	4	5
Geometry and Measurement - Circles		1	1
Geometry and Measurement - Coordinates/slope	2	2	4
Geometry and Measurement - Perimeter	1		1
Geometry and Measurement - Polygons		1	1
Geometry and Measurement - Pythagorean Theorem	1		1
Geometry and Measurement - Volume	1		1
Number Theory - Common factors	1		1
Number Theory - Primes	1		1
Number Theory - Reciprocals	1	1	2
Numeration - Equivalent forms	1		1
Numeration - Place value/decimals	1		1
Numeration - Reduce fractions		1	1
Numeration - Signed numbers	1	1	2
Numeration - Roots/radicals	2		2
Total	23	27	50

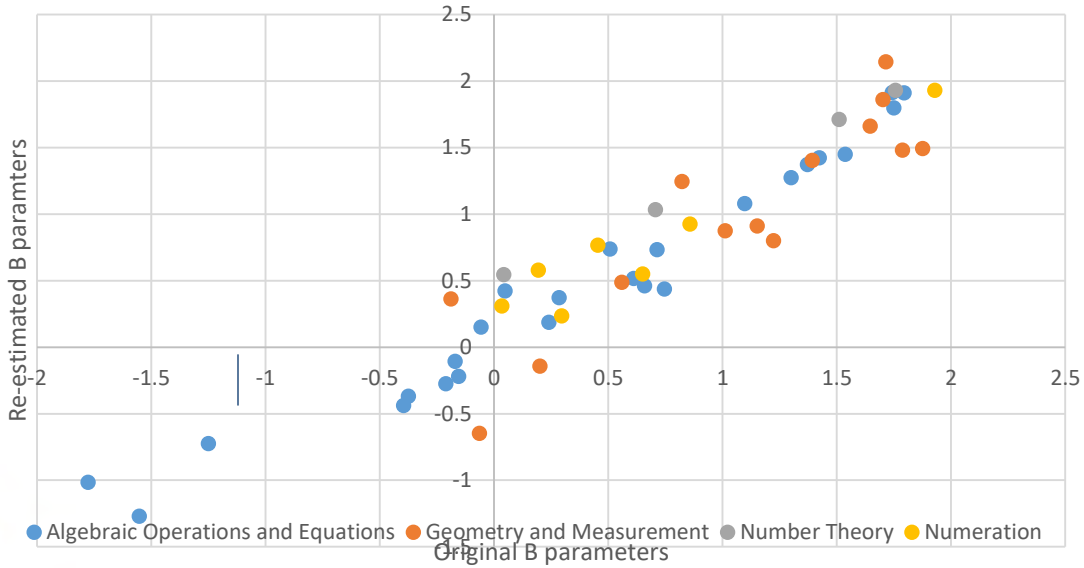
Evaluation Criteria

- Evaluation at the item level
  - How many of the 250 items are within difficulty and discrimination ranges?
- Evaluation at the model level
  - What proportion of models behave as expected?
- Cost effectiveness
  - What is the cost of each generated items?

Aside on Parameter Estimate Stability







r^2 Correlations

a	*0.34*
b	*0.91*
c	*0.33*

Decline in discrimination over time?

Increase in difficulty over time?

Difference O-R	Count	A Parameter		B Parameter		C Parameter	
Algebraic Operations and Equations	25	M 0.06	SD 0.18	M -0.08	SD 0.11	M -0.01	SD 0.00
Geometry and Measurement	14	0.22	0.11	0.07	-0.08	-0.02	-0.04
Number Theory	4	0.02	-0.06	-0.30	0.15	-0.02	-0.02
Numeration	7	0.00	-0.11	-0.12	0.07	-0.05	-0.02
All Items	50	0.09	0.11	-0.06	0.06	-0.02	-0.02

Possible Contributors to Parameter Estimate Variability

- Methodological: Estimation details
  - LOGIST vs BILOG (Yen,1987; Mislevy & Stocking, 1989) equally accurate on longer tests
- Population effects (composition of incoming testing population)
  - Economy's effect on recruitment (Warner, 2012)
  - Seasonality (Wyse & Babcock, 2016)
- Incidental effects
  - Position and context effect (Pommerich & Harris, 2003)
- Radical effects: Curricular trends
  - NCLB (Dee & Jacob, 2011), evidence from NAEP, math wars (Schoenfeld, 1985)
- Time allowance, was it comparable?
- Medium Effect

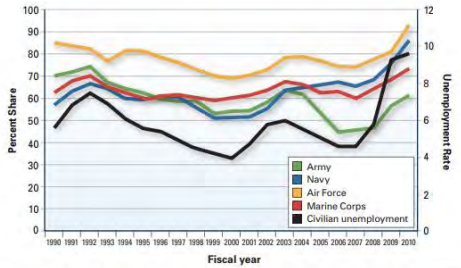


Figure 1. High-Quality Enlistments by Service and Unemployment, Fiscal Year 1990–2010

Warner, J. T. (2012). The Effect of the Civilian Economy on Recruiting and Retention. In *The 11th Quadrennial Review of Military Compensation: Supporting Research Papers: https://militarypay.defense.gov/Portals/3/Documents/Reports/11th_QRMC_Supporting_Research_Papers_(932pp)_Linked.pdf.*

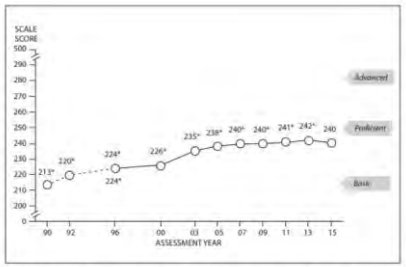


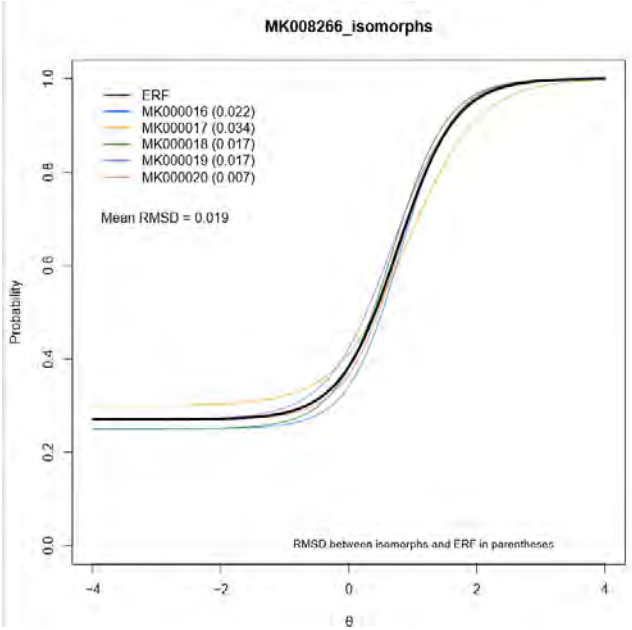
Fig. 3. Trend in fourth-grade NAEP mathematics mean scale scores. Graphic from the National Center for Education Statistics, *The Nation's Report Card: A First Look—Mathematics and Reading 2015*. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics, 2015d. Used with permission.

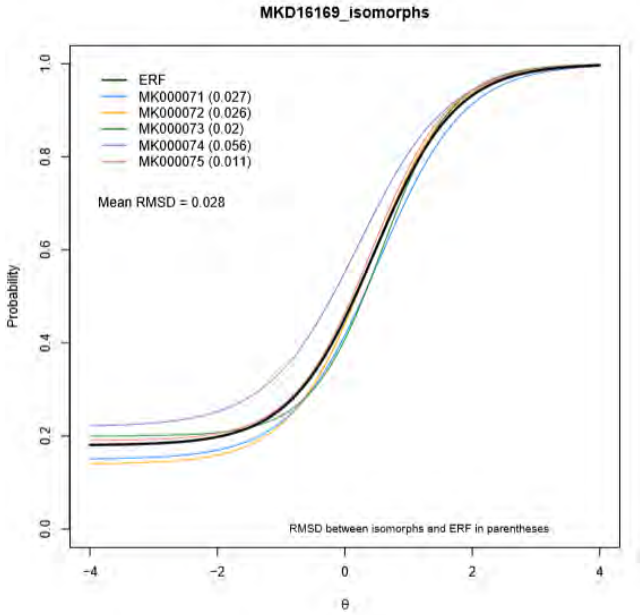
Item Modeling Results

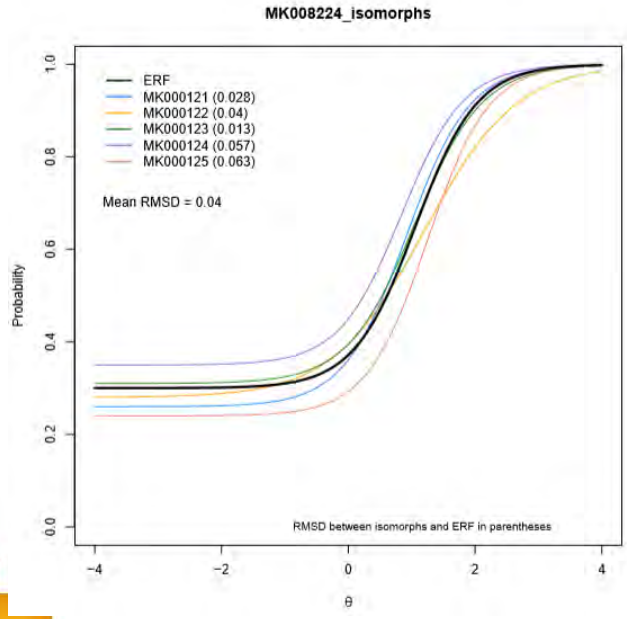
1) Evaluation at Item Level

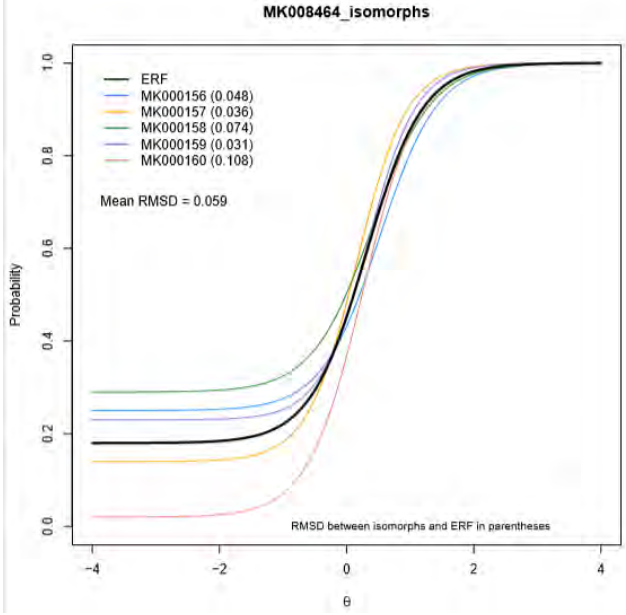
- All but 13 items out of 250 met difficulty and discrimination criteria

2) Model-Level Analysis

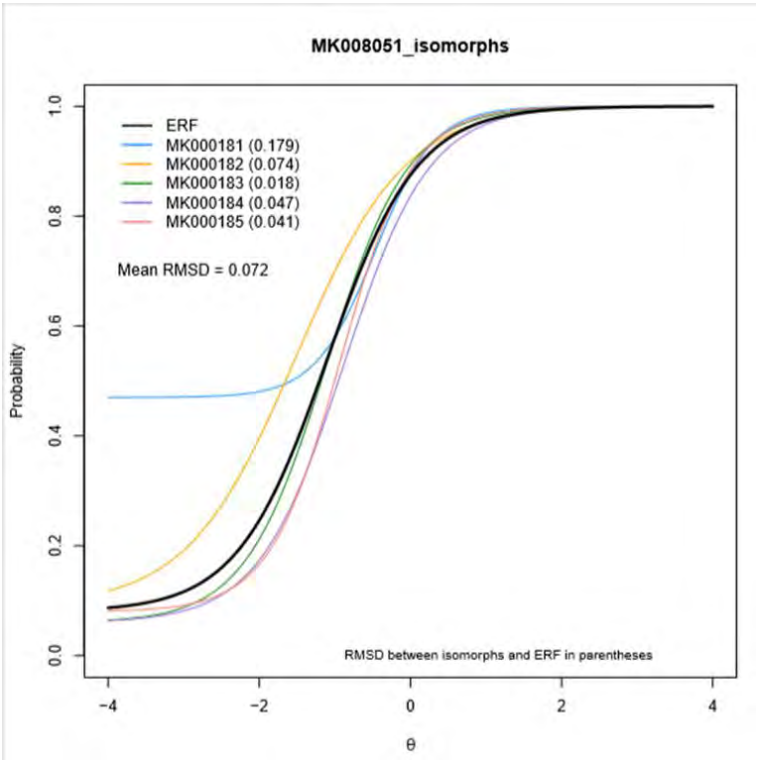




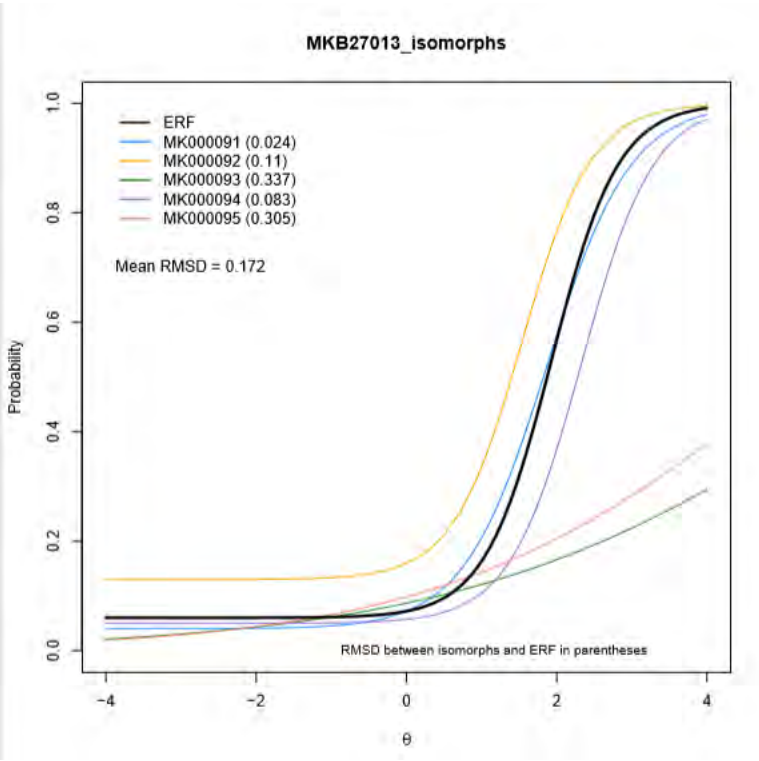




Potentially fixable or hopeless?



☹



RMSD Histogram

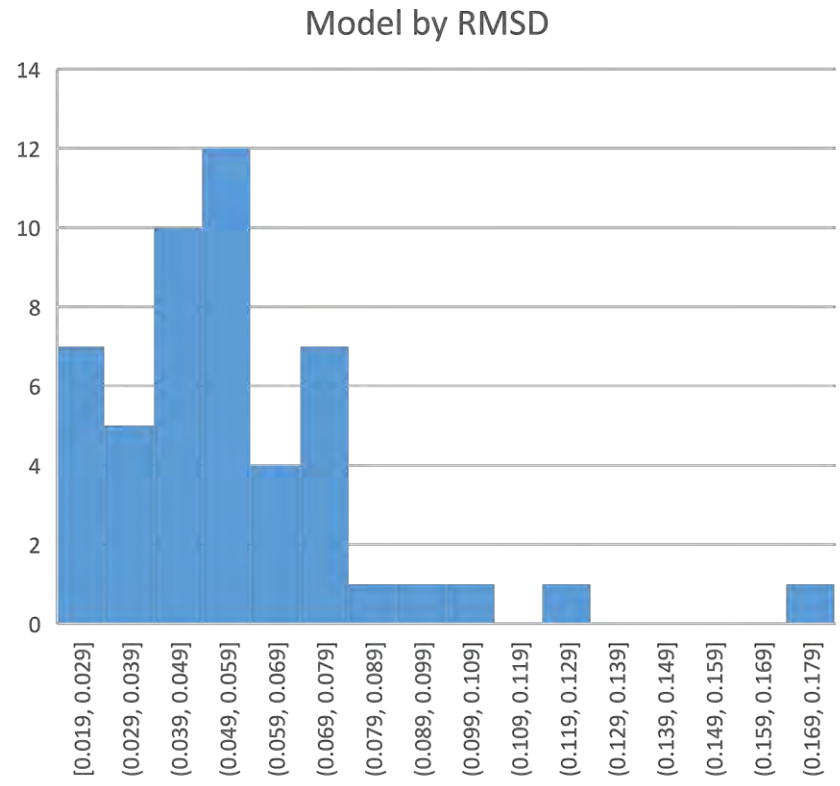
$$RMSD_i = \sqrt{\frac{\sum_{k=1}^{K} w_k [P(\theta_k)_i - P^*(\theta_k)]^2}{\sum_{k=1}^{K} w_k}}$$

θ_k = Quadrature Point

w_k= Quadrature Weight

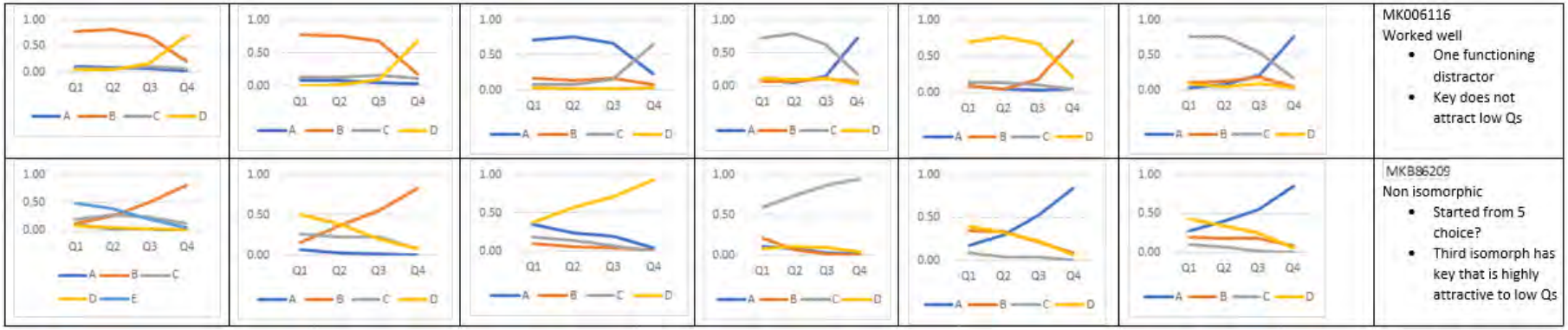
$P(\theta_k)_i$ = Probability of a correct response for item *i*

$P^*(\theta_k)$ = Probability of a correct response for the expected response function



parent id	Original pool 3	Original five choice	Retest did not work	Original Low A	Original High A	Original Low B	Original High B	Original Low C	Original High C	Model Distractor issues	Too much model numerical variation	Variation in irt C not clear why	
MKC27184													4
MKD17308													4
MKC27305													3
MKA37004													3
MKB27013													5
MKC77055													3
MKD97100													4
MKD87167													3
MKC36266													3
MKC16035													2
MKA56178													3
MKC26289													2
MKB86158													2
MKB86209													2
MK008051													4
MK008316													3
MK008320													2
MK008373													1
MK008697													2
MK008667													3
MK008998													1
MK008391													3
MK008941													2
MK008464													1
MK008468													1
	14	9	2	5	4	4	3	3	2	5	2	13	

Distractor Analysis



Graphics Items

Interactive Content Here

3) Cost Effectiveness Analysis (MK)

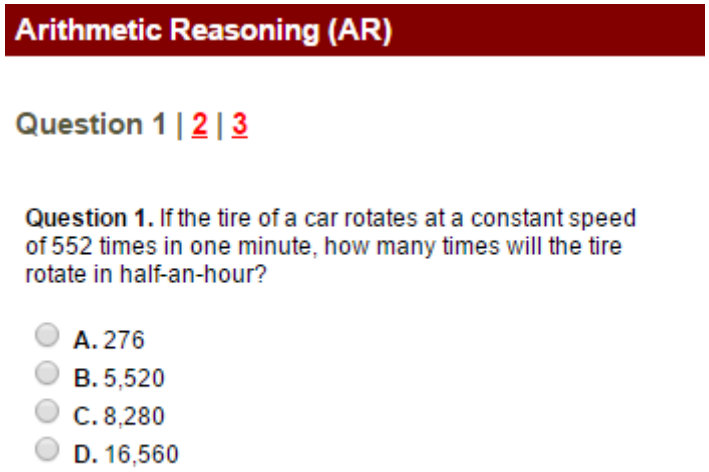
- The total number of items for the 50 models is, conservatively, 500.
- But, note that isomorphs of the same model could not appear in the same pool
- The total number of hours (MK + AR)
  - Authoring: 400 hours, assume 200 were for MK
  - Review: 200 hours, assume 100 were for MK
  - Authoring and review, 300hs
  - Graphic modeling included
- The time per item is (500 items with 300 hours of labor) .6 hour
- If we use only items from working models, 1.2 hour per item
- Actual cost depend on salary rates, overhead etc.

MK conclusions

- Nearly 100% of the items generated were functional.
- The % of acceptable models (50%) could have been higher:
  - Acceptability would increase with what we now know:
    - Pool 3 had many more 5-choice items
    - Success rate was 60% with 4-choice items
    - Rotating distractors could have an effect on c
    - Review distractor analysis prior to modeling
    - Avoid modeling items with *a*'s that are lower than 1
      - (None of the low *a* models worked)
    - Low *b*'s were much less likely to work
      - At this level numbers matter: incidentals could become radicals
    - The cutoff for declaring a model unsuccessful requires additional research
      - It is potentially possible to compensate for *random* variation

AR

Sample AR Item



AR Field Test (ongoing, projected completion 11/30/2019)

- 45 models
- 225 items

Family	Difficulty			
	Lower	Medium	Higher	Total
1	5	5	5	15
….				
15	5	5	5	15
				225

Criteria for Evaluation

- Item-level yield
- Model-level yield
  - Within-model: Was difficulty successfully held constant?
  - Between-model: Was difficulty successfully manipulated?
- Cost effectiveness

AR parent items				
id	original pool 3	original 5-choice	low C	high C
C17195				
E47192				
C17135				
B27088				
C16384				
B26637				
C16348				
D16416				
008471				
008578				
008445				
008595				
008352				
008336				
008195				

GS

GS Milestones

Task	GS
4.3.1 Review of literature relevant to mathematics	☺
4.3.2 Model GS items from existing items	☺
4.3.3 Construct item generation software	☺
4.3.4 Generate pilot items:	Jan 2020
4.3.5 Assess quality of parameter accuracy; provide software and documentation	Sep 2020

Scope

- GS covers several domains
- Scope of field test limited to Anatomy and Physiology, and Zoology

Question 1. Air is less dense than water because

- **A.** it is lighter.
- **B.** its molecules are further apart.
- **C.** its molecules are closer together.
- **D.** it moves more quickly and easily.

Question 2. 100° C is equal to

- **A.** 32° F.
- **B.** 100° F.
- **C.** 200° F.
- **D.** 212° F.

Question 3. Salt helps to melt ice because it

- **A.** dissolves in water to form an acid.
- **B.** chemically destroys the water molecules.
- **C.** lowers the temperature at which water freezes.
- **D.** is attracted to concrete sidewalks below the ice.

Approach

- Infer the construct from analysis of items by analyzing GS 3 and 7 into "item models"
- Based on GS 3 and GS 7, we identified the following high-level item models:

			GS3			Total			GS 7			Total	Grand Total
ITEM MODELS	Astronomy	Biology	Chemistry	Geology	Physics		Astronomy	Biology	Chemistry	Geology	Physics		
category example		3	3		2	8		5				5	13
category part		8	1			9	1	2	3			6	15
cause effect	3	3	1	2	7	16	1			1	3	5	21
concept application	2	5	9	2	5	23		1	5	1	3	10	33
concept definition	2	26	9	9	11	57	1	3	2	1	4	11	68
concept identification		4	3		2	9		9	2	3	2	16	25
fact identification		2		1		3				1		1	4
structure function		6		1	2	9		3			1	4	13
term synonym			1			1		1				1	2
(blank)													
Grand Total	7	57	27	15	29	135	3	24	12	7	13	59	194

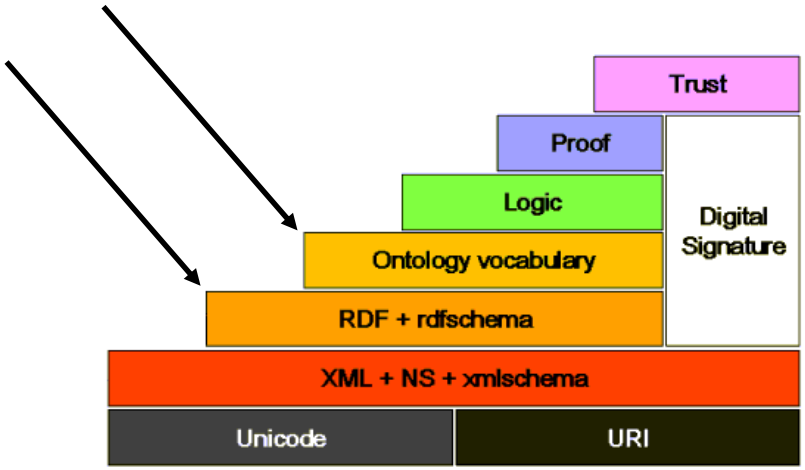
Category	Criteria	Sample Item
category example	Student matches a scientific category (e.g., animal tissue) with a member of that category (e.g., muscular, nervous, epithelial, connective).	Item 1 Which of the following is NOT a type of animal tissue? A. muscular tissue B. nervous tissue C. vascular tissue* D. epithelial tissue
category part	Student matches a category (e.g., digestive system) with a key component or part (e.g., liver).	Item 2 The liver is part of the A. digestive system* B. excretory system C. pulmonary system D. circulatory system
cause effect	Student matches a described set of conditions (e.g., reduced sweat evaporation on a hot and humid day) with a corresponding effect (e.g., heat stroke).	Item 3 Sweat does not evaporate well on a hot and humid day. This results in an increased risk of A. sunburn B. hypothermia C. sun blindness D. heat stroke*
concept application	Student applies understanding of a focal concept (e.g., digestion) to reasoning about a situation (e.g., boy eats sandwich).	Item 4 A boy is hungry and eats a sandwich for lunch. After several hours,
concept definition	Student matches a focal concept from the domain (e.g., respiration) with its definition (e.g., exchange of oxygen and carbon dioxide between cells and the external environment).	Item 5 Diabetes is a disease in which A. blood sugar levels are not controlled* B. blood vessels that supply blood to the heart become narrowed C. a blood vessel in the brain is blocked or leaks D. blood does not clot properly
concept identification	Student matches a focal concept from the domain (e.g., homeostasis) with a specific instantiation (e.g., heart rate slowing in response to high blood pressure).	Item 6 Heart rate slowing in response to high blood pressure is an example of A. homeostasis* B. cardiac arrest C. hemoglobin D. capillary action
fact identification	Student completes a correct scientific fact.	Item 7 The largest human organ is the A. brain B. small intestine C. stomach D. skin*
structure function	Student matches a component structure (e.g., kidneys) with its function in the within the whole (e.g., remove waste products from the blood).	Item 8 The kidneys are a pair of organs that A. remove waste products from the blood* B. secrete acid and enzymes that digest food C. absorb nutrients from digested food D. store solid waste before it is expelled
term synonym	Student matches a technical scientific term (e.g., cranium) with its everyday equivalent (e.g., skull).	Item 9 The cranium is also known as A. the heart B. the stomach C. the spine D. the skull*

Approach to generation

- Generation of GS items requires a biology knowledge base or ontology (suitable to K-12!)
- We tried a K-12 ontology that was developed for the purpose of *answering* K-12 science items: it did not generate suitable items
- Conclusion: We need to generate our own ontology

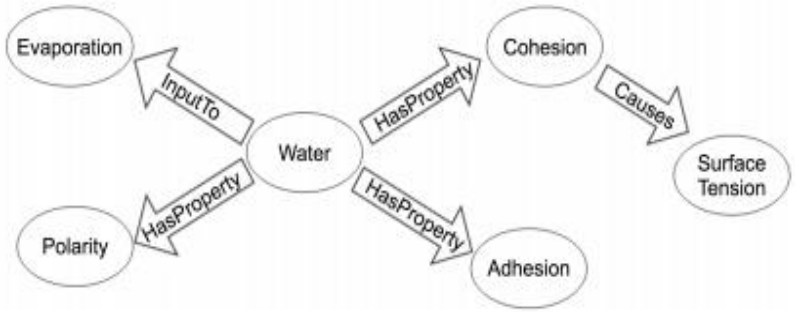
Approach to Knowledge Representation

- Semantic Web
- Uses RDF
  - Resource Description Framework (semantic triples)
  - Ontology and Vocabulary



Semantic triples in GS

- Top-down
  - Starting from an ontology,
  - (node,relation,node) triples
  - Ontology (partially) represented in textbook: manually extract triples

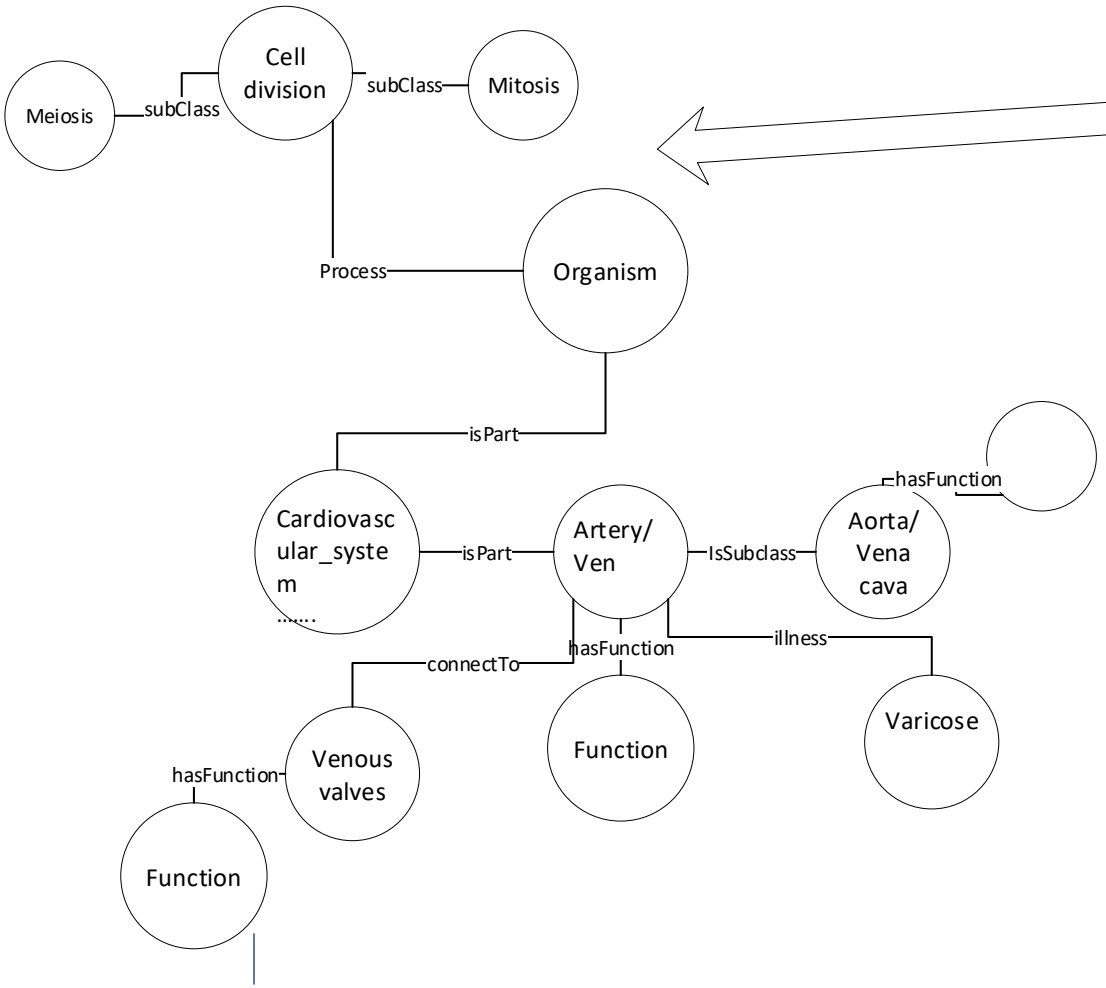


Stasaski, K., & Hearst, M. (2017). *Multiple choice question generation utilizing an ontology. Paper presented at the Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications.*

- Bottom-up,
  - Information extraction
  - Use text patterns to identify (subject, predicate, object) triples
    - Identify sentences with target vocabulary
    - **Apply extraction patterns**
    - Collect triples
    - SME reviews

Corresponding triples		
Water	InputTo	Evaporation
Water	HasProperty	Polarity
Water	HasProperty	Cohesion
Water	HasProperty	Adhesion
Cohesion	Causes	Surface Tension

Semantic triples in GS (Anatomy and Physiology and Zoology)

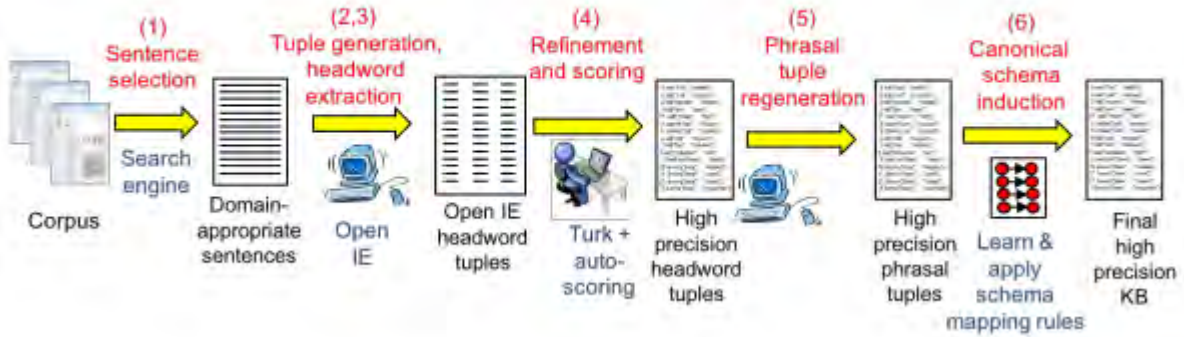


Triple Representation

Meiosis | is_a_subclass |Cell Division
Mitosis |is_a_subclass |Cell Division.
Cell Division | is_a | Process
Organism |undergoes | Cell Division

GS Ontology Development

- Development of vocabulary
  - Current status
- SME-curated triples
- Experiment with information extraction



Dalvi Mishra, B., Tandon, N., & Clark, P. (2017). Domain-Targeted, High Precision Knowledge Extraction. *Transactions of the Association for Computational Linguistics, 5, 233-246.* *doi:https://transacl.org/ojs/index.php/tacl/article/view/1064*

GS Field Test Design

Anatomy-Concept Definition-Lower Difficulty item model

	Difficulty		
Content	Lower	Medium	High
Anatomy A3.1			
Concept definition	5	5	5
Concept example	5	5	5
Physiology A3.2			
Concept definition	5	5	5
Concept example	5	5	5
Zoology A2			
Concept definition	5	5	5
Concept example	5	5	5

Current Status

- SME-curated triples in progress
- Upload items for field testing: January 1, 2020
- Information extraction will be implemented for comparison
  - Compare the efficacy of the two approaches

Q&A

Tab J

OPA

OFFICE OF PEOPLE ANALYTICS

ASVAB Career Exploration Program

August 2019

OPA

OFFICE OF PEOPLE ANALYTICS

ASVAB

CAREER EXPLORATION PROGRAM

Discussion Topics

- ASVAB CEP Usage Metrics
- Expert Panel Recommendations and Progress
- State Usage and ESSA
- PTI Proficiency Training
- #OptionReady

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

ASVAB CEP Usage Metrics

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

ASVAB CEP Numbers and Metrics

Year*	Number of Students Tested
2013	670,836
2014	690,950
2015	687,900
2016	706,200
2017	684,223
2018	713,777
2019	**786,807**

Year*	Number of Schools Tested	Percentage of Schools Tested
2013	12,613	56%
2014	12,731	56.4%
2015	12,929	56.6%
2016	13,169	57.2%
2017	12,870	55.5%
2018	12,380	55%
2019	**13,976**	**60.6%**

*School year runs from July 1- June 30. Data as of 30 June of respective year.

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

School Year 18-19

Paper and Pencil Numbers

	Examinees 17-18	Examinees 18-19
TOTAL	662,564	**714,333**

CEP iCAT Numbers

	Examinees 17-18	Examinees 18-19
TOTAL	51,213	**72,474**

*Total students as of 30 June each year.

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Leads

Year*	Leads Provided to Military Services
2014	492,419
2015	470,229
2016	478,196
2017	440,542
2018	433,317
2019	**468,003**

Total students as of 30 June each year.

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Accessions By Service:
Number of students using their ASVAB CEP score for enlistment

Year*	ARMY	NAVY	AIR FORCE	MARINE CORPS	COAST GUARD	TOTAL
2014	14,513	4,439	3,677	5,474	130	28,233
2015	15,156	4,731	3,669	5,682	285	29,523
2016	14,449	4,990	4,121	5,655	310	29,525
2017	15,053	4,310	4,465	6,037	392	30,257
2018	14,432	4,699	4,234	5,370	405	29,140
2019	**13,430**	**4,963**	**4,700**	**5,163**	**358**	**28,614**

*School year runs from July 1- June 30.
**ASVAB CEP Score is usable for two years.

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Website Utilization: www.asvabprogram.com
(July 1 – June 30)

	17-18	18-19
Unique Visitors	440,882	582,162 ↑
Returning Visitors	203,357	225,100 ↑
Page Views	6,747,160	8,550,582 ↑
Bounce Rate	31.41%	28.18% ↓
Average Time Per Session	12:55	11:43
Number of Pages Per Session	10.49	10.10
Tablet/Mobile Visitors	212,870	332,850 ↑

Access Code Utilization
(July 1, 2018 – June 30, 2019)

Code Type	Visitors	Repeat Visitors
Marketing	2,068	925
Counselor	2,022	1,014
Student	246,895	99,452
Reserve	719	340
Total Number of Logins	**251,704**	**101,731**

The number of student access codes used indicates a heavy reliance on the paper based booklets, schools not requesting a post test interpretation, resource limitations for a second visit, or a failure to educate the schools about the PTI option.

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Website Utilization: www.careersinthemilitary.com
(July 1 – June 30)

	17-18	18-19
Unique Visitors	72,230	104,531
Returning Visitors	39,515	46,320
Page Views	2,003,165	1,976,405
Bounce Rate	31.90%	24.29%
Average Time Per Session	4:34	5:53
Number of Pages Per Session	17.93	13.12
Tablet/Mobile Visitors	26,330	50,316*

*The new site was built using Angular JS, a promising technology for interactive websites. However, Google indexing services are not up to speed with tracking content on sites built with Angular JS. As a result, Google search was not crawling our site, significantly reducing our organic search results.

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Contact Us: www.asvabprogram.com
(July 1, 2018 – June 30, 2019)

- **Inquiries: 1,212**
- **Bring ASVAB CEP to Your School: 1,469**
  - **Student or Parent: 822**
  - **Counselor: 647**
- **Score Requests: 2,038**

Total: 4,719

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Contact Us: www.careersinthemilitary.com
(July 1, 2018 – June 30, 2019)

- Army: 32
- Marine Corps: 17
- Navy: 25
- Air Force: 15
- Coast Guard: 10

Total: 99

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Expert Panel/Needs Assessment: Recommendations and Progress

Program Processes:

- Develop electronic reminder system to maintain contact with students and parents between testing and interpretation.
- Develop a mechanism for including parents in the career exploration process.
- Develop a protocol for sending personalized reminders of available resources and activities that can further engage the students in reflecting on their assessment results and career/educational exploration plans.
- Session numbers and other manual processes hindering effectiveness of program (Needs Assessment Finding)
    - **Business Modernization Contract**
        - **Includes a look at manual processes that occur throughout the lifecycle of the CEP, from scheduling to accountability. Includes requirements gathering to update legacy systems. Joint project with USMEPCOM. June 2019-2020**

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Expert Panel/Needs Assessment: Recommendations and Progress

Assessments:

- Expert panel recommended a review/evaluation of the current FYI item pool to achieve an inventory that encompasses critical, occupationally relevant tasks for high school students and is culturally appropriate
    - **FYI Revision Efforts**
        - **Presentation by Olga Friedman to follow on current state of FYI. Future efforts will commence September 2019-2021**

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Expert Panel/Needs Assessment: Recommendations and Progress

Websites:

- Under Resources, some are .pdf files and some are Word documents. The panel recommends converting all resources into writeable .pdf files. The website also should provide a link to Adobe for those who might not have downloaded the software to read pdfs.
- The Resources section also contains links to items that are not relevant to students (i.e., ASVAB CEP Counselor Guide). The panel recommends that the website is configured such that the available resources are specific to each user population (students, educators, parents).
    - **Reconfigured and implemented a new resource center on asvabprogram.com, includes integration of twitter feed and other social media.**
    - **Website reconfiguration planned with website refresh (2021-2022)**

Expert Panel/Needs Assessment: Recommendations and Progress

Marketing

- The ASVAB CEP and the FYI should be marketed to professional journals and textbooks, as well as integrated into the National Certified Counselors curriculum.
    - **ASVAB CEP included in: A comprehensive Guide to Career Assessment, 7th Edition, Published by NCDA**
    - **https://www.ncda.org/aws/NCDA/pt/sd/product/11018/_PARENT/layout_products/false**

Part IV – Career Assessment Instrument Reviews

Ability Explorer – Kathy M. Evans (Both Print and Online)
Ashland Interest Inventory – Darrin L. Carr, Pamela McCoy, & Alyssa West
ASVAB – Laith G. Mazahreh
California Psychological Inventory – Rebekah Reysen
Career Decision Self- Efficacy Scale – Joshua C. Watson (Both Print and Online)
Career Occupational Preferences System Interest Inventory (COPS) – Jenna Crabb
Career Thoughts Inventory – Brian M. Calhoun
Jackson Career Explorer – Justin R. Fields
Jackson Vocational Interest Inventory – Julie Aitken Schermer
Kuder Career Planning System – Melinda M. Gibbons & Charmayne R. Adams (Both Print and Online)
NEO-4 – Brian J. Taber
Occupational Aptitude Survey and Interest Schedule (OASIS-3) – Amanda G. Flora
Self- Directed Search – Chad Luke & Zach Budesa
Work Values Inventory – S. Autumn Collins
Career Construction Interview – Louis A. Busacca (Both Print and Online)
Career Genogram – Tina M. Anctil
Knowdell Card Sorts – Tanya M. Campos

Expert Panel/Needs Assessment: Recommendations and Progress

Reports and Interpretation

- Post Test Interpretations, inconsistencies, lack of website usage (Needs Assessment)
- Identify strategies to increase engagement with the CEP to increase the amount of time spent exploring occupational information and other resources to strengthen the depth of processing and personal relevance. For example, involve more interactive online activities, and link to the portfolio within the CITM with internet resources and activities that supplement guided exploration with counselors and parents.
    - **PTI Training Effort**
        - **Detailed information on the virtual and in-person training to follow. February 2019-August 2019 (As needed)**

Expert Panel/Needs Assessment: Recommendations and Progress

- Identify strategies to increase engagement with the CEP to increase the amount of time spent exploring occupational information and other resources to strengthen the depth of processing and personal relevance.
- Conduct a thorough review of training provided to Education Services Specialists and of post-test interpretation processes, to include updates to the websites, measures, and activities.
- Expand the interpretative information for ASVAB results to provide detailed suggestions for understanding and strengthening skills. Consider including additional reflective questions to enhance understanding of results and facilitate action steps for the career exploration process.
- Make it a priority to increase the number of computerized administration options beyond the current CEP, including smartphone applications.
    - **Classroom activity inclusion and PTI Training**
        - **By educating the field on the information included in the websites, and inviting recruiting commands to participate in the train-the-trainer model, we have built in multiple opportunities for schools, counselors, and recruiters to use the CEP to inform students of their options. (Ongoing)**

Expert Panel/Needs Assessment: Recommendations and Progress

Other Comments, Suggestions, Ideas for the Program:

- Develop a work values measure that links an individual's work values to occupations
    - **Under contract September 2019-2020**
- Develop a Successful Job Search toolbox on the asvabprogram.com website that includes how to develop a resume, use social media, search for a job, and interview for a job
    - **Under contract June 2020**
- Provide uniform credentialed career development training to ASVAB CEP administrators and interpreters
    - **USMEPCOM investigating funding options and feasibility of incorporation in the CP-31 Program.**
    - **Investigating offering CEUs for pre-conference workshops**

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

State Usage and ESSA

State Legislature Schedules for 2020

- http://www.statescape.com/resources/legislative/session-schedules

- Another useful site (but current schedules only)
  - http://www.ncsl.org/research/about-state-legislatures/2019-state-legislative-session-calendar.aspx

Contracting Effort: State Usage of ASVAB

- Monitoring websites
  - State Boards and Departments of Education
- Goal: Glean any mention of their use of ASVAB or CEP
- Method
  - Excel file with links to each state's Board or Department of Education news/press release website, dummy variable tracking if state websites offer any information on ASVAB CEP
  - Websites checked weekly for any updates/changes
  - Offer notes on ESSA and other career exploration information from the state

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

States that reference ASVAB CEP or career development in legislation

Indiana
Kentucky
Michigan
Minnesota
North Dakota
New Jersey
Virginia
West Virginia
Wyoming

- Indiana, Minnesota, New Jersey, North Dakota, and West Virginia, mention ASVAB CEP as a career exploration tool
- Pending legislation in Kentucky will require ASVAB testing with follow-up counseling in high schools
- ASVAB CEP endorsed online by Superintendent of Education in Wyoming

ASVAB CAREER EXPLORATION PROGRAM

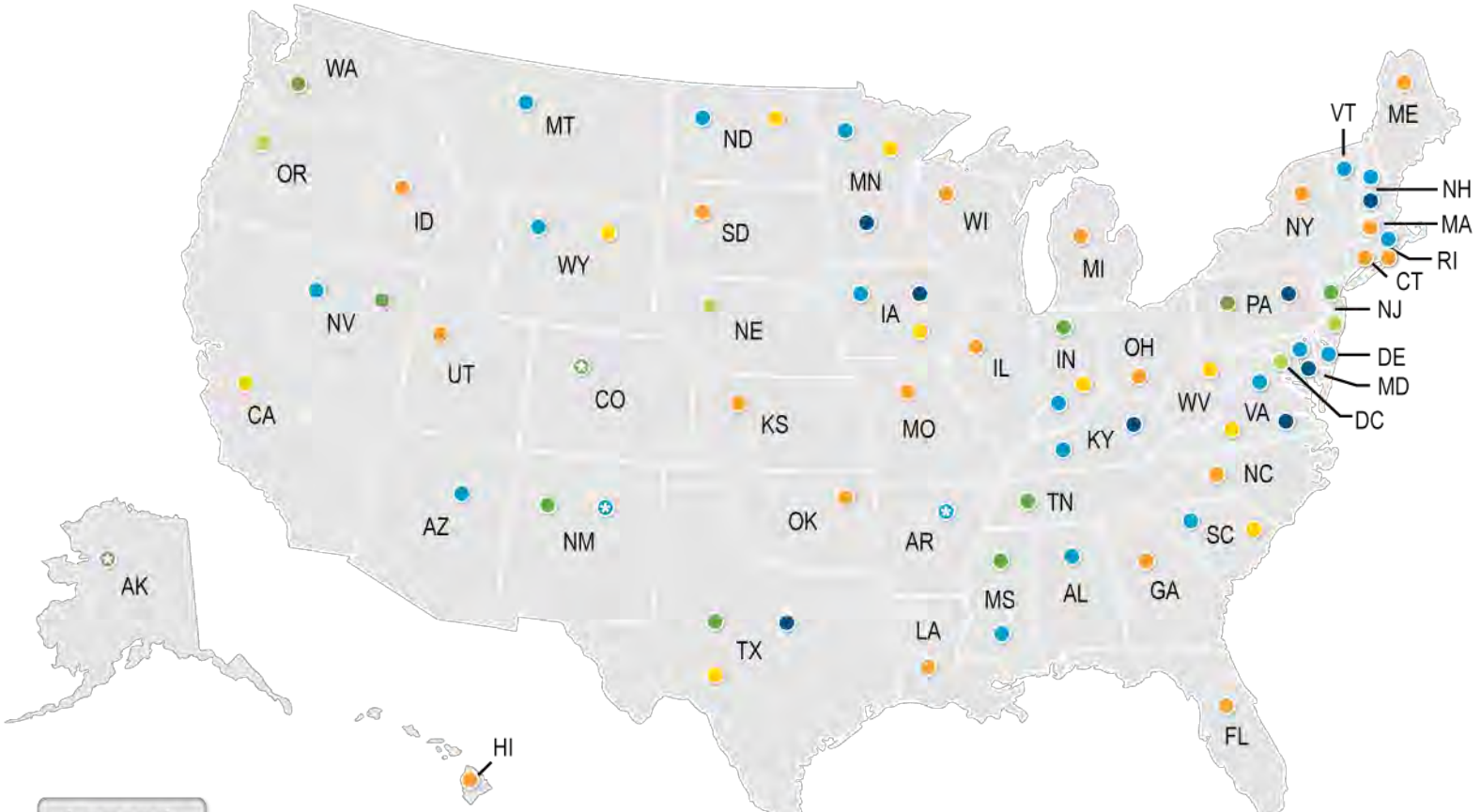
States that mention ASVAB only in legislation

Alaska
Colorado
Maryland
Missouri
Mississippi
New Hampshire
New Mexico
Nevada
New York
Oregon
South Carolina
Tennessee
Texas
Vermont

- Alaska, Colorado, Maryland, Mississippi, Missouri, Nevada, New Hampshire, New Mexico, and Vermont indicate use of ASVAB as one indicator of college and career readiness
- Michigan lists ASVAB as a "suggested strategy" to meet Target 2 (Contextualized Academics) in the Michigan Career Development Model
- New York gives option of offering ASVAB as a school choice, and Oregon provides links to ASVAB websites for more information
- South Carolina ESSA Plan indicates use of ASVAB AFQT score as an indicator of military readiness
- Legislation passed in Texas will allow students to use ASVAB as vocational test

Option to Meet Graduation Requirement	Military and Career Readiness Indicator	Required/ Recommended for Career Exploration	Legislation Activity	Mentions ASVAB CEP on State Website	Limited or Specialized Use
Alaska	Alabama	California	Iowa	Nebraska	Connecticut
Colorado	Arkansas	Indiana	Kentucky	New Jersey	Florida
Indiana	Arizona	Iowa	Maryland	Oregon	Georgia
Mississippi	Delaware	Minnesota	Minnesota	Washington, DC	Hawaii
New Jersey	Iowa	North Dakota	New Hampshire		Idaho
New Mexico	Indiana	South Carolina	Pennsylvania		Illinois
Nevada	Kentucky	Texas	Texas		Kansas
Pennsylvania	Maryland	Virginia	Virginia		Louisiana
Tennessee	Minnesota	West Virginia			Massachusetts
Texas	Mississippi	Wyoming			Maine
Washington	Montana				Michigan
	North Dakota				Missouri
	New Hampshire				North Carolina
	New Mexico				New York
	Nevada				Ohio
	Rhode Island				Oklahoma
	South Carolina				Rhode Island
	Virginia				South Dakota
	Vermont				Utah
	Wyoming				Wisconsin

Level of Engagement by State



OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

ASVAB CEP and ESSA

- Several States have passed legislation requiring schools to provide ASVAB CEP to high school students, however, legislation is not worded accurately.
- Military Services have been speaking to legislators about the ASVAB CEP and the benefits of the program. But many of them are not aware of the program updates.
- Stakeholder meeting held August 15 in Alexandria, VA which included participants from AP, USMEPCOM, Military Service Liaisons, and members from the Defense State Liaison Office.
- Agenda: Review of States and how they use ASVAB CEP, Service initiatives, review of legislative calendar, AP guidance, data and processes for obtaining data, social media and tool kits

ASVAB CEP Orientation Event in Indiana

- ASVAB CEP is included as a Pathway to Graduation in Indiana
  - IDOE hosted a three-hour overview of program components
  - 125 attendees
    - Most were previously unaware of the website offerings
    - 70% said they will utilize the PTI in the future
  - Opportunity for follow up webinar series
    - How planning tools map to pathways

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

ESS and Recruiting Commands Engagement with State BOEs

- Supplied a memo to the field regarding the appropriate uses of the ASVAB CEP (Approved by AP)
- Included guidance during State presentations, Q/A sessions, National Conferences on appropriate uses of ASVAB CEP
- Will continue to monitor State legislation and utilization of ASVAB CEP

PTI Proficiency Training

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

PTI Proficiency Training

Because:

1. We have more States looking at the ASVAB CEP as a program to give for career exploration, we have an increased pressure to visit schools more than once (additional work load), and deliver a standard program. (Needs Assessment, Expert Panel Recommendation)
2. Because we have added so much new functionality to asvabprogram.com and careersinthemilitary.com, MEPS ESSs (as a whole) have not been adequately trained to use the websites effectively or to train others to use them. (Needs Assessment, Expert Panel Recommendation)
3. We have not had a way to track our national work force for ASVAB CEP in delivering PTIs. With the introduction of virtual training, a standard metric, and training, we can now establish this. (DAC Recommendation)
4. This training will be a stepping stone to the Certified Career Counselor Credential offered by the National Career Development Association. (Needs Assessment, Expert Panel Recommendation, DAC Recommendation)
5. The standard metric of required elements, with behavioral anchors, removes most subjectivity of the training process, and allows us to use it as a learning and evaluation tool. (Expert Panel Recommendation)

PTI Proficiency Requirements

1. Be Nominated to become proficient
2. Complete virtual training modules
3. Be observed effectively conducting a PTI
4. Load proof of proficiency into Moodle

Post-Test Interpretation Proficiency Training

Goal: Standardize the process by which post-test interpretation (PTI) sessions are conducted. Serve as a workforce multiplier (using a train the trainer approach) by including Recruiting Service Partners to satisfy demand for program in schools.

Purpose: Address expert panel recommendations to orient attendees to the ASVAB CEP enhancements, and teach attendees the strategic purposes of collaborating with others operating within their territory to achieve missions.

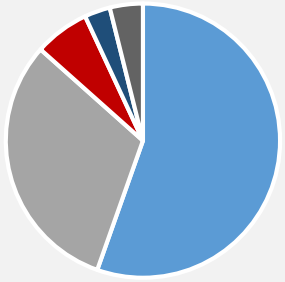
Metrics to Gauge Success: Increased utilization of ASVAB CEP related websites, increased testing numbers, virtual training use, in-person training attendance, additional access opportunities for recruiters.

Next Session

None Scheduled

In-Person PTI Training Participation

Total:
231



USMEPCOM	128
Army	72
Navy	15
Air Force	7
Coast Guard	9

1,487
Total User Accounts
(Virtual Accounts Created)

922
Total Nominees Since Training

304
Total Workforce Strength

Steps to Become PTI Proficient

Step 1: Virtual Training

Step 2: In-person training

Step 3: Nominate, Observe, Evaluate others

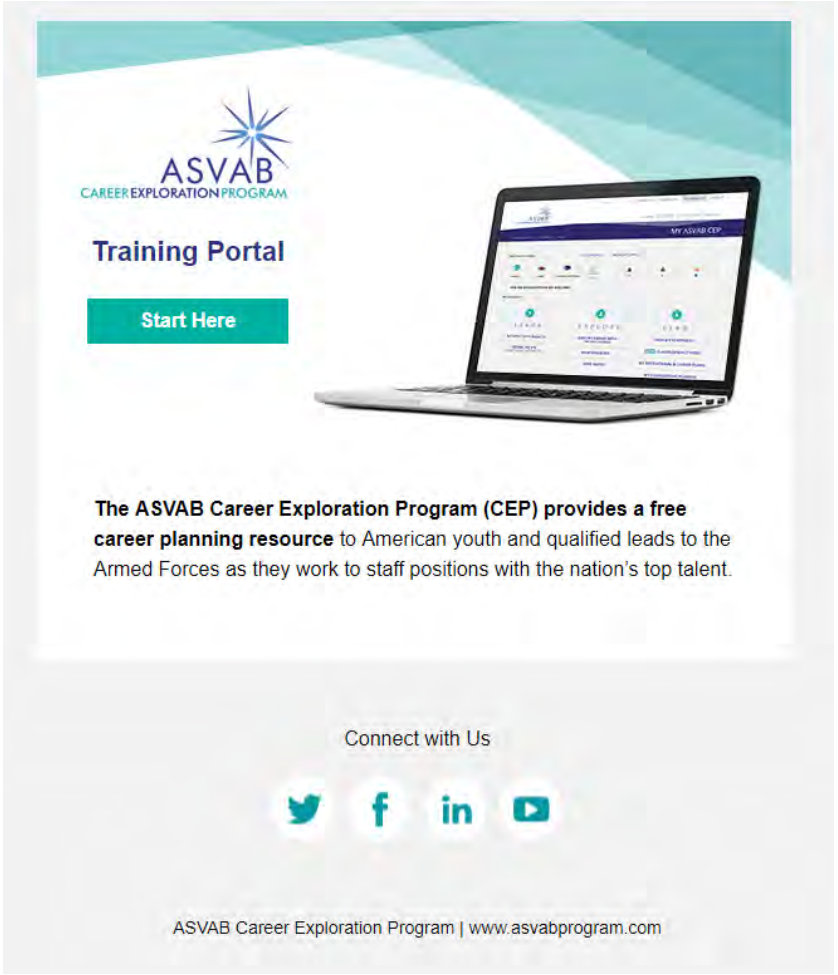
(Those trained at OPA/MEPCOM formal in-person training sessions can train and evaluate others)

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Virtual Training Consists of:

- User authentication
- Learning Objectives
- Multimedia Content: including videos, print materials, social media, etc.
- Concept Checks, Reflection and Application Activities
- Area to upload supporting documentation
- List of all people who are PTI proficient, regardless of job function or affiliation
- Area to assign three-year access codes that are pre-populated with scores
- Communication system for all people who are conducting PTIs across the US
- Ability to collect information about training needs



OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Virtual Training Topics Covered

ASVAB Measurement, Data, and Use

- ASVAB History and Validity
- ASVAB Score Release Options for Schools
- ASVAB Score Use Policies

Interpreting & Discussing ASVAB Scores

- Basic Testing Theory & The ASVAB
- Understanding the ASVAB Summary Results Sheet
- Preparing Students to Work with the ASVAB Scores
- Discussing Students' ASVAB Scores

ASVAB CEP Components

- The ASVAB CEP
- The ASVAB
- Find Your Interests
- OCCU-Find and Career Planning Tools
- Additional Resource – Careers in the Military

Conducting a Post Test Interpretation

- Overview of Essential PTI Components
- Online Post Test Interpretation
- Interpreting ASVAB Results
- Administering the FYI Online
- Demonstrating the OCCU-Find
- Demonstrating Careers in the Military
- Additional Tools and Materials

Becoming PTI Proficient

- How to Use the PTI Proficiency Evaluation Metric
- Lesson

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

In-Person Group Composition

- Small Group
  - Maximum 6
  - Unrelated Facilitator
  - Mixed experience and job function
  - Mix of geographic locations
  - No supervisor/supervisees in the same group
- Homework Group
  - Maximum 8
  - No one from small group
  - Close geographic proximity
  - Mix of job functions
  - Homework assignments were centered around topics that required discussion and collaboration (website analytics, ethical case studies)

Attendee Feedback

"Having very little experience with the ASVAB, I found it easy to retain information and overcome the learning curve. This was because of well-prepared content and a thoughtful and engaging structure"

"The metrics are easy to follow and organized logically. They will make this process to nominate other personnel for the training to be simple."

"As a recruiter, the content of this training is, without a doubt, completely game changing. It will allow me to access non-military friendly schools with ease and provide future generations with valuable tools."

"I do not, and likely will not conduct PTIs as a recruiter, but by now knowing this procedure exists, I will look to escort my MEPS TCs and ESSs for their PTIs."

"A lot of information provided throughout the presentation and interaction with the variety of participants from different agencies and services."

75%

did not know much about the site **until** they participated in the online training.

93%

were **satisfied** or very satisfied with the training.

Theme

Many participants were **unaware** of the nature and use of Career Exploration Scores and the training **provided this framework** for them.

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

#OptionReady

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

What is #optionready?

- The ASVAB CEP can be used in various ways:
    - Post-Secondary Options
    - Career Options
    - Score Release Options
    - Scalable Implementation Options

- Being #optionready is being informed about the options
- Landing page: asvabprogram.com/option-ready
    - Sharable content school counselors can easily use to inform their community about benefits of ASVAB CEP participation and encourage sign up
    - Sharing portal for those who wish to upload photos and videos from PTI workshops to encourage peer-to-peer sharing

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

#optionready Campaign

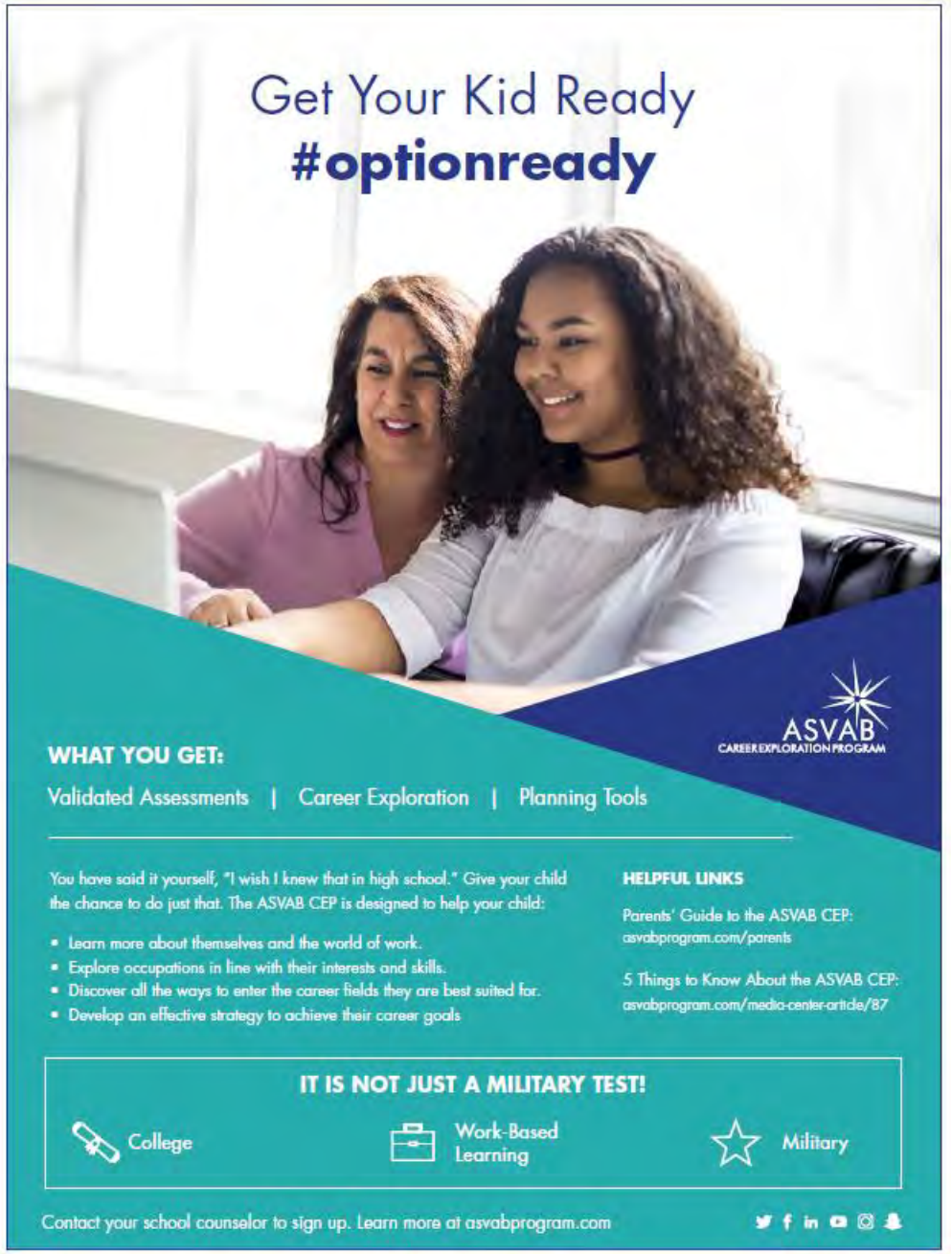
Goal: Reach 1 million participants in the ASVAB CEP within one academic year

Purpose: Correct misconceptions to improve reputation. Build awareness of the benefits of participation to increase participation.

Metrics to Gauge Success:

- Number of landing page visits
- Number of toolkit downloads and content shares
- Number of #optionready engagements on social
- Number of photos and videos submitted via sharing portal
- Increase in participation
- Increase in bring it to your school requests

OPA
OFFICE OF PEOPLE ANALYTICS



Monthly Toolkit

Two-part Goal:

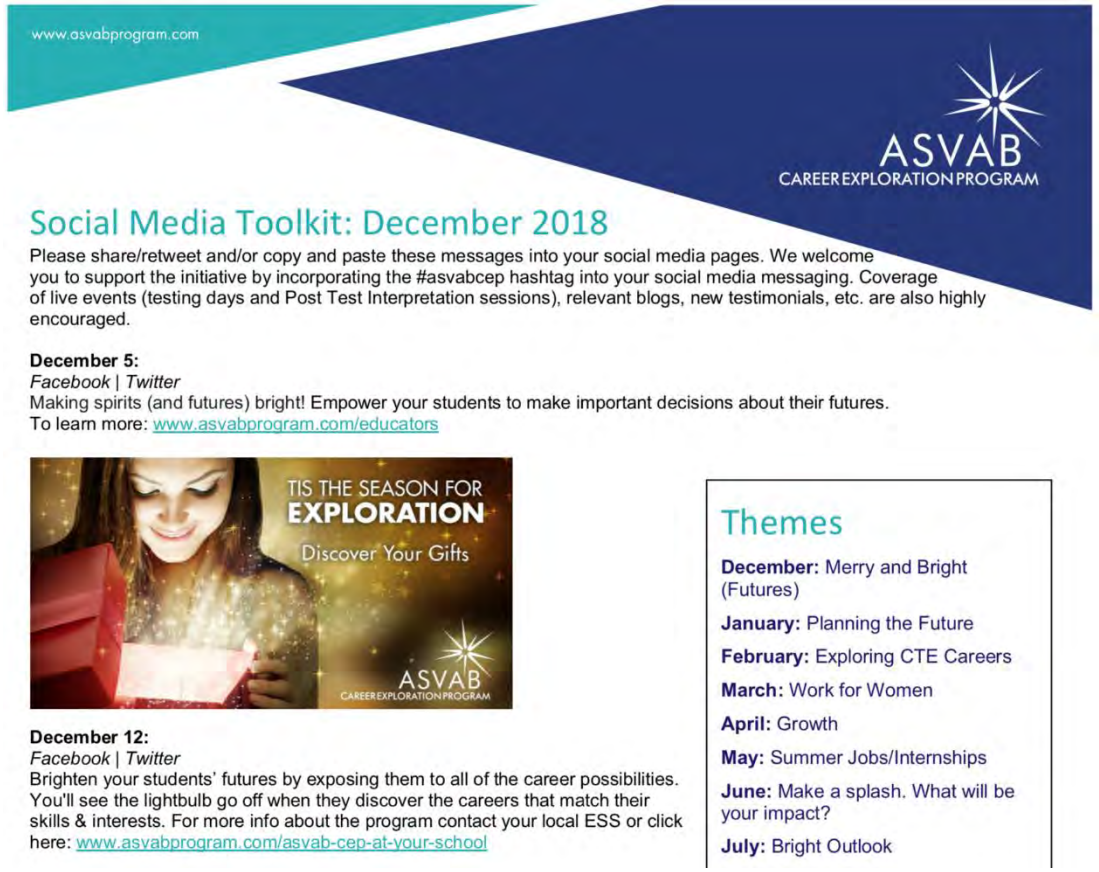
1. Make it simple to engage with ASVAB CEP on social media
2. Increase student participation

#whatsyourdreamjob

#optionready

Key Messages

- Sample Posts
- Links to all channels



www.asvabprogram.com

ASVAB CAREER EXPLORATION PROGRAM

Social Media Toolkit: December 2018

Please share/retweet and/or copy and paste these messages into your social media pages. We welcome you to support the initiative by incorporating the #asvabcep hashtag into your social media messaging. Coverage of live events (testing days and Post Test Interpretation sessions), relevant blogs, new testimonials, etc. are also highly encouraged.

December 5:
Facebook | Twitter
Making spirits (and futures) bright! Empower your students to make important decisions about their futures.
To learn more: www.asvabprogram.com/educators

December 12:
Facebook | Twitter
Brighten your students' futures by exposing them to all of the career possibilities. You'll see the lightbulb go off when they discover the careers that match their skills & interests. For more info about the program contact your local ESS or click here: www.asvabprogram.com/asvab-cep-at-your-school

Themes

December: Merry and Bright (Futures)
January: Planning the Future
February: Exploring CTE Careers
March: Work for Women
April: Growth
May: Summer Jobs/Internships
June: Make a splash. What will be your impact?
July: Bright Outlook

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

National Events

Marketing Events

- **American Counseling Association, March 28-30**
  - Booth: 277 Leads
- **National Career Development Association, June 26-30;**
  - Booth: 32 Leads
- **National Charter Schools Conference, June 29-July 1;**
  - Booth: 265 Leads
- **American School Counselors Association, June 30-July 3;**
  - Booth: 525 Leads
- **National Principals Conference, July 17-21**
  - Booth: 47 Leads

Upcoming Events

- **National Career Pathways Network, October 12-13**
  - Booth & Presentation
- **Association for Career and Technical Education, Dec 4-7**
  - Booth, Exhibitor Presentation, Two Panel Presentations

Education/Research Industry

Upcoming Events

- **IHIET, Aug 24**
  - **Presentation:** Enhancing the Military Civilian Crosswalk
- **IMTA, Oct 12**
  - **Presentation:** Symposium on ASVAB Career Exploration Program - Broadening Options at All Stages of a Career; Enhancing the Military Civilian Crosswalk; UNIFORM Web-based Application: Meeting DoD Occupational Data Needs

Stakeholder Engagement

- **PTI Proficiency Training February-March, August**
- **Arizona DOE, April 2019**
- **Georgia ACTE, JROTC July 2019**
- **JAMINAR, July**
- **US Army National Educator Tour, May**

Upcoming Events

- **Indiana DOE, Sept 12, 2019**

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Inquiries

National Program Director

Shannon.d.salyer.civ@mail.mil

Sign Up to Receive Social Media Toolkit

kelly@writtenllc.com

Like, Follow, Share @asvabcep

#asvabcep | #youdecide | #optionready

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Tab K

Evaluation of the Find Your Interests Inventory

Presented to the MAPWG

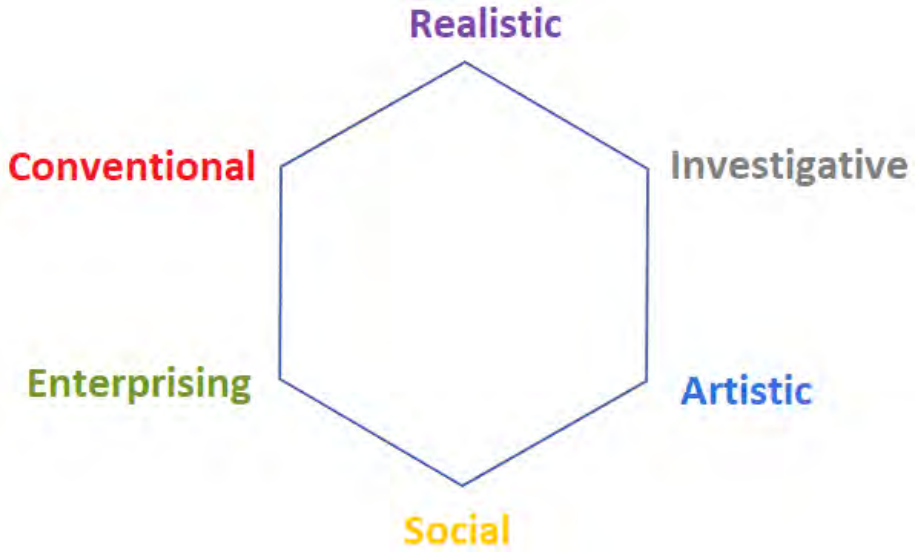
Olga Fridman
Mary Pommerich
Shannon Salyer
Defense Personnel Assessment Center
August 13, 2019 | Alexandria, VA

OVERVIEW

- **Background**
- **Goals and Methods**
  - Multidimensional Scaling
  - Factor Analysis
- **Conclusions**

BACKGROUND

- The Find Your Interests inventory (FYI, Baker, Styer, Harmon, & Pommerich, 2010) was developed in 2005 for the ASVAB Career Exploration Program (CEP).
- The FYI inventory was designed to measure interest in six areas in accordance with Holland's (1985) theory of career choice.
- Holland's theory states that people generally fall into one of six personality types (Realistic, Investigative, Artistic, Social, Enterprising, and Conventional; RIASEC), and that work environments can also be categorized into one of the six RIASEC categories.



Schematic relationship among Holland's personality/work environment categories

BACKGROUND

- Shannon Salyer organized archiving FYI inventory data in the HumRRO database.
- There were 505,109 records of the FYI inventory responses collected during years 2015–2017 in the HumRRO database.
- 321,687 records were retained for this study after removing duplicates.
- If there was more than one record with the same access code, all such records were removed.
- A record contains 90 responses to 6 groups of questions with 15 items per group.
- For each question there are three response alternatives: “Like,” “Indifferent,” or “Dislike.”

GOALS AND METHODS

- Now, 14 years after development, we are performing an evaluation of the FYI inventory in view of accumulated data.
- Such an evaluation is desirable because tremendous changes in technology, the economy, and social structure could make some content of the inventory outdated and even irrelevant.

- We aimed to answer the following questions:

1. How well (if at all) does the interest inventory match Holland's hexagon structure nowadays?
2. Does the FYI apply equally to men and women?
3 What can be done to improve FYI quality?

- We employed two statistical methods to aid us in this analysis
  - **Multidimensional scaling**
  - **Factor analysis**

GOALS AND METHODS

- In February 2019, HumRRO issued a report, "*Initial Research on the Revision of the Find Your Interests Inventory for ASVAB CEP*."
- HumRRO made several recommendations. We focused on the following two:
  - Minimize possible differences related to gender.
  - Use multidimensional scaling method to achieve a more robust RIASEC model in a revised FYI inventory.

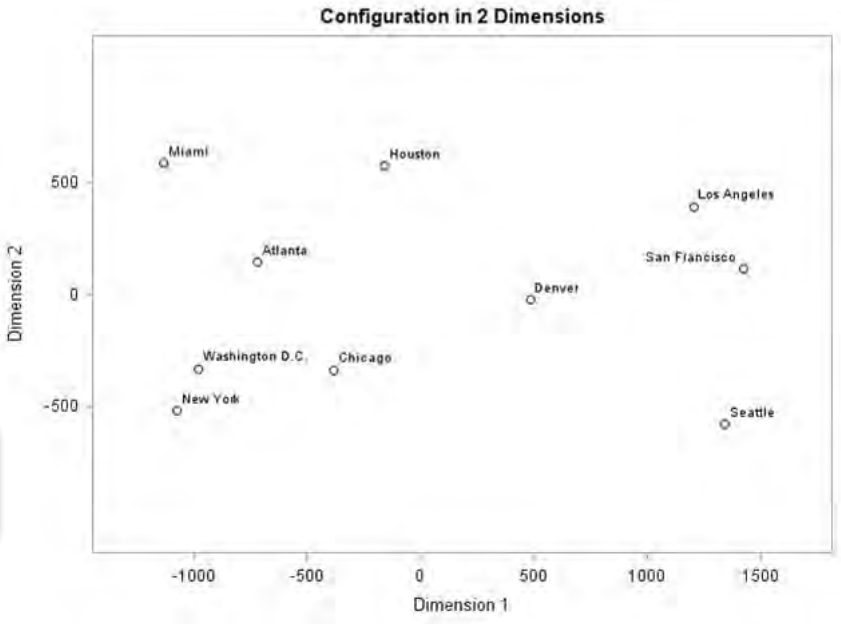
MULTIDIMENSIONAL SCALING

- Multidimensional scaling (MDS) can be considered an alternative to factor analysis. In general, the goal of the analysis is to detect meaningful underlying dimensions that allow the researcher to explain observed similarities or dissimilarities (distances) between the investigated objects.

- One example of an application of MDS is to reconstruct a map from a table of distances between points on the map. While MDS can recover the relative positions of the cities, it cannot determine absolute location or orientation. In this case, east is on the left, west is on the right of the plot.

0										Atlanta
587	0									Chicago
1212	920	0								Denver
701	940	879	0							Houston
1936	1745	831	1374	0						Los Angeles
604	1188	1726	968	2339	0					Miami
748	713	1631	1420	2451	1092	0				New York
2139	1858	949	1645	347	2594	2571	0			San Francisco
2182	1737	1021	1891	959	2734	2408	678	0		Seattle
543	597	1494	1220	2300	923	205	2442	2329	0	Washington D.C.

In our case, the similarities between objects are expressed in a correlation matrix.



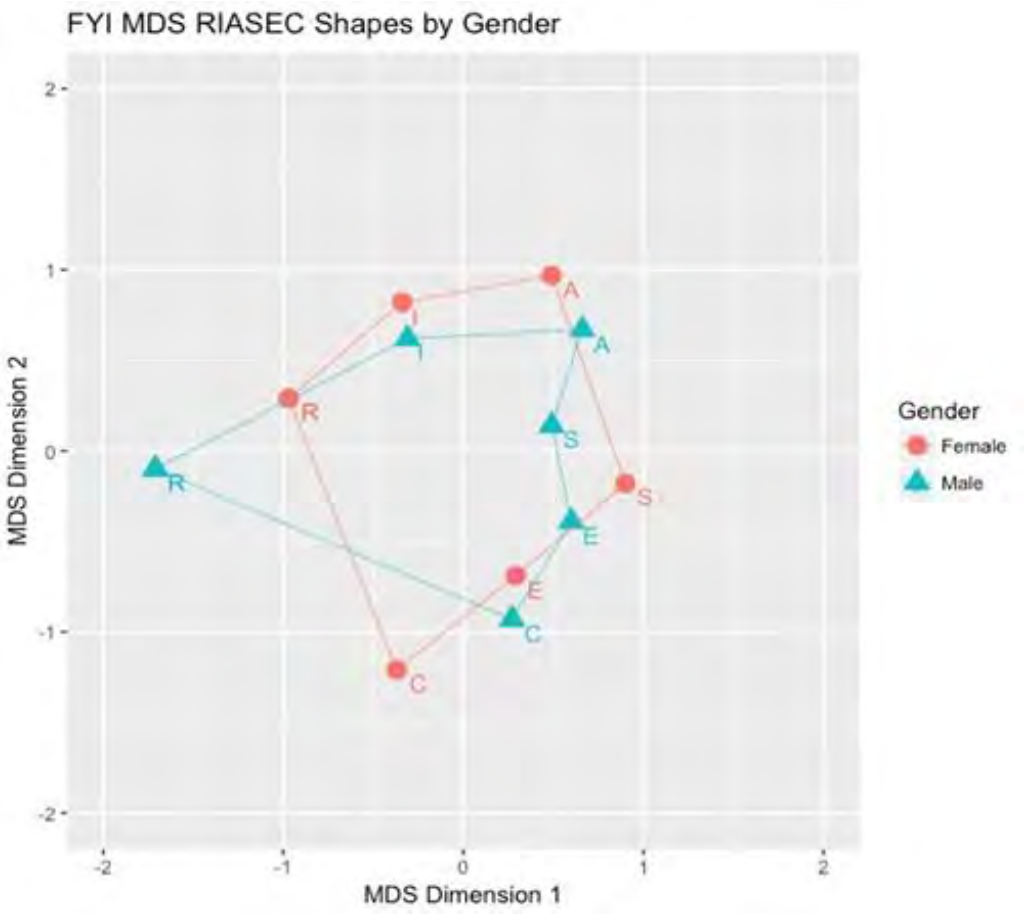
2-D MDS ANALYSIS FROM HumRRO's REPORT

Table 5. Intercorrelations for FYI RIASEC Scales for Females

Scale	Realistic	Investigative	Artistic	Social	Enterprising	Conventional
Realistic	--					
Investigative	.28	--				
Artistic	.23	.31	--			
Social	.13	.13	.27	--		
Enterprising	.16	.16	.25	.36	--	
Conventional	.25	.11	-.02	.21	.60	--

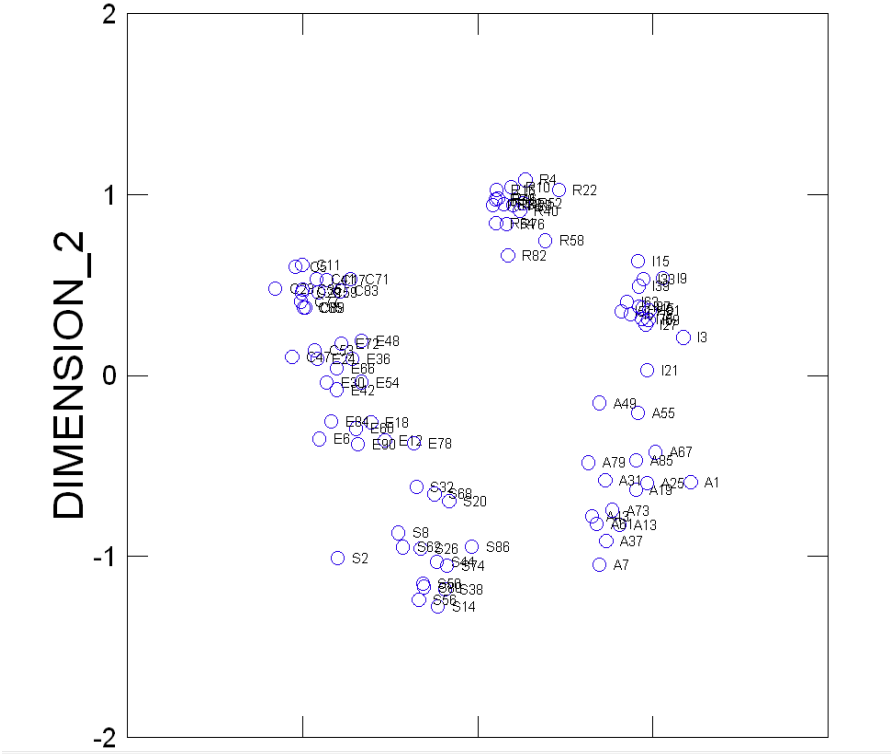
Table 6. Intercorrelations for FYI RIASEC Scales for males

Scale	Realistic	Investigative	Artistic	Social	Enterprising	Conventional
Realistic	--					
Investigative	.15	--				
Artistic	.05	.34	--			
Social	.14	.30	.48	--		
Enterprising	.06	.26	.36	.48	--	
Conventional	.13	.23	.15	.30	.64	--

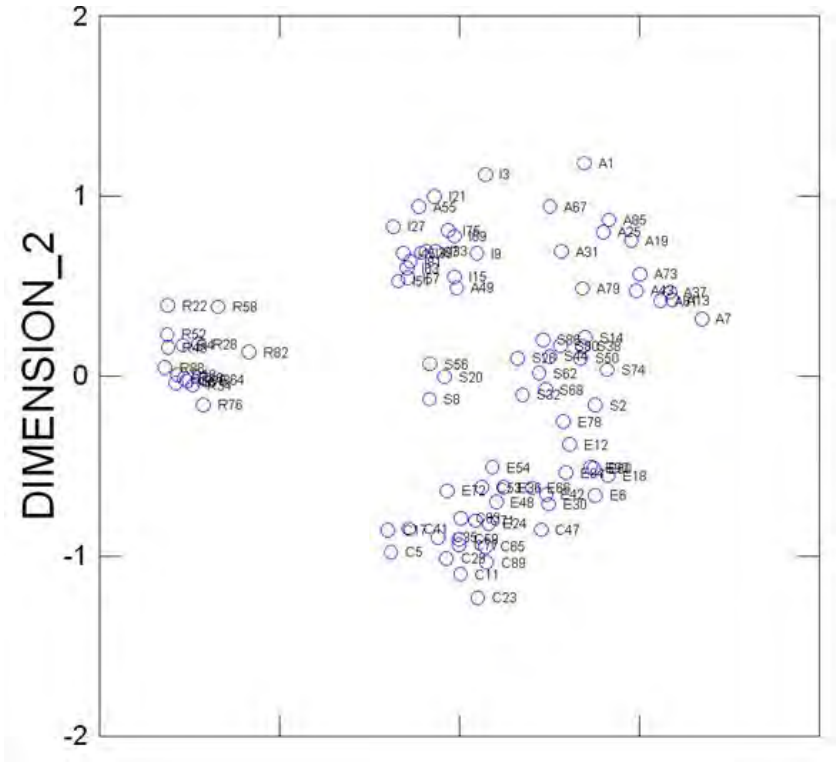


"There is a noticeable difference, with males having a Realistic scale pulled away from the remaining scales and the Social scale slightly closer to the Realistic scale than the Artistic and Enterprising scales. Both configurations have a gap between the Conventional and Realistic scale."

2-D MDS ANALYSIS FROM HumRRO'S REPORT



Full-item raw score two-dimensional multidimensional scaling solution for females.



Full-item raw score two-dimensional multidimensional scaling solution for males.

The report concludes that FYI inventory items poorly fit the RIASEC model for males.

MULTIDIMENSIONAL SCALING

MDS tries to find points that have a given set of pairwise distances. When no set of points satisfies distance constraints, MDS finds the best solution in the least squares sense. In our case, the distances between objects are expressed in the correlation matrix.

obs	R	I	A	S	E	C
R	1	0.28428	0.23027	0.13083	0.16109	0.24562
I	0.28428	1	0.30744	0.12863	0.15648	0.10916
A	0.23027	0.30744	1	0.271	0.24494	-0.02295
S	0.13083	0.12863	0.271	1	0.35775	0.20653
E	0.16109	0.15648	0.24494	0.35775	1	0.59791
C	0.24562	0.10916	-0.02295	0.20653	0.59791	1

Table 1. Correlation coefficients matrix of raw score for Females

R-R	R-I	R-A	R-S	R-E	R-C
0	0.71572	0.76973	0.86917	0.83891	0.75438
0.71572	0	0.69256	0.87137	0.84352	0.89084
0.76973	0.69256	0	0.729	0.75506	1.02295
0.86917	0.87137	0.729	0	0.64225	0.79347
0.83891	0.84352	0.75506	0.64225	0	0.40209
0.75438	0.89084	1.02295	0.79347	0.40209	0

Table 2. Dissimilarity matrix of raw score for Females

obs	R	I	A	S	E	C
R	1	0.14781	0.04599	0.13403	0.05719	0.12446
I	0.14781	1	0.33545	0.29676	0.25698	0.22569
A	0.04599	0.33545	1	0.47391	0.36151	0.14383
S	0.13403	0.29676	0.47391	1	0.47337	0.29438
E	0.05719	0.25698	0.36151	0.47337	1	0.63905
C	0.12446	0.22569	0.14383	0.29438	0.63905	1

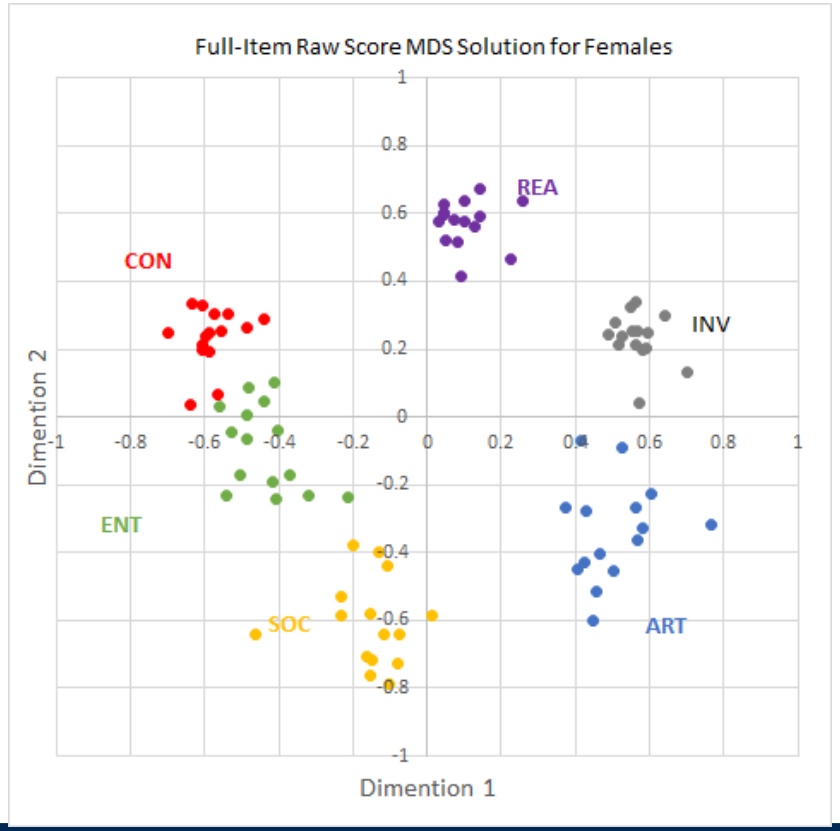
Table 3. Correlation coefficients of raw score for Males

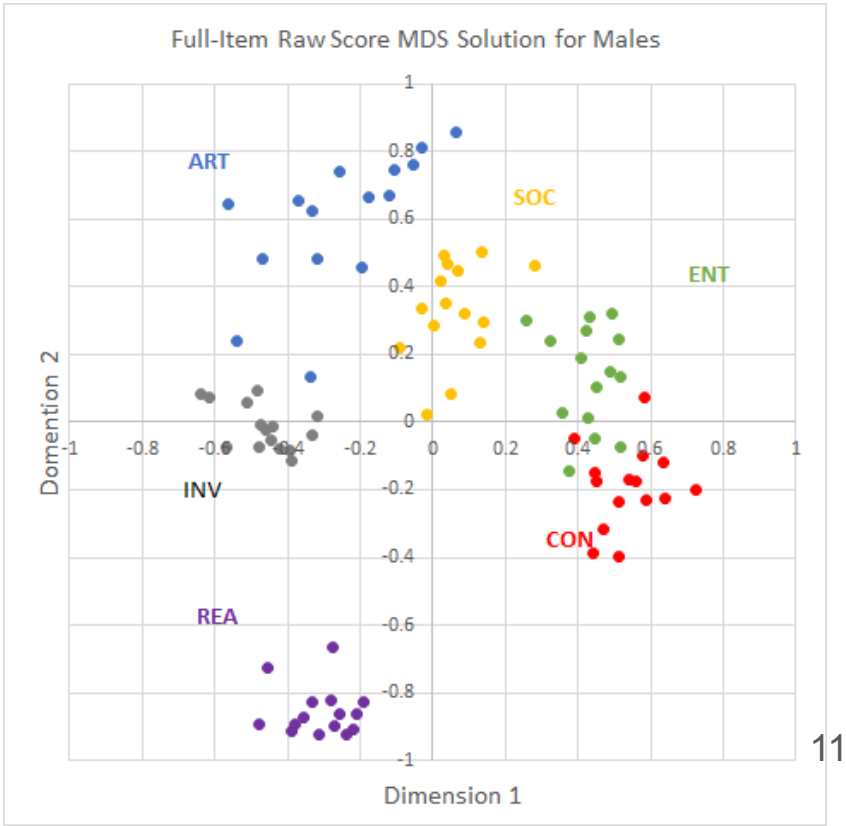
R-R	R-I	R-A	R-S	R-E	R-C
0	0.85219	0.95401	0.86597	0.94281	0.87554
0.85219	0	0.66455	0.70324	0.74302	0.77431
0.95401	0.66455	0	0.52609	0.63849	0.85617
0.86597	0.70324	0.52609	0	0.52663	0.70562
0.94281	0.74302	0.63849	0.52663	0	0.36095
0.87554	0.77431	0.85617	0.70562	0.36095	0

Table 4. Dissimilarity matrix of raw score for Males

MULTIDIMENSIONAL SCALING ANALYSIS BY OPA

- Our goal was to follow the experts panel's recommendation to "use a multidimensional scaling method to achieve a more robust RIASEC model."
  - First, we replicated the 2-D MDS calculations.
  - Each dot on the plots below represents an item.
  - As expected, the dots form 6 clusters.
  - Noticeably, the clusters fail to form a hexagonal structure.
  - But one can think, by removing some items from the Inventory, we may achieve better results.





MULTIDIMENSIONAL SCALING ANALYSIS BY OPA

F:

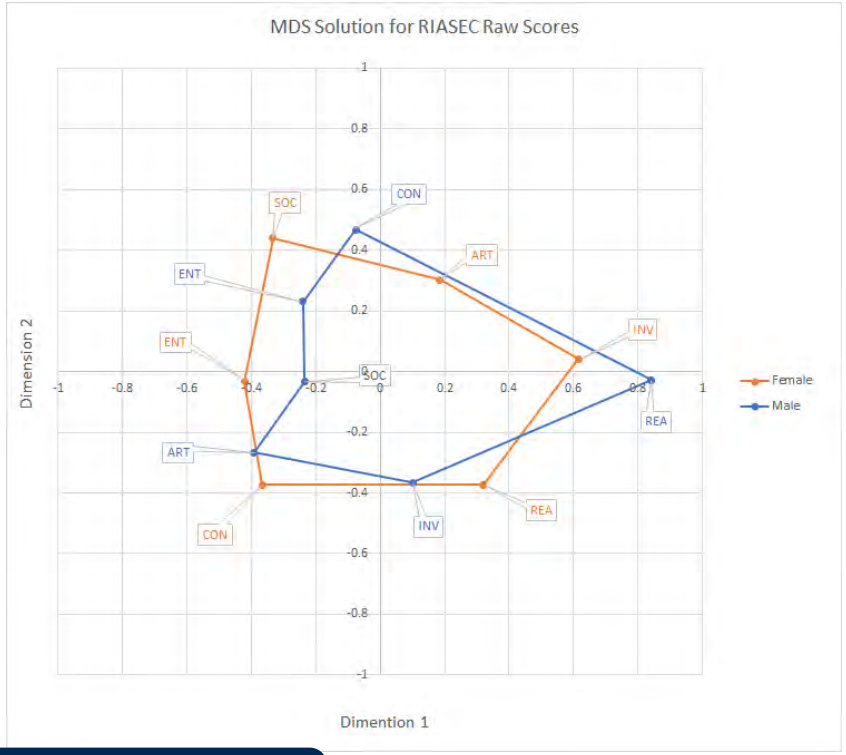
R-R	R-I	R-A	R-S	R-E	R-C
0	0.71572	0.76973	0.86917	0.83891	0.75438
0.71572	0	0.69256	0.87137	0.84352	0.89084
0.76973	0.69256	0	0.729	0.75506	1.02295
0.86917	0.87137	0.729	0	0.64225	0.79347
0.83891	0.84352	0.75506	0.64225	0	0.40209
0.75438	0.89084	1.02295	0.79347	0.40209	0

Dissimilarity matrix of raw score for Females

M:

R-R	R-I	R-A	R-S	R-E	R-C
0	0.85219	0.95401	0.86597	0.94281	0.87554
0.85219	0	0.66455	0.70324	0.74302	0.77431
0.95401	0.66455	0	0.52609	0.63849	0.85617
0.86597	0.70324	0.52609	0	0.52663	0.70562
0.94281	0.74302	0.63849	0.52663	0	0.36095
0.87554	0.77431	0.85617	0.70562	0.36095	0

Dissimilarity matrix of raw score for Males



Is the 2-D MDS fit any good in this case?

F	Dim1	Dim2	distances	
R	0.31941	-0.3742	0	R - R
I	0.61526	0.04063	0.5095205	R - I
A	0.18476	0.30201	0.6894857	R - A
S	-0.33241	0.43884	1.0420669	R - S
E	-0.42054	-0.03464	0.8141419	R - E
C	-0.36648	-0.37263	0.6858918	R - C

Coordinates produced by MDS of raw Scores for Females

M	Dim1	Dim2	distances	
R	0.84133	-0.02905	0	R - R
I	0.10031	-0.36669	0.81432	R - I
A	-0.39136	-0.26787	1.25561	R - A
S	-0.23488	-0.03437	1.07622	R - S
E	-0.23989	0.23093	1.11204	R - E
C	-0.07552	0.46705	1.04246	R - C

Coordinates produced by MDS of raw Scores for Males

MULTIDIMENSIONAL SCALING ANALYSIS BY OPA

Next we considered the norm of the MDS fit error.

F =

R-R	R-I	R-A	R-S	R-E	R-C
0	0.71572	0.76973	0.86917	0.83891	0.75438
0.71572	0	0.69256	0.87137	0.84352	0.89084
0.76973	0.69256	0	0.729	0.75506	1.02295
0.86917	0.87137	0.729	0	0.64225	0.79347
0.83891	0.84352	0.75506	0.64225	0	0.40209
0.75438	0.89084	1.02295	0.79347	0.40209	0

M =

R-R	R-I	R-A	R-S	R-E	R-C
0	0.85219	0.95401	0.86597	0.94281	0.87554
0.85219	0	0.66455	0.70324	0.74302	0.77431
0.95401	0.66455	0	0.52609	0.63849	0.85617
0.86597	0.70324	0.52609	0	0.52663	0.70562
0.94281	0.74302	0.63849	0.52663	0	0.36095
0.87554	0.77431	0.85617	0.70562	0.36095	0

$Fmds$ =

2D MDS distances for Females						
R	I	A	S	E	C	
0	0.509521	0.689486	1.042067	0.814142	0.685892	R
0.5095205	0	0.503637	1.027935	1.038531	1.065175	I
0.6894857	0.503637	0	0.534965	0.692619	0.871209	A
1.0420669	1.027935	0.534965	0	0.481612	0.812185	S
0.8141419	1.038531	0.692619	0.481612	0	0.342286	E
0.6858918	1.065175	0.871209	0.812185	0.342286	0	C

$Mmds$ =

2D MDS dissmilarities for Males						
R	I	A	S	E	C	
0	0.814317	1.25561126	1.076223	1.112037	1.042463	R
0.814317	0	0.50150252	0.472005	0.687667	0.852079	I
1.255611	0.501503	0	0.281084	0.521291	0.799914	A
1.076223	0.472005	0.28108405	0	0.265347	0.526135	S
1.112037	0.687667	0.52129128	0.265347	0	0.287698	E
1.042463	0.852079	0.79991394	0.526135	0.287698	0	C

F – $Fmds$ =

0	0.243701	0.026766	0.098929	0.050111	0.031677
0.2437009	0	0.212361	0.065097	0.112187	0.180715
0.0267657	0.212361	0	0.126438	0.058042	0.077143
0.0989286	0.065097	0.126438	0	0.246268	0.007894
0.0501107	0.112187	0.058042	0.246268	0	0.006219
0.031677	0.180715	0.077143	0.007894	0.006219	0

M – $Mmds$ =

0	0.037873	0.30160126	0.210253	0.169227	0.166923
0.037873	0	0.16304748	0.231235	0.055353	0.077769
0.301601	0.163047	0	0.245006	0.117199	0.056256
0.210253	0.231235	0.24500595	0	0.261283	0.179485
0.169227	0.055353	0.11719872	0.261283	0	0.073252
0.166923	0.077769	0.05625606	0.179485	0.073252	0

F – M =

0	0.13647	0.18428	0.0032	0.1039	0.12116
0.13647	0	0.02801	0.16813	0.1005	0.11653
0.18428	0.02801	0	0.20291	0.11657	0.16678
0.0032	0.16813	0.20291	0	0.11562	0.08785
0.1039	0.1005	0.11657	0.11562	0	0.04114
0.12116	0.11653	0.16678	0.08785	0.04114	0

$\frac{1}{30}\sum_{i,j=1}^{6} ||F_{i,j} - Fmds_{i,j}||$ = 0.1

$\frac{1}{30}\sum_{i,j=1}^{6} ||M_{i,j} - Mmds_{i,j}||$ = 0.15

$\frac{1}{30}\sum_{i,j=1}^{6} ||F_{i,j} - M_{i,j}||$ = 0.11

The difference between Females and Males dissimilarity matrixes is comparable to the errors of the MDS fit.

MULTIDIMENSIONAL SCALING ANALYSIS BY OPA

- In other words, the difference between Females and Males is lost in the errors of the 2-D MDS method.
- 2-D MDS graphs give desirable visualizations of data in some cases (such as distances between points on a map), but in our 5-dimensional case, 2-D MDS gives a rather poor representation of dissimilarities between scales.
- However, we can increase the dimensionality of the method.
  - We calculated the errors for 4 different dimensionalities.
  - The next slide demonstrates that the higher the dimensionality, the lower the error.

MULTIDIMENSIONAL SCALING ANALYSIS BY OPA

5D MDS coordinates for Females					
	Dim1	Dim2	Dim3	Dim4	Dim5
R	-0.14864	-0.378	0.32487	-0.021	0
I	-0.34779	-0.19643	-0.31844	0.176	0
A	-0.40373	0.25143	0.01867	-0.242	0
S	0.08279	0.39598	0.15712	0.275	0
E	0.32884	0.12617	-0.13202	-0.18	0
C	0.48853	-0.19914	-0.0502	-0.007	0

$3D: Fmds =$

3D MDS distances for Females						
R	I	A	S	E	C	
0	0.697478	0.74499	0.825073	0.831218	0.760693	R
0.697478	0	0.56334	0.873216	0.772432	0.878289	I
0.744991	0.563339	0	0.526084	0.758325	1.00194	A
0.825073	0.873216	0.52608	0	0.465768	0.749516	S
0.831218	0.772432	0.75832	0.465768	0	0.371513	E
0.760693	0.878289	1.00194	0.749516	1.00194	0	C

$4D: Fmds =$

4D MDS distances for Females						
R	I	A	S	E	C	
0	0.7247	0.77704	0.876521	0.846317	0.760831	R
0.7247	0	0.701254	0.878824	0.850484	0.896998	I
0.77704	0.701254	0	0.737418	0.760831	1.029215	A
0.876521	0.878824	0.737418	0	0.651147	0.800589	S
0.846317	0.850484	0.760831	0.651147	0	0.410097	E
0.760831	0.896998	1.029215	0.800589	0.410097	0	C

$2D: \frac{1}{30}\sum_{i,j=1}^{6} ||F_{i,j} - Fmds_{i,j}|| = 0.1$

$3D: \frac{1}{30}\sum_{i,j=1}^{6} ||F_{i,j} - Fmds_{i,j}|| = 0.07$

$4D: \frac{1}{30}\sum_{i,j=1}^{6} ||F_{i,j} - Fmds_{i,j}|| = 0.006$

$5D: \frac{1}{30}\sum_{i,j=1}^{6} ||F_{i,j} - Fmds_{i,j}|| = 0.0$

MULTIDIMENSIONAL SCALING ANALYSIS BY OPA

5D MDS coordinates for Males					
	Dim1	Dim2	Dim3	Dim4	Dim5 m5
R	0.64747	0.04156	-0.1277	-0.03621	0.01614263
I	0.05892	-0.23496	0.4132	0.06768	0.01482717
A	-0.19957	-0.36234	-0.12019	-0.1997	-0.03607733
S	-0.14772	-0.11672	-0.22175	0.25099	-0.01838039
E	-0.26671	0.23227	-0.03905	-0.05181	0.10938759
C	-0.09239	0.4402	0.09548	-0.03095	-0.08589999

$2D$: $\frac{1}{30}\sum_{i,j=1}^{6} ||M_{i,j} - Mmds_{i,j}||$ = 0.15

$3D$: $\frac{1}{30}\sum_{i,j=1}^{6} ||M_{i,j} - Mmds_{i,j}||$ = 0.05

$4D$: $\frac{1}{30}\sum_{i,j=1}^{6} ||M_{i,j} - Mmds_{i,j}||$ = 0.007

$5D$: $\frac{1}{30}\sum_{i,j=1}^{6} ||M_{i,j} - Mmds_{i,j}||$ = 0.0

$3D: Mmds =$

3D MDS distances for Males						
R	I	A	S	E	C	
0	0.845829	0.938439	0.816226	0.938059	0.869549	R
0.845829	0	0.606257	0.678117	0.727234	0.761368	I
0.938439	0.606257	0	0.270799	0.603865	0.837897	A
0.816226	0.678117	0.270799	0	0.4115	0.643317	S
0.938059	0.727234	0.603865	0.4115	0	0.302854	E
0.869549	0.761368	0.837897	0.643317	0.837897	0	C

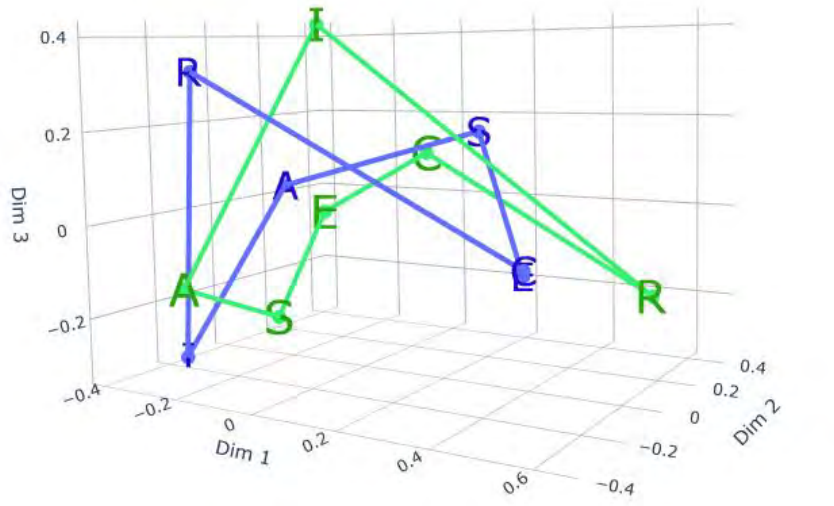
$4D: Mmds =$

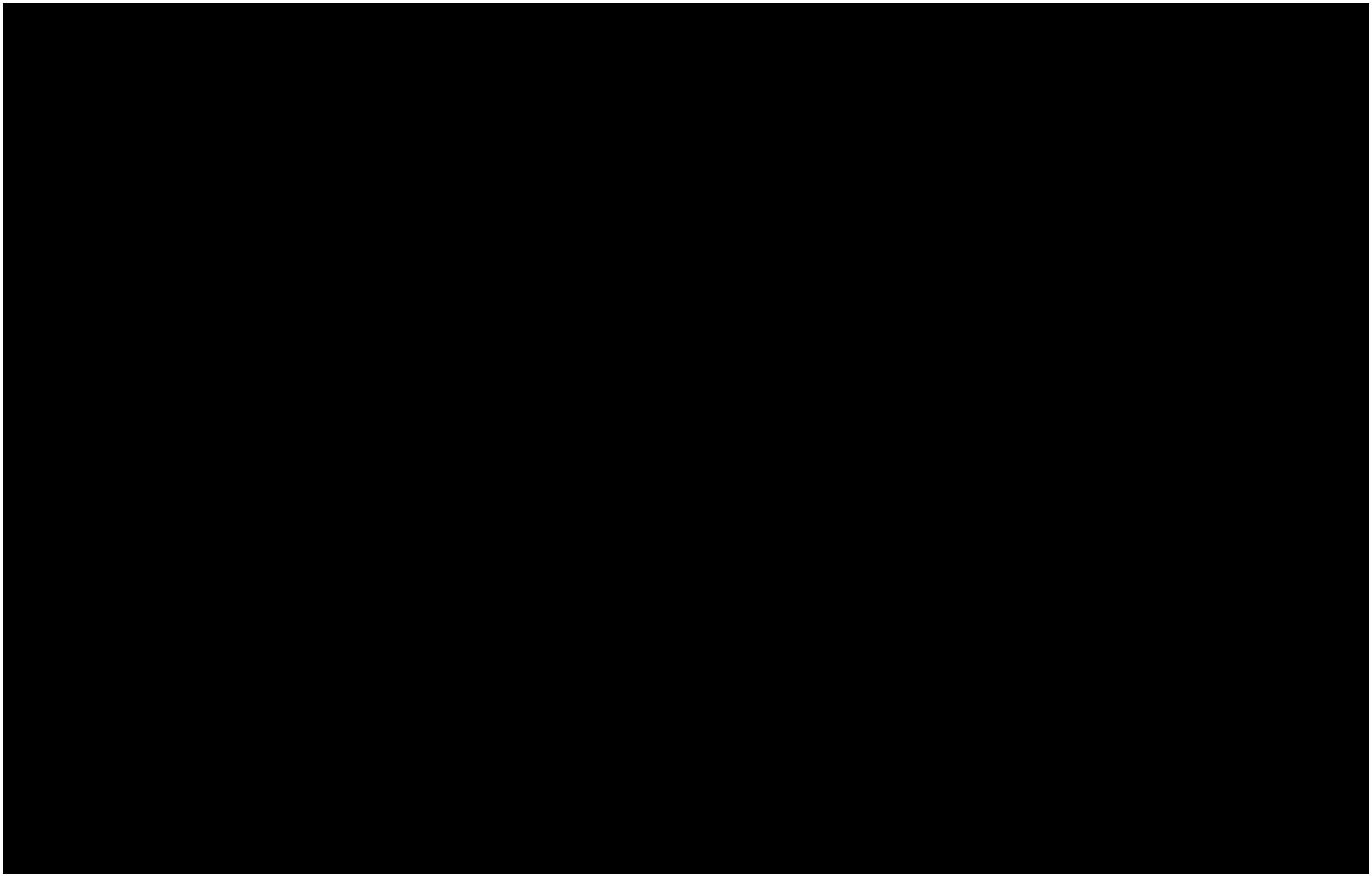
4D MDS distances for Males						
R	I	A	S	E	C	
0	0.852186	0.952574	0.86528	0.938188	0.869565	R
0.852185634	0	0.662601	0.702456	0.736985	0.76773	I
0.952574066	0.662601	0	0.525788	0.621711	0.854721	A
0.865279687	0.702456	0.525788	0	0.510901	0.702386	S
0.938188445	0.736985	0.621711	0.510901	0	0.303572	E
0.869565207	0.76773	0.854721	0.702386	0.303572	0	C

3-D MDS produces smaller errors:

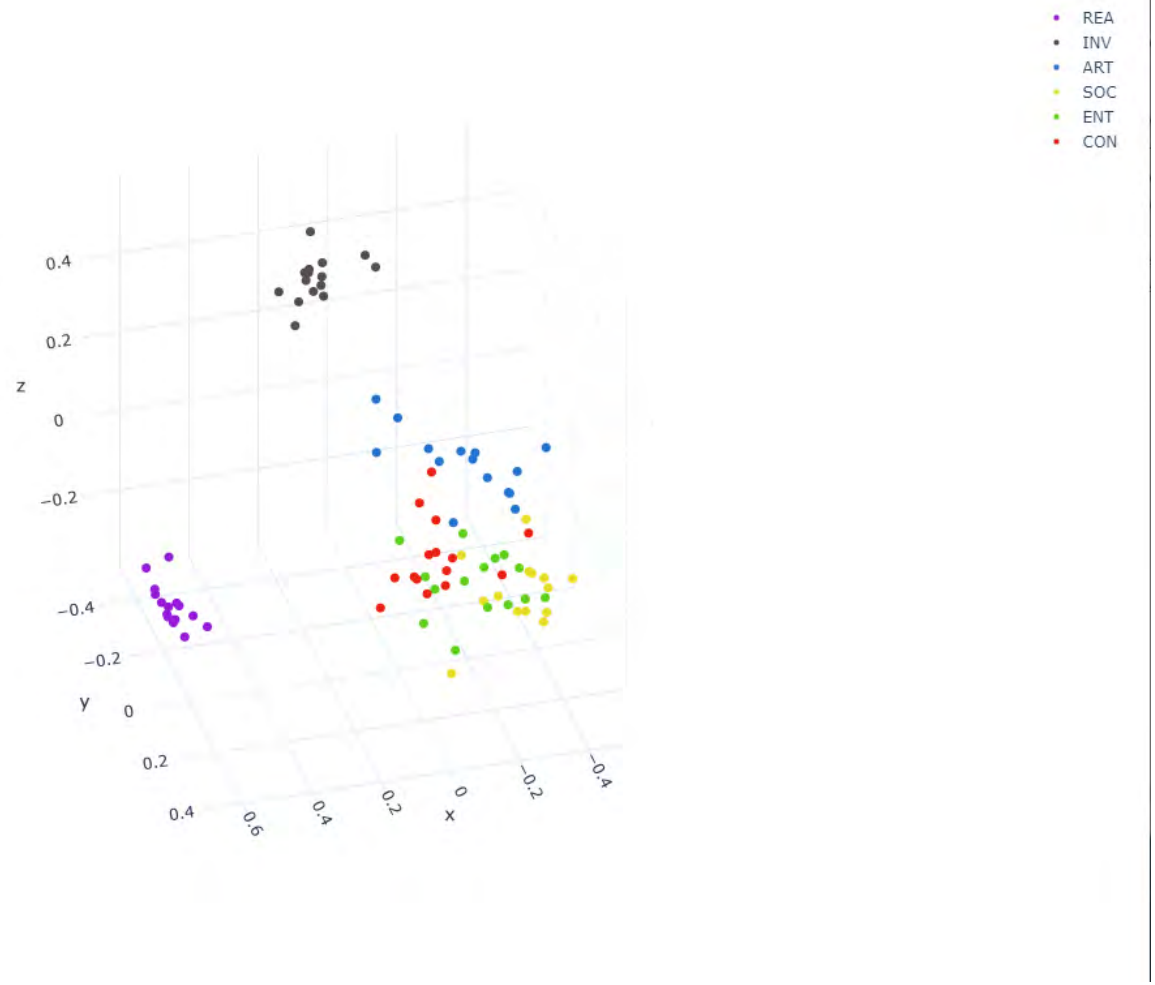
- 30% lower than 2-D MDS for female respondents
- 60% lower than 2-D MDS for male respondents

Next slides demonstrate 3-D graphs of the interest inventory structure.


F-15
M-15
Dim 3
0.4
0.2
0
−0.2
Dim 1
−0.4
−0.2
0
0.2
0.4
0.6
Dim 2
0.4
0.2
0
−0.2
−0.4



RIASEC 90 items

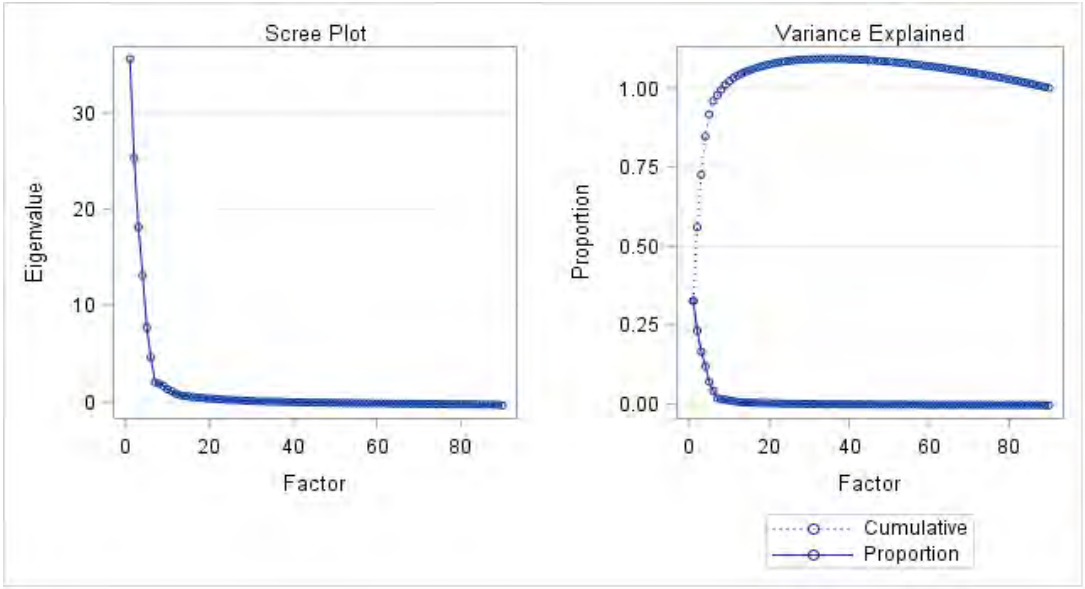


MULTIDIMENSIONAL SCALING ANALYSIS BY OPA

- **Conclusions on MDS analysis:**
  1. Holland's hexagon is a symbolic illustration of the mutual association of the six interest categories.
     - We do observe that mutual associations are consistent with Holland's diagram.
     - The fact that dissimilarities fail to form a 2-D hexagon is not a violation of Holland's model.
  2. 2-D MDS is not a convincing tool for analyzing the quality of the FYI inventory. For this task, it is inconclusive and misleading.

FACTOR ANALYSIS BY OPA

Factor analysis helps us to determine how well the measured variables (90 items) represent the number of categories in the RIASEC representation. Factor analysis aims to find independent latent variables.



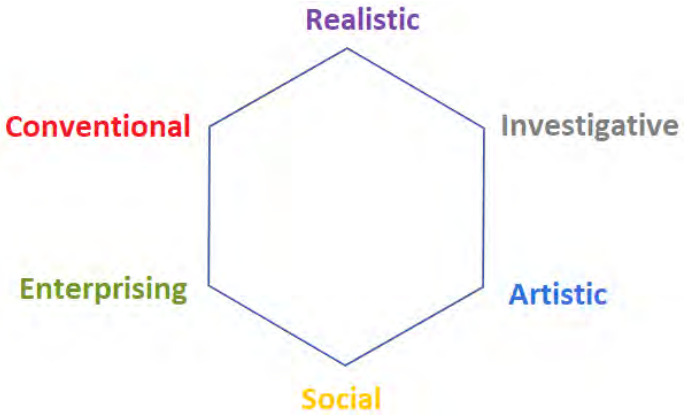
All	Eigenvalue	Difference	Proportion	Cumulative
1	35.5977351	10.305547	0.3271	0.3271
2	25.2921878	7.2016991	0.2324	0.5595
3	18.0904886	4.9471376	0.1662	0.7258
4	13.1433511	5.4459586	0.1208	0.8466
5	7.6973925	3.1050246	0.0707	0.9173
6	4.5923679	2.5689268	0.0422	0.9595
7	2.023441	0.1635895	0.0186	0.9781
8	1.8598515	0.2119057	0.0171	0.9952
9	1.6479459	0.3641429	0.0151	1.0103
10	1.283803	0.1965745	0.0118	1.0221
11	1.0872285	0.2567654	0.01	1.0321
12	0.8304631	0.1650102	0.0076	1.0397

Preliminary Eigenvalues

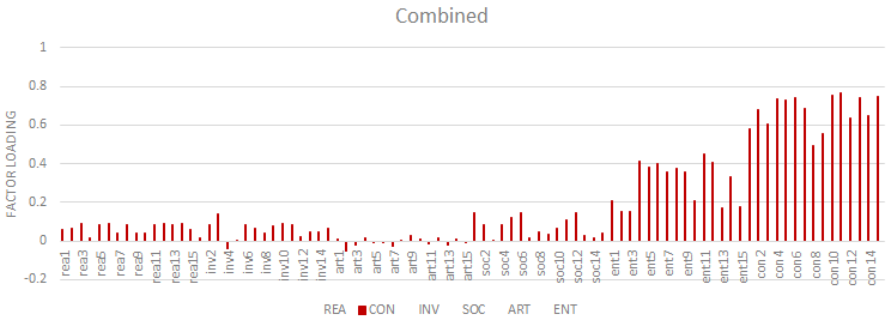
Eigenvalues measure the amount of variation in the total sample accounted for by each factor. If a factor has a low eigenvalue, then it is contributing little to the explanation of variances in the variables and may be ignored as less important than the factors with higher eigenvalues.

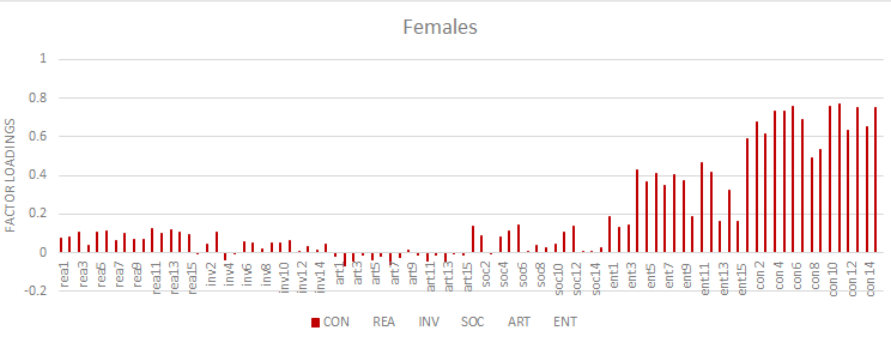
The first 6 largest positive eigenvalues account for 96% of the common variance. The analysis suggests that 6 factors are present.

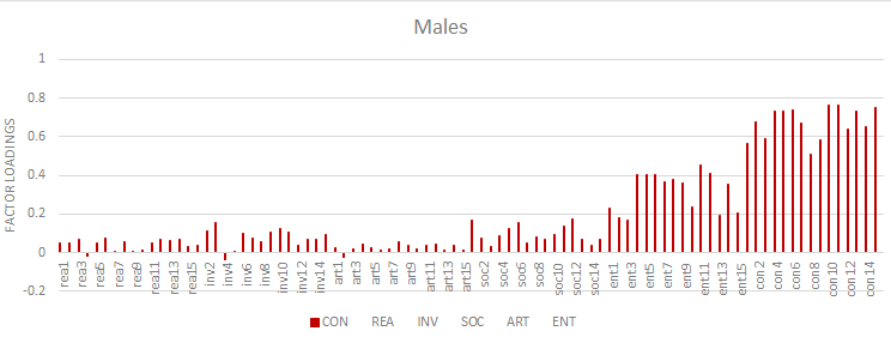
FACTOR ANALYSIS BY OPA



Bar-plots of factor loadings for Conventional items







The results of the factor analysis allow us to conclude that:

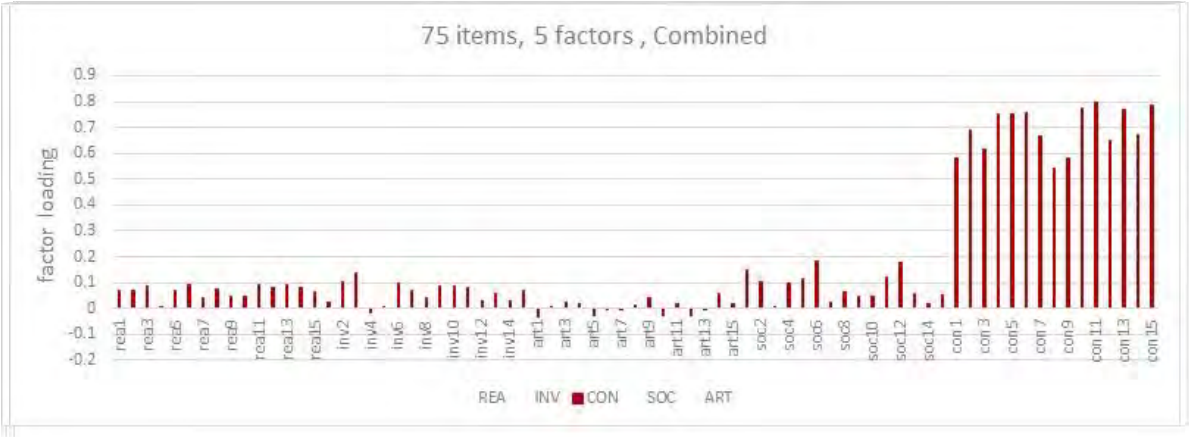
1. There are 6 clearly isolated well-defined factors.
2. The *Enterprising* category demonstrates excessively strong cross-loadings with the *Conventional* domain.
3. The bar-plots for male and female factor loadings are not identical but fairly similar.

Factor loadings for each scale are presented for clarity.

FACTOR ANALYSIS BY OPA

Since the previous analysis demonstrated an excessive overlapping between the last two scales, we decided to run an experiment: We removed ***Enterprising*** items in one case and ***Conventional*** items in another, and we re-ran the factor analysis with 5 factors.

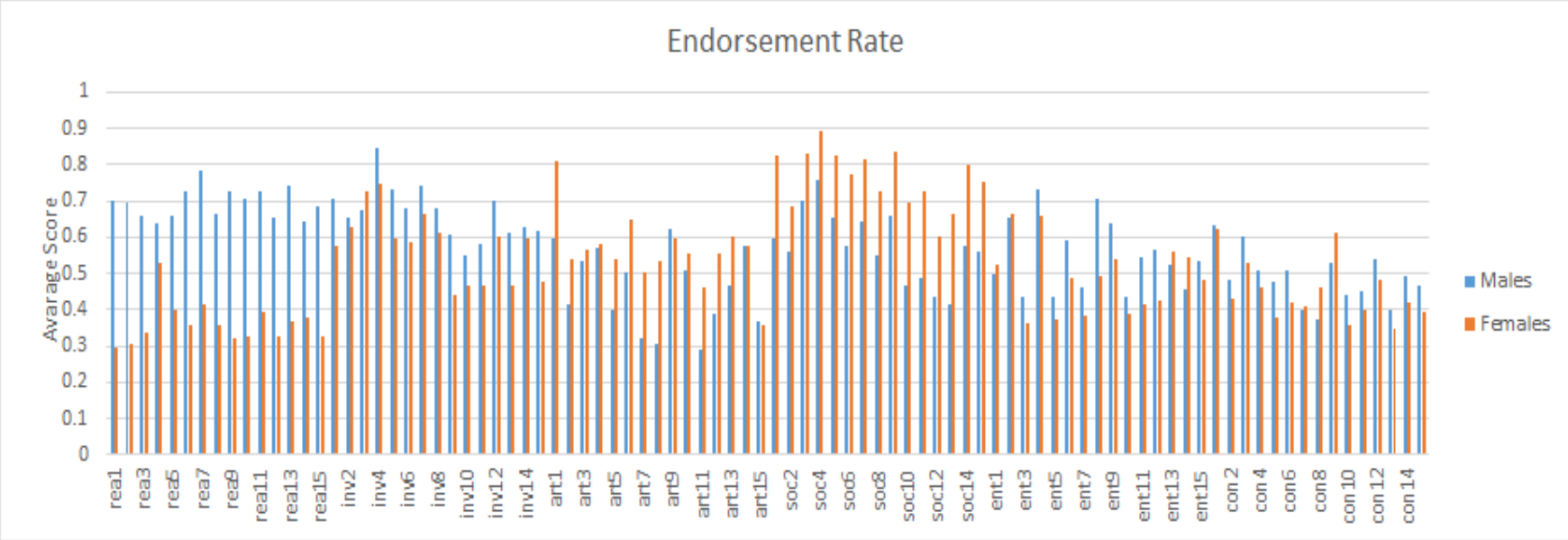
Factor loadings for 75 items in ***Realistic, Investigative, Artistic, Social, Conventional***



Factor loadings became remarkably well distinct.

Factor loadings for 75 items in ***Realistic, Investigative, Artistic, Social, Enterprising***

ENDORSEMENT RATE



- Although the male/female factor structure is very similar, we do see somewhat different endorsement rates across males and females for Realistic items (but not for the other domains).

- Thus, we can conclude that separate norms may not be needed for males and females except, perhaps, for the Realistic domain, and that we might want to revisit the Realistic items to see whether we can reduce the differential endorsement rates that are currently exhibited.

CONCLUSIONS

- 2-D MDS does not seem to be a proper tool for evaluation of the FYI inventory.
- Factor analysis shows that the FYI inventory satisfactorily represents Holland's RIASEC structure.
- Factor analysis suggests fairly little difference between females and males.
- Based on the endorsement rate analysis, we may want to revisit the Realistic items to see whether we can reduce the differential endorsement rate that is currently exhibited.
- The items in the Enterprising category appear in need of improvement in terms of fit to the RIASEC model.
- Separate norms may not be needed for males and females.

Tab L

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

ASVAB Validity Argument Briefing

September, 2019

For more information,
please contact :

Arthur Thacker

66 Canal Center Plaza, Suite 700, Alexandria, VA 22314-1578 | Phone: 703.549.3611 | Fax: 703.549.9025 | www.humrro.org

Presentation Overview

- Overview of validity argument approach to validation
- Applying the validity argument approach to validation of AFQT and ASVAB
- Theory of action (TOA) drafts for AFQT
- Draft claims structures (interpretive argument) for AFQT
- Specific validity evidence
- Next steps
- Challenges associated with collecting and categorizing validity evidence for ASVAB

Validity Argument Overview

- The validity of an assessment cannot be summarized via a single statistic or coefficient. Validation depends on the assessments purpose, the inferences made from assessment results, and the uses of those results.
- Argument-based validation tests the underlying claims that must be true to support the inferences made from assessment information (scores).
- An assessment score may be valid for multiple purposes.
- When assessments are used for multiple purposes, it is rare that the assessment is equally valid for all of them.

Tiered Evidence

- Evidence is collected for a validity argument to support claims in a chain of reasoning, where any claim in the chain found to be weak may undermine subsequent claims.
  - Example 1—Poor item quality can undermine all results from an assessment
  - Example 2—Even if all aspects of a test seem supported and strong validity evidence for use of scores is available for a given year, poor year-to-year equating can undermine cross-year comparisons of scores.
- If multiple inferences are drawn from a single assessment score (or event), each inference may have its own unique validity argument.

Where Do We Start

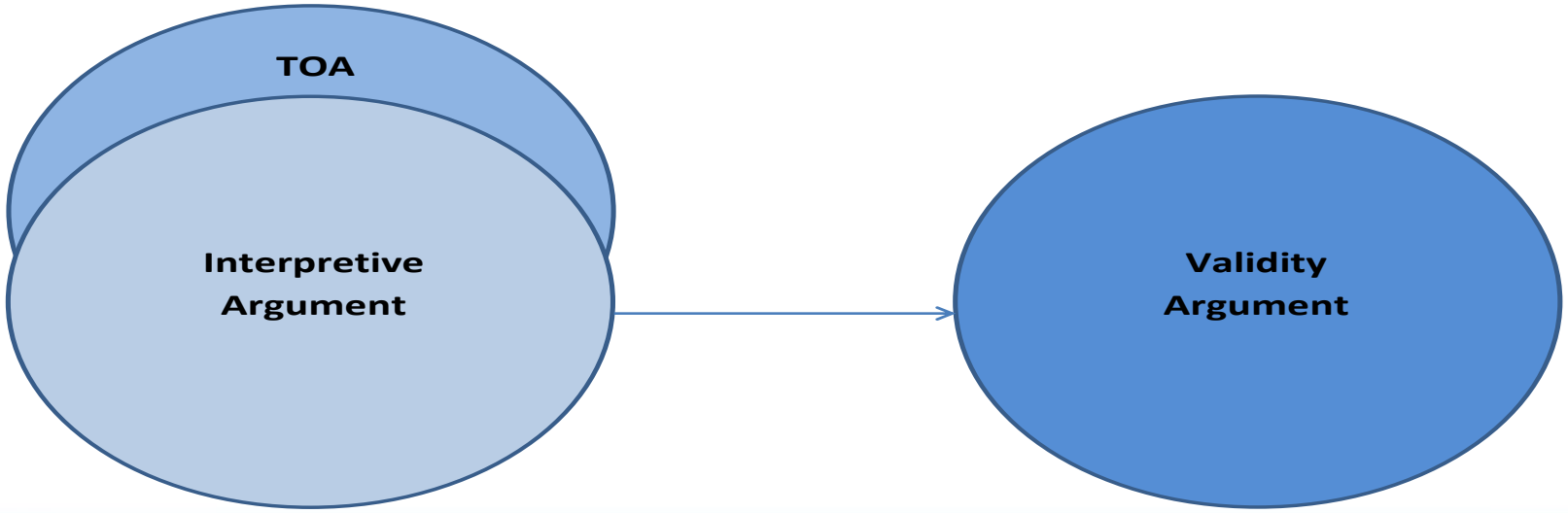
- What are the most important inferences we want to make?
  - Admission into military branches (current AFQT focus)
  - Placement into training programs or advanced educational opportunities (focus of ASVAB)
- Establishing Draft TOAs for ASVAB
  - ASVAB primarily relies on an informal reasoned approach
  - Evidence is not currently tied to organized claims or assumptions
  - A TOA (or similar) is required to frame interpretive and validity arguments
- Bounding the Argument (Limitations)
  - Will not address admittance to specific training programs or MOS (each would require its own body of evidence which is beyond the scope here)

Validity Argument Illustration

TOA—Theory of Action (all the things the test and test scores are expected to be used for and the expected advantages of using the test for those purposes)

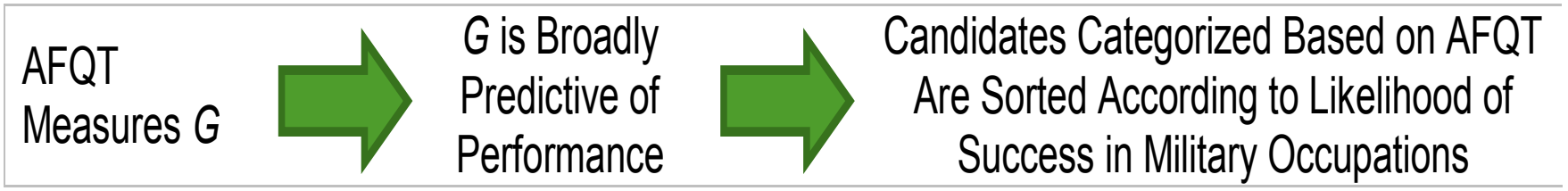
Interpretive Argument—a description of the inferences that the test scores support

Validity Argument—evidence providing justification for the inferences in the interpretive argument.



Draft AFQT Theory of Action

The AFQT measures *G*, and because *G* is predictive of a broad range of future performance, the AFQT will broadly predict candidates' success in military occupations.



We can then develop claims that must be supported for each step in the TOA to be true.

AFQT Assumptions

I. AFQT Subtests Measures G

1. G is a broad stable construct underlying cognitive test scores.
2. Each of the four AFQT scores is a reliable measure of its intended construct.
3. The four AFQT subtests are the best options for a G proxy.

II. Derivation of the AFQT Category Scores Supports Their Use for Recruit Selection

4. AFQT scores measure G in the intended population.
5. The predictive nature of G is continuous for nearly the full scale of the AFQT (i.e. higher scores always yield a better predicted outcome, irrespective of the area of the scale the score falls in).
6. The AFQT categories represent important differentiators among applicants.

AFQT Assumptions (continued)

7. AFQT scores have high overall reliability and lower error, especially near the cut points for the categories.
8. AFQT scores have high classification accuracy.
9. AFQT scores are unbiased with regard to race/ethnicity, gender, etc.

III. G and the AFQT Predict Important Training and Job Performance Criteria

10. AFQT is a measure of G, so AFQT scores should demonstrate a pattern of correlations with different types of job and training performance criteria similar to G's predictive pattern.
11. AFQT category scores are associated with important outcomes.

AFQT Assumptions (continued)

IV. G and the AFQT Scores Yield Similar Patterns for Subgroup Differences

12. G and the AFQT scores yield similar patterns for subgroup differences.

V. Contextual Factors Support the Use of the AFQT

13. Users of the AFQT scores and/or the AFQT category scores understand the meaning and use/outcome of the scores.
14. The ASVAB is administered appropriately.

VI. Candidates Scores are interchangeable irrespective of the Version of the AFQT they Take

15. The paper-and-pencil and CAT versions of the AFQT yield interchangeable scores.
16. Unproctored verified and proctored versions of the AFQT yield interchangeable scores.
17. AFQT delivery on other devices (e.g. tablets, cell phones) yield interchangeable scores compared to proctored CAT versions.

AFQT Draft Validity Argument Excerpt

Assumption	Claim	Summary of Evidence (with links to write-up)
I. AFQT Subtests Measure G		
I.1. G is a broad, stable construct underlying cognitive test scores.	I.1.a. Scores on tests of mathematics and verbal skills are generally accepted measures of G.	• G is a broad general factor underlying cognitive test scores (Humphreys, 1979; Hunter, 1986; Jensen, 1986; Spearman, 1927). • Tests of verbal and mathematical skills are measures of G to the extent that they rely more heavily on reasoning ability than accumulated knowledge (Carroll, 1993; Johnson & Bouchard, 2005a, 2005b; Vernon, 1965).
	I.1.b. G is relatively stable over time.	• Cognitive ability test scores are typically highly stable over time as evidenced by high test-retest reliability (Cronbach, 1984; Eichorn, Hunt, & Honzik, 1981; Honzik, MacFarlane, & Allen, 1948; Humphreys, 1986). • Scores on cognitive ability tests improve upon retesting (Cronbach, 1984; Vernon, 1979) and practice (Kulik, Kulik, & Bangert, 1984), although the magnitude of the gain varies with the type of test and the retest interval. • ***"The higher a test's "g" loading, the less susceptible it is to a practice effect" (Jenson, 1998, p. 315).*** • Test preparation results in modest gains in scores beyond that provided by retest alone (Kulik, Bangert-Drowns, and Kulik (1984). • Retesting does not appear to result in changes in "g" (Reeve & Lamm 2005).

AFQT Draft Validity Argument Excerpt (Cont.)

Assumption	Claim	Summary of Evidence (with links to write-up)
I. AFQT Subtests Measure G		
I.2. Each of the four AFQT subtests is a reliable measure of its intended construct.	I.2.a. AFQT subtest items reflect well-defined **content domains** that are appropriate for the (a) construct, (b) intended population and (c) purpose.	I.2.a.1. • The taxonomic definitions of AFQT subtests and item writing guidelines (DPAC, 2014a, 2014b, 2014c, 2016) indicate that the subtests are designed to measure verbal and mathematics reasoning skills that typically load strongly on G in factor analytic research (Carroll, 1993; Johnson & Bouchard, 2005a, 2005b).
		I.2.a.2. • The AFQT subtest item development process is consistent with professional guidance (AERA, APA, & NCME, 2014; Lane, Raymond, Haladyna, & Downing, 2016; Wise & Plake, 2016). • DPAC provides recent, thorough item writing guidelines for each AFQT subtest (DPAC, 2014a, 2014b, 2014c, 2016) that are consistent with professional guidance on item writing (Haladyna & Rodriguez, 2013). • DPAC item writing guidelines (DPAC, 2014a, 2014b, 2014c, 2016) and test development procedures (HumRRO, 2019) address fairness and bias in order to minimize construct irrelevant variance in item content (Ziecky, 2006). Content specifications for ASVAB subtests have been updated in response to data and other evidence (Waugh, Knapp, Ramsberger, & Caramagno, 2015).
		I.2.a.3. • Subtest items are drawn from the item pool to adequately (a) sample underlying domains and (b) create interchangeable forms.

Links from Validity Argument

- Each claim will include a link to more specific documentation about that claim.
- All claim links are organized using a common structure
- Links include:
  1. Restatement of the claim
  2. Evidence categories (main headings for organizing evidence)
  3. Summary for Validity Argument (repeats summary)
  4. Literature review
  5. Reference list

Excerpt from Claim Link

- **Claim**
- I.1.b. G is relatively stable over time.
- **Evidence**
- I.1.b.1. Longitudinal studies of G; test-retest and alternate forms reliabilities for measures of G
- **Summary for Validity Argument**
- Cognitive ability test scores are typically highly stable over time as evidenced by high test-retest reliability (Cronbach, 1984; Eichorn, Hunt, & Honzik, 1981; Honzik, MacFarlane, & Allen, 1948; Humphreys, 1986).
- Scores on cognitive ability tests improve upon retesting (Cronbach, 1984; Vernon, 1979) and practice (Kulik, Kulik, & Bangert, 1984), although the magnitude of the gain varies with the type of test and the retest interval.
- *"The higher a test's "g" loading, the less susceptible it is to a practice effect" (Jenson, 1998, p. 315).*
- Test preparation results in modest gains in scores beyond that provided by retest alone (Kulik, Bangert-Drowns, and Kulik (1984).
- Retesting does not appear to result in changes in "g" (Reeve & Lamm 2005).

Excerpt from Claim Link (continued)

Literature Review

- ***Cognitive abilities tend to be highly stable over time.***

A large body of research suggests that intelligence and cognitive ability test scores are highly stable over time (Cronbach, 1984; Eichorn, Hunt, & Honzik, 1981; Honzik, MacFarlane, & Allen, 1948; Humphreys, 1986), as evidenced by high test-retest reliability (i.e., stability coefficients). The GATB (Department of Labor, 1970) provides a useful example. High school freshmen, sophomores, and juniors were tested and then retested when they were seniors; there were about 7,000 participants in each group. The test-retest reliabilities were highest for juniors (who had been retested after only one year)—ranging from .74 for spatial aptitude to .83 for general intelligence. For those who took the test for the first time as sophomores, coefficients ranged from .73 to .82, and for those who were freshmen on first testing, coefficients ranged from .70 to .79. As Peterson, Hanson, and Wolfe (1996) noted, the age at initial testing is confounded with the length of the test interval making it impossible to disentangle the two sources of attenuation. Even so, the data demonstrate reasonably good stability over time for the GATB tests in spite of maturational and educational influences.

Excerpt from Claim Link (continued)

References

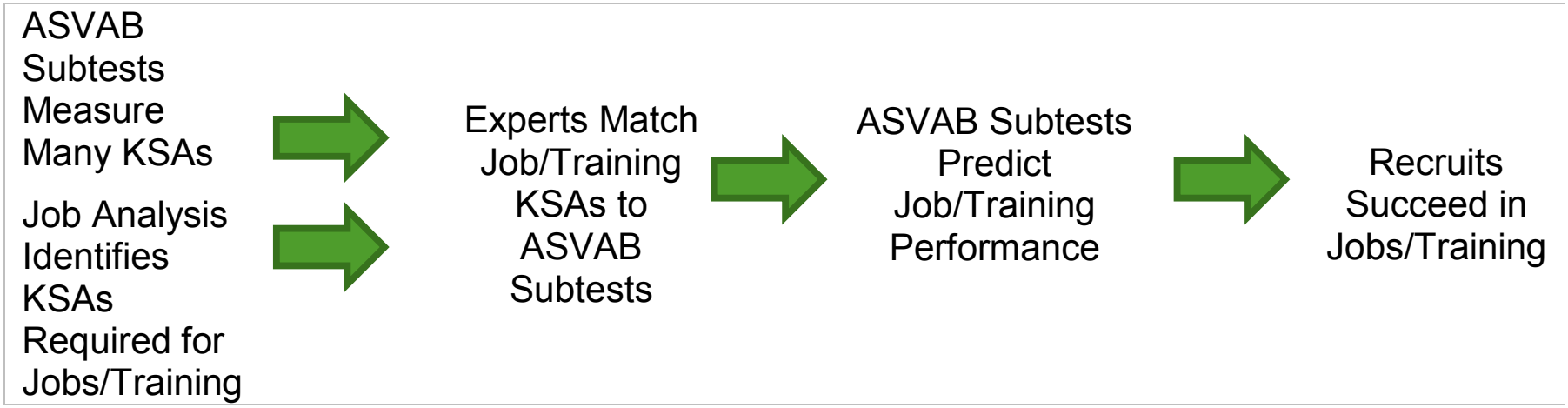
- Johnson, W., & Bouchard Jr., T. (2005a). Testing the grand old models of the structure of human intelligence: It's verbal, perceptual, and visualization (VPZ), not fluid and crystallized. *Intelligence. 33*, 393-416.
- Johnson, W., & Bouchard Jr., T. (2005b). Constructive replication of the visual-perceptual-image rotation model in Thurstone's (1941) battery of 60 tests of mental ability. *Intelligence. 33*, 417-430.
- Kyllonen, P. C., & Christal, R. E. (1990). Reasoning is (little more than) working memory capacity. *Intelligence, 14*, 389-433.
- Linn, R. L. (1986). Comments on the *g* factor in employment testing. *Journal of Vocational Behavior, 29*, 340-362.
- Ree, M. J., & Earles, J. A. (1991, April). *Estimating psychometric g: An application of the Wilk's theorem*. Paper presented at the annual meeting of the American Psychological Association, San Francisco, California.
- Spearman, C. (1927). *The abilities of man*. New York: MacMillan Co.
- Vernon, P. E. (1950). *The structure of human abilities*. London: Methven.

Next Steps

1. Revise TOAs to better reflect the logic model underlying AFQT and ASVAB (Iterative)
2. Define/revise assumptions associated with the revised TOA
3. Develop/revise specific claims that support the assumptions
4. Indicate the required evidence necessary to support validity claims
5. Reference evidence for specific validity claims from the literature and from ASVAB documentation (e.g. technical manuals)
6. Identify evidence gaps or weaknesses and commission analyses/studies to address them
7. Maintain and update validity argument as necessary

Draft ASVAB Technical Test Theory of Action #1

Job Analysis Model



Model relies on clear linkages between KSAs required for military jobs/training and KSAs measured by ASVAB.

Draft ASVAB Technical Test Theory of Action #2

Why ASVAB Predicts Success

- AFQT and ASVAB measure G
- ASVAB technical tests may reflect course-taking patterns or life experiences of candidates (e.g. a candidate who worked in automotive repair might score better on AI or SI sub-tests)
- ASVAB technical tests may act as an interest inventory (e.g. candidates who spend time on an interest would be expected to score better on associated sub-tests)
- Preparation may improve sub-test scores, and could be associated with conscientiousness/motivation (e.g. practice tests, test prep courses)

Unvalidated Potential Uses of ASVAB

- Inclusion on state summative assessments as an indicator of student readiness for careers
- School-level accountability measure
- As alternate evidence of high school preparation (for students who do not pass the state's graduation assessment)
- Alternative language versions of the ASVAB as a better G measure for non-native English speakers

Ongoing Challenges

1. Lack of models from comparable assessment systems
2. 50 years of history
3. ASVAB was not created using Evidence Centered Design (ECD) or a similar approach based on claims and inferences
4. Multiple users
5. Varied score information
6. Multiple inferences need to be supported
7. Discerning which ASVAB literature is relevant for the validity argument is not always straightforward
8. ASVAB literature is not always unbiased

Recommendations Preview

1. Establish an AFQT/ASVAB technical manual
   a) Start from the “ASVAB Standards” from the field test checklist to create categories for the technical manual.
   b) Indicate how often each type of evidence included in the technical manual should be updated.
   c) Update data tables only unless a substantive change is made in the assessment
2. Estimate classification accuracy at the selection decision cut score
3. Investigate comparability of devices as new technology becomes available (phones, tablets, etc.)

Thank you!

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

Tab M

DOD-wide First-Term Enlisted Servicemember Criterion Measurement

Presenter : Dr. Laura Ford

September 27, 2019

Headquarters: 66 Canal Center Plaza, Suite 700, Alexandria, VA 22314-1578 | Phone: 703.549.3611 | www.humrro.org

Overall Project Objective

To document the test evaluation criteria unique to each Service's accession and classification testing efforts, and to identify and / or develop a unified set of criterion instruments which can be used by all Services.

- Document current practices
- Standardize criteria where possible and document remaining differences
- Standardizing measures across Services is desirable to facilitate (a) interpretation of validation studies and (b) generalizability of results.

Overview

- Organized the Criterion Measures Advisory Panel (CMAP)
- Developed taxonomy to define job domain of first-term, enlisted Servicemembers (Slides 5-14)
- Documented existing criterion measures (Slides 15-19)
- Recommended development of a set of Service-wide test evaluation criteria to be used for validation research (Slides 20-22)
- Current Status of Measures Development (Slides 23-27)
- Questions/Discussion

Criterion Measures Advisory Panel (CMAP)

- Air Force – Dr. Tom Carretta (COR), Mr. Johnny Weissmuller
- Army – Dr. Cristina Kirkendall (Technical Advisor)
- Coast Guard – Dr. Donna Duellberg
- Marine Corps – CPT Alex Ryan, Dr. Eric Charles
- Navy – CDR Henry Phillips, Mr. Tom Blanco
- Office of the Under Secretary for Personnel and Readiness (OUSD P&R) – Dr. Sofiya Velgach

Developed Taxonomy - Three Criterion Domains

- **Job performance** – individuals' behaviors that are relevant to the Services' goals and that can be scaled in terms of each individual's proficiency
  - Four factors at the highest level
    - A. Technical proficiency
    - B. Organizational citizenship & peer leadership
    - C. Psychosocial Well-being
    - D. Physical Performance
  - 12 mid-level performance dimensions
  - 33 sub-dimensions

Early career, enlisted job performance in training thru the end of the 1st term

- **Attitudes** – cognitions that are relevant to individuals' job plans and performance (e.g., commitment, satisfaction, career intentions).
- **Organizational outcomes** – outcomes that are important to the Services at an organizational level such as reducing attrition and enhancing reenlistment.

Job Performance Taxonomy Development

Develop a job performance taxonomy to describe the domain of early career, enlisted DOD servicemember job performance

Accomplished by:

1. Review of extant literature
2. Initial taxonomy development
3. Retranslation exercise
4. Finalize taxonomy and definitions
5. Subject Matter Expert (SMEs) evaluation

1 & 2: Initial Taxonomy Development

- Objectives
  - Comprehensiveness, Efficiency, Hierarchical, Relevance
- Literature review **Full references provided in supporting slides
- 8-dimension Campbell (2012) model served as a scaffold, then incorporated content from other models
  - (1) Job-specific and (2) non-job-specific technical task proficiency
  - (3) Written and oral communication task proficiency
  - (4) Demonstrating effort
  - (5) Maintaining personal discipline
  - (6) Facilitating peer and team performance
  - (7) Supervision/leadership and (8) management/administration
- Process resulted in 33 draft dimensions

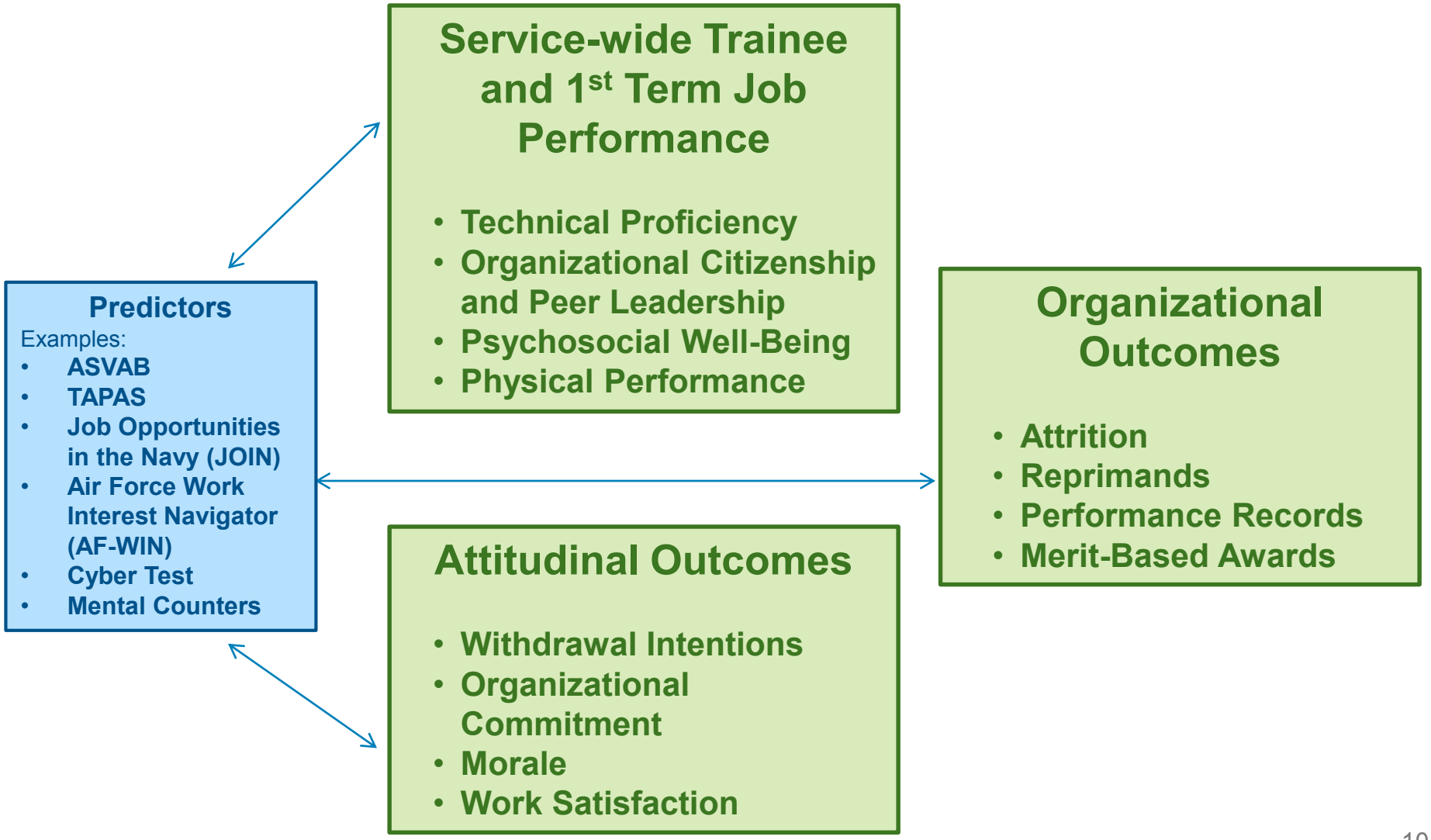
3: Retranslation

- Purpose
  - Evaluate clarity of 33 performance dimensions
  - Determine appropriate hierarchical structure
- Procedure
  - 17 researchers with extensive experience in performance measurement and/or military criterion development
  - Structures evaluated:
    - 2-dimension – Task (Can-do) vs. Contextual (Will-do)
    - 4-dimension – Technical, Counterproductive Work Behavior, Citizenship & Peer Leadership, Physical
    - 10-dimension
  - Rated (2-dimension) or sorted (4, 10-dimension)
  - Reliability / agreement high
    - ICC(C,*k*) = .95
    - Mean $\%_{agree}$ = 86% (10-dimension), 88% (4-dimension)

3: Retranslation

- Results
  - 2-dimension structure did not work well, but 4- and 10-dimension did
  - Used ratings to make refinements, for example:
    - Used 2-dimension ratings to help with sorting decisions
    - Moved some dimensions to other categories based on ratings
    - Refined definitions where there was a lack of clarity
    - Dimensions that did not sort well into any of the 10 dimensions were broken into their own
  - Final model had three levels:
    - 4 Categories
      - 12 Dimensions
        - » 33 Sub-Dimensions

4. Finalize Taxonomy



5: SME Evaluation

- Purpose was to evaluate:
  - *Generalizability* – the extent to which the constructs measured are important DOD-wide.
  - *Relevance* – the extent to which the constructs measured are relevant to the organization's occupations and goals.
- Procedure
  - Asked military experts with broad knowledge of service job requirements to rate the following for each sub-dimension:
    - Importance across enlisted, first term occupations
    - Criticality to service's mission accomplishment
  - Received sufficient data from four Services to proceed with analysis (Army, Navy, Air Force, Marine Corps)
    - Interrater reliabilities (ICC[C,*k*]) within service ranged from .73 to .98

5: SME Evaluation

Trainee and 1st Term Enlisted Servicemember Performance:
A Survey of the Importance and Criticality of Performance Dimensions

46% Completed

Navigation Instructions | Continue Later

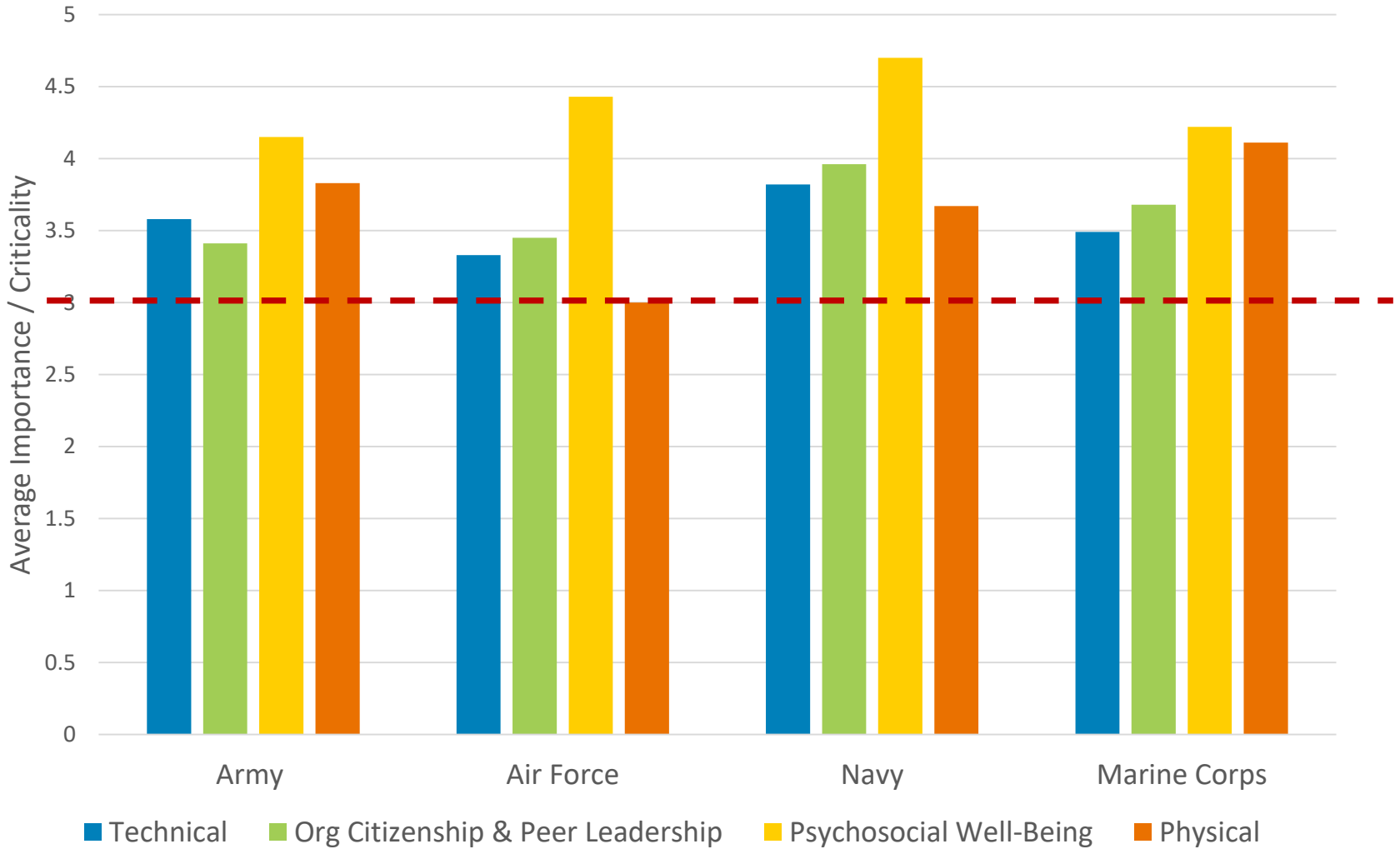
1. Background | 2. Technical Performance

Please rate the importance and criticality of each element.

	How important is this performance element across enlisted, first-term occupations?					To what extent is successful enlisted, first-term performance in this element critical to your service's mission accomplishment?				
	Not important	Somewhat important	Important	Very important	Extremely important	Not at all critical	Somewhat critical	Critical	Highly critical	Extremely critical
a) **Job-Specific Proficiency** - Being able to perform job-specific tasks at the appropriate skill level.	○	○	○	○	○	○	○	○	○	○
b) **General Proficiency** - Being able to perform service-wide tasks at the appropriate skill level (e.g., navigation in the Army and Marine Corps).	○	○	○	○	○	○	○	○	○	○
c) **Oral Communication** - Conveying information in a clear, understandable, organized manner when speaking.	○	○	○	○	○	○	○	○	○	○

5: SME Evaluation

- High level summary of results:



5: SME Evaluation

- Summary of results
  - With one exception, broad performance factors were relevant and generalizable across services
  - Eight of 33 performance sub-dimensions had an average importance / criticality rating less than 3.0
  - Psychosocial Well-Being rated highest across services (i.e., most generalizable)
  - Physical Performance sub-dimensions rated least highly, also the most variable across services
  - Organizational Citizenship / Peer Leadership and Technical Performance rated comparably to one another overall, but there was substantial variation across elements
    - e.g., job-specific proficiency and safety rated highly, nonverbal and written communication rated lower in the Technical Performance category
- Conclusion
  - 4-dimension and 11-dimension levels relevant and generalizable across services

Collect Information on Current Criteria

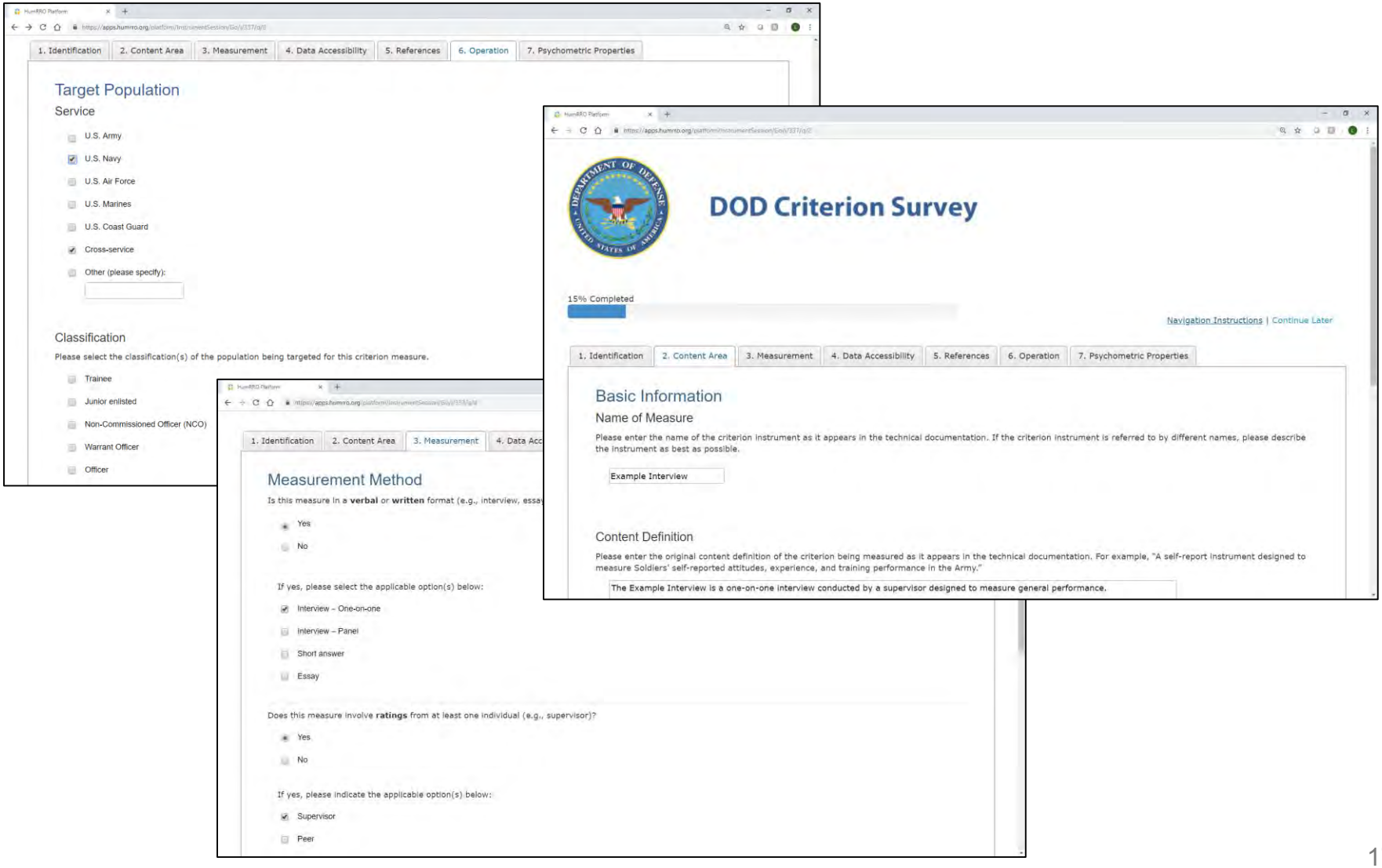
Develop a criterion database containing the descriptive information needed for available criterion instruments.

1. Conducted criterion review
   - Set parameters for collection of existing criterion instruments
   - Identification of “operational” and “exploratory” criteria
   - Met with CMAP members to finalize list
2. Developed an online data entry tool
   - Developed and refined metadata elements
   - Developed web-based survey
3. Populated criterion database
   - Over 230 operational and exploratory criteria
   - Finalized database format for easy discovery
4. Mapped criterion instruments against taxonomy and identified gaps or problem areas

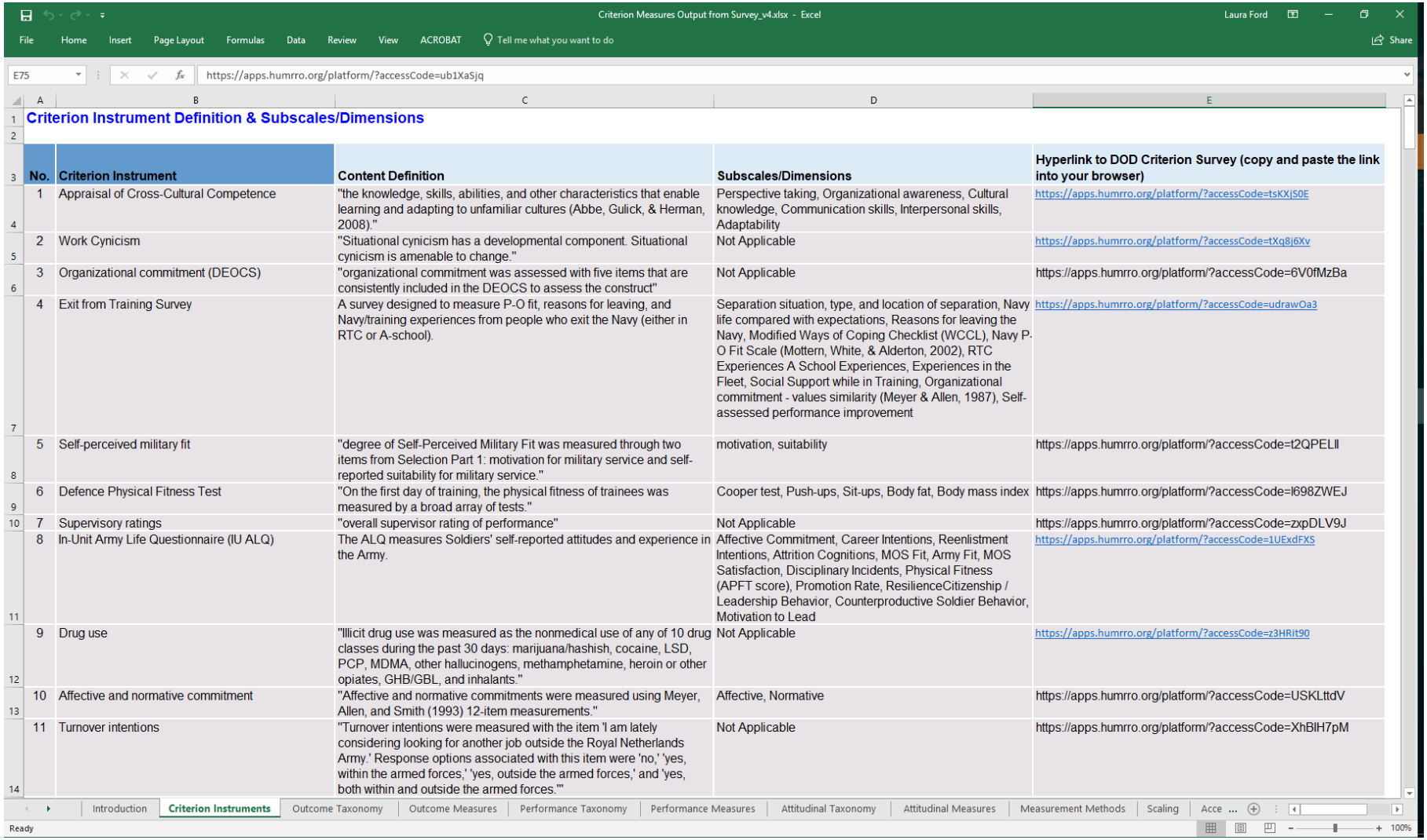
1. Parameters for Documenting Criterion Instruments

- Applicable across Services and Service components
  - Excluded job-specific instruments but included service-specific measures (e.g., general technical performance rating scales)
- Developed 1980 or later
  - Tagged to the Joint Performance Measurement (JPM) project
- Focus on first-term enlisted outcomes
  - Exclude officers / NCOs from primary search of “operational” criteria
- Operational emphasis
  - Expanded to an “exploratory” search as well (Military Psychology, IMTA) to capture other parts of the job performance taxonomy

2. On-line Survey Tool to Document Metadata



3. Created Database of Criterion Instruments


Criterion Measures Output from Survey_v4.xlsx - Excel

Laura Ford

File Home Insert Page Layout Formulas Data Review View ACROBAT Tell me what you want to do Share

E75 https://apps.humrro.org/platform/?accessCode=ub1XaSjq

Criterion Instrument Definition & Subscales/Dimensions

No.	Criterion Instrument	Content Definition	Subscales/Dimensions	Hyperlink to DOD Criterion Survey (copy and paste the link into your browser)
1	Appraisal of Cross-Cultural Competence	"the knowledge, skills, abilities, and other characteristics that enable learning and adapting to unfamiliar cultures (Abbe, Gulick, & Herman, 2008)."	Perspective taking, Organizational awareness, Cultural knowledge, Communication skills, Interpersonal skills, Adaptability	https://apps.humrro.org/platform/?accessCode=tsKXjS0E
2	Work Cynicism	"Situational cynicism has a developmental component. Situational cynicism is amenable to change."	Not Applicable	https://apps.humrro.org/platform/?accessCode=tXq8j6Xv
3	Organizational commitment (DEOCS)	"organizational commitment was assessed with five items that are consistently included in the DEOCS to assess the construct"	Not Applicable	https://apps.humrro.org/platform/?accessCode=6V0fMzBa
4	Exit from Training Survey	A survey designed to measure P-O fit, reasons for leaving, and Navy/training experiences from people who exit the Navy (either in RTC or A-school).	Separation situation, type, and location of separation, Navy life compared with expectations, Reasons for leaving the Navy, Modified Ways of Coping Checklist (WCCL), Navy P-O Fit Scale (Mottern, White, & Alderton, 2002), RTC Experiences A School Experiences, Experiences in the Fleet, Social Support while in Training, Organizational commitment - values similarity (Meyer & Allen, 1987), Self-assessed performance improvement	https://apps.humrro.org/platform/?accessCode=udrawOa3
5	Self-perceived military fit	"degree of Self-Perceived Military Fit was measured through two items from Selection Part 1: motivation for military service and self-reported suitability for military service."	motivation, suitability	https://apps.humrro.org/platform/?accessCode=t2QPELII
6	Defence Physical Fitness Test	"On the first day of training, the physical fitness of trainees was measured by a broad array of tests."	Cooper test, Push-ups, Sit-ups, Body fat, Body mass index	https://apps.humrro.org/platform/?accessCode=I698ZWEJ
7	Supervisory ratings	"overall supervisor rating of performance"	Not Applicable	https://apps.humrro.org/platform/?accessCode=zxpDLV9J
8	In-Unit Army Life Questionnaire (IU ALQ)	The ALQ measures Soldiers' self-reported attitudes and experience in the Army.	Affective Commitment, Career Intentions, Reenlistment Intentions, Attrition Cognitions, MOS Fit, Army Fit, MOS Satisfaction, Disciplinary Incidents, Physical Fitness (APFT score), Promotion Rate, ResilienceCitizenship / Leadership Behavior, Counterproductive Soldier Behavior, Motivation to Lead	https://apps.humrro.org/platform/?accessCode=1UExdFXS
9	Drug use	"Illicit drug use was measured as the nonmedical use of any of 10 drug classes during the past 30 days: marijuana/hashish, cocaine, LSD, PCP, MDMA, other hallucinogens, methamphetamine, heroin or other opiates, GHB/GBL, and inhalants."	Not Applicable	https://apps.humrro.org/platform/?accessCode=z3HRit90
10	Affective and normative commitment	"Affective and normative commitments were measured using Meyer, Allen, and Smith (1993) 12-item measurements."	Affective, Normative	https://apps.humrro.org/platform/?accessCode=USKLttdV
11	Turnover intentions	"Turnover intentions were measured with the item 'I am lately considering looking for another job outside the Royal Netherlands Army.' Response options associated with this item were 'no,' 'yes, within the armed forces,' 'yes, outside the armed forces,' and 'yes, both within and outside the armed forces.'"	Not Applicable	https://apps.humrro.org/platform/?accessCode=XhBlH7pM

Introduction | Criterion Instruments | Outcome Taxonomy | Outcome Measures | Performance Taxonomy | Performance Measures | Attitudinal Taxonomy | Attitudinal Measures | Measurement Methods | Scaling | Acce ...

Ready 100%

4. Mapped criterion instruments against the taxonomy

- Identified 74 current criterion instruments for focus:
  - 13 job performance rating scales
  - 13 performance tests (including knowledge, physical, and situational judgment)
  - 20 attitudinal surveys
  - 28 variables from administrative data (e.g., attrition, performance records)
- Mapped these criterion instruments to the three taxonomic domains – initial look highlighted constructs to which few or no instruments were mapped

Recommendation Goals

1. *Relevant and Generalizable* - Cover important components of the criterion space.
2. *Feasible* - Relatively easy to develop, administer, score, manage, and maintain.
3. *Psychometrically Sound* - Yield data that are reliable, sufficiently variable, and relatively free from contaminating variance.
4. *Flexible* - Across services to enhance Service-specific use, while ensuring that we support needed criterion-related validity inferences.
5. *Future-oriented* - What services could accomplish with additional effort, rather than solely on current practices.
6. *Utilitarian* - Making the most of the Services' current practices and procedures.

Develop Recommendations

Involved four steps:

1. Evaluated the generalizability and relevance of criterion constructs
   - (a) conducted CMAP-supported survey (n=26) to rate the importance and criticality of 33 job performance sub-dimensions (Slide 13).
   - (b) used internal HumRRO staff with extensive military testing experience to rate the generalizability and relevance of attitudes and organizational outcome constructs.
2. Evaluated the psychometric quality and feasibility of criterion measurement methods.
3. Computed criterion composites and applied to the criterion instruments in our database.
4. Reviewed and applied all of the data gathered, and assembled recommendations. Draft recommendations were reviewed and discussed with our contract monitors (Army and Air Force), a HumRRO panel of experienced military measurement experts, and the CMAP.

Recommendations

Administrative Data

- Focus on available administrative data
- Align outcome criteria across Services

Attitude Measurement Tools

- Develop cross-service self-assessments
- Align personnel assessments/surveys

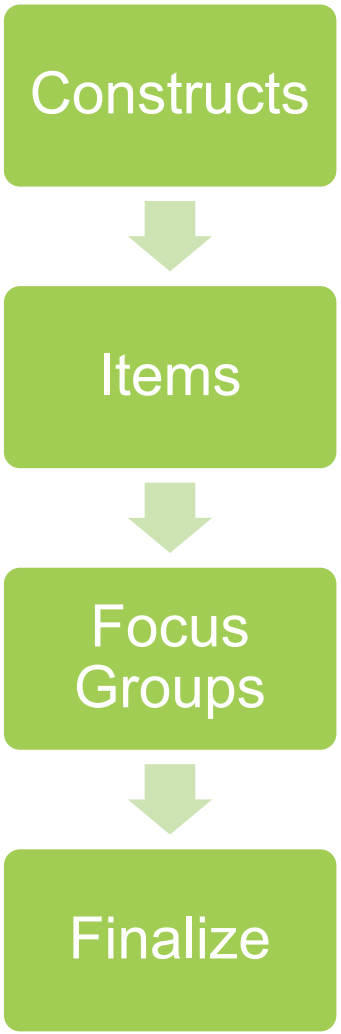
Performance Measurement Tools

- Develop standardized job performance rating scales, job knowledge assessments, and a situational judgement test (SJT)

Align Outcome Criteria Across Services

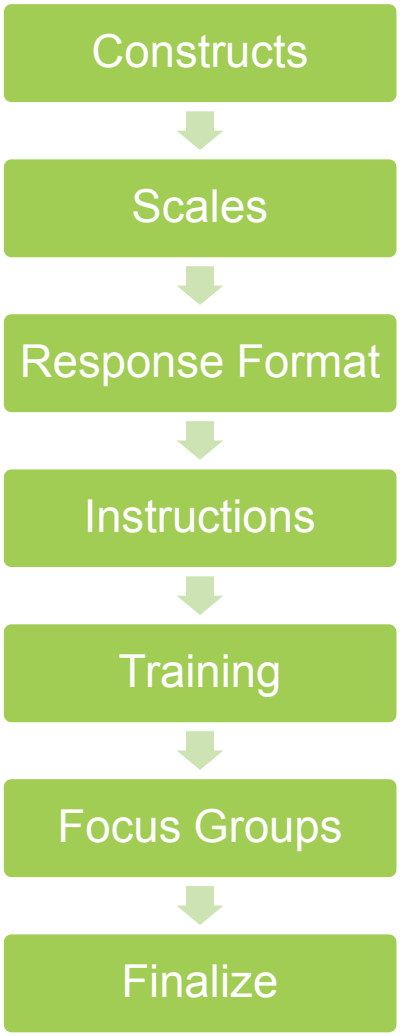
- Objective
  - Develop guides for constructing outcome criterion variables from administrative records
- Purpose
  - Develop step-by-step guides for constructing administrative outcome variables
  - Focus on developing shared guidelines for constructing **attrition** variables when used in validation research
  - Develop guides for other administrative variables depending on available data (training outcomes, disciplinary incidents, physical fitness test scores)
- Deliverables
  - Summary tables of cross-Service approaches
  - Step-by-step attrition construction guide
  - Step-by-step guides for other variables as needed

Develop Cross-Service Self-Assessments



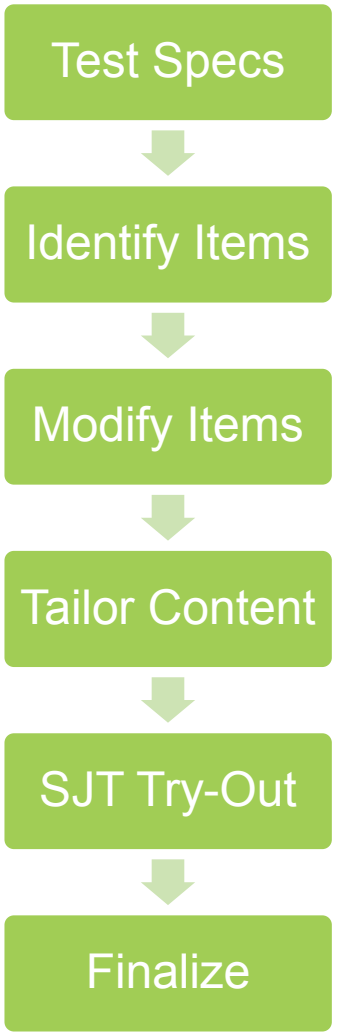
- Objective
  - Develop three short self-report assessments (end-of-training, in-unit, exit) tailored to each Service
- Purpose
  - Self-report assessments measure attitudes, performance outcomes, and reasons for attrition
- Deliverables
  - Cross-Service self-report assessments

Develop Performance Rating Scales



- Objective
  - Develop four standardized performance rating scales tailored to each Service for:
    - End-of-training
    - In-unit for supervisors and peers
- Purpose
  - Performance rating scales to measure job performance
  - Measures designed to enable efficient collection of criteria for use in selection/classification test validation studies
- Deliverables
  - Cross-Service performance rating scales

Develop DOD-wide SJTs



- Objective
  - Develop a DOD-wide SJT tailored to each Service
- Purpose
  - Develop an SJT for Services to use in validation research
  - Constructs:
    - Organizational Citizenship and Peer Leadership
    - Decision Making, Problem Solving, and Innovation
  - Level: end of first-term (18-36 months experience)
- Deliverables
  - Cross-Service SJT content, tailored to each Service
    - Ready for pilot-testing
  - Scoring protocols for SJT

Prepare Research Plan to Pilot Test/Validate Measures

- Objective
  - Develop a plan for evaluating the new measures
- Purpose
  - Take stock of data collection facilities, and equipment and capabilities within Services
  - Recommend a strategy of conducting data collections
  - Identify obstacles, for example:
    - Differences in training philosophy across Services that affect attrition rates and their interpretation
    - Low response rates to occupational and exit surveys
    - Avoiding redundancy of proposed measures (e.g., Exit survey) with current Service practices (e.g., USAF already has an exit survey)
    - Different IRB requirements among the Services
- Deliverable
  - Research plan (not to be executed under current contract)

HumRRO Project Task Leads

- Dr. Matt Allen
- Dr. Laura Ford (project manager)
- Mr. Chris Graves
- Dr. Chris Huber
- Dr. Teresa Russell (consultant)
- Dr. Martin Yu

HumRRO Military/Measurement Expert Panel

- Dr. Dave Dorsey
- Dr. Deirdre Knapp
- Dr. Rod McCloy
- Dr. Dan Putka
- Dr. Matt Trippe

Thank You

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

Reference Information

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

Taxonomies

Target	Dimension Set	Key References
Many constructs	The Campbell Model	Campbell, 2012; Campbell, Hanson, & Oppler, 2001; Campbell & Wiernik, 2015; Campbell, McCloy, Oppler, & Sager, 1993
	The Great Eight	Bartram, 2005
	Model of Work Role Performance	Griffin, Neal, & Parker, 2007
	Attributes of Successful Leaders	Zaccaro, Laport, & Jose, 2012
Specific constructs	Teamwork	O'Shea, Goodwin, Driskell, Salas, & Ardison, 2009; Shuffler, Pavlas, & Salas, 2012
	Task Performance	Borman, Grossman, Bryant, & Dorio, 2017
	Adaptability	Pulakos, Arad, Donovan, & Plamondon, 2000
	Self-Directed/Active Learning	Garrison, 1997; Russell, Rosenthal, Paullin, & Putka, 2006
	Employee Engagement	Macey & Schneider, 2008
	Organizational Citizenship	Dorsey, Cortina, Allen, Waters, Green, & Luchman, 2017; Goffin, Woycheshin, Hoffman, & George, 2013; Organ, 1988
	Counterproductive Work Behavior	Dalal, 2005; Rotundo & Spector, 2017; Spector, Bauer, & Fox, 2010; Spector et al., 2006
	Ethical Performance	Russell, Sparks, Campbell, Ramsberger, Handy, & Grand, 2017
	Cross Cultural Performance	Klafehn, Anderson, Taylor, Ingerick, & Ford, 2018, February
Military-specific constructs	Combat Performance	Wasko, Owens, Campbell, & Russell, 2012
	Situational Awareness	Matthews, Eid, Johnsen, & Boe, 2011
	1st term Performance	Campbell, Hanson, & Oppler, 2001; Sager, Russell, Campbell, & Ford, 2005
	Air Force-Wide Rating Dimensions	Lance, Teachout, & Donnelly, 1992
	Training Performance	Waugh & Russell, 2005

References

Bartram, D. (2005). The great eight competencies: A criterion-centric approach to validation, *Journal of Applied Psychology, 90*, 6, 1185-1203.

Borman, W. C., Grossman, M. R., Bryant, R. H., & Dorio, J. (2017). The measurement of task performance as criteria in selection research. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd edition). New York: Routledge.

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

Campbell, J. P. (2012). Behavior, performance, and effectiveness in the 21st century. In S. Kozlowski (Ed.), *The Oxford handbook of organizational psychology*: Volume 1. New York, NY: Oxford University Press.

Campbell, J. P., Hanson, M. A., & Oppler, S. H. (2001). Modeling performance in a population of jobs. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification.* New York: Lawrence Erlbaum Inc.

Campbell, J. P., & Knapp, D. J. (Eds.). (2001). *Exploring the limits in personnel selection and classification*. New York: Lawrence Erlbaum Inc.

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

Campbell, J. P., & Wiernik, B. M. (2015). The modeling and assessment of work performance. *Annual review of organizational psychology and organizational behavior*, *2*:191-19.28. doi: 10.1146/annurev-orgpsych-032414-111427

Dalal, R. S. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology, 90*(6), 1241-55.

Dorsey, D. W., Cortina, J. M., Allen, M. T., Waters, S. D., Green, J. P., & Luchman, J. (2017). Adaptive and citizenship-related behaviors at work. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd edition). New York: Routledge.

Garrison, D. R. (1997). Self-directed learning: Toward a comprehensive model. In *Adult Education Quarterly*, Fall 97 v 48 n 1, p18, 16p.

Goffin, R. D., Woycheshin, D. E., Hoffman, B. J., & George, K. (2013). The dimensionality of contextual and citizenship performance in military recruits: Support for nine dimensions using self-, peer, and supervisor ratings. *Military Psychology, 25*, 478-488.

Griffin, M. A., Neal, A., & Parker, S. K. (2007). A new model of work role performance: positive behavior in uncertain and interdependent contexts. *Academy of management journal*, *50*(2), 327-347.

Klafehn, J., Anderson, L., Taylor, W., Ingerick, M., & Ford, L. (2018, February). *Criterion development for evaluation of the cross-cultural Competence Assessment System* (Draft). Unpublished document.

Lance, C. F., Teachout, M. S., & Donnelly, T. M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology, 77,* 437-452.

References (cont.)

Macey, W. H., & Schneider, B. (2008). The meaning of employee engagement. *Industrial and Organizational Psychology*, *1*, 3-30.

Matthews, M. D., Eid, J., Johnsen, B. H., & Boe, O. C. (2011). A comparison of expert ratings and self-assessment s of situation awareness during a combat fatigue course. *Military Psychology*, *23*, 125-136.

Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome*. Lexington, MA: Lexington Books D.C. Heath and Company.

O'Shea, P. G., Goodwin, G. F., Driskell, J. E., Salas, E., & Ardison, S. (2009). The many faces of commitment: Facet-level links to performance in military contexts. *Military Psychology*, *21*(1), 5-23.

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

Rotundo, M., & Spector, P. E. (2017). New perspectives on counterproductive work behavior including withdrawal. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd edition). New York: Routledge.

Russell, T. L., Sparks, T. E., Campbell, J. P., Handy, K., Ramsberger, P., & Grand, J. A. (2017). Situating ethical behavior in the nomological network of job performance. *Journal of Business and Psychology*, *32* (3), 253-271.

Sager, C. E., Russell, T. L., Campbell, R. C., & Ford, L. A (2005). *Future soldiers: Analysis of entry-level performance requirements and their predictors* (Technical Report 1169). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Shuffler, M. L., Pavlas, D., & Salas, E. (2012). Teams in the Military: A review and emerging challenges. In J. H. Laurence & M. D. Matthews (Eds.), *The Oxford handbook of military psychology*. New York, NY: Oxford University Press.

Spector, P. E., Bauer, J. A., & Fox, S. (2010). Measurement artifacts in the assessment of counterproductive work behavior and organizational citizenship behavior: Do we know what we think we know? *Journal of Applied Psychology*, *95*(4), 781-790. doi: http://dx.doi.org/10.1037/a0019477

Spector, P. E., Fox, S., Penney, L. M., Brursema, K., Goh, A., & Kessler, S. (2006). The dimensionality of counterproductivity: Are all counterproductive behaviors created equal? *Journal of Vocational Behavior*, *68*, 446-460.

Wasko, L., Owens, K. S., Campbell, R., & Russell, T. (2012). Development of the combat/deployment performance rating scales. In D. J. Knapp, K. S. Owens, & M. T. Allen (Eds.), *Validating future force performance measures (Army Class): In-unit performance longitudinal validation* (Technical Report 1314). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Waugh, G. W., & Russell, T. L. (2005). Predictor situational judgment test. In D. J. Knapp, C. E. Sager, & T. R. Tremble (Eds.), *Development of experimental Army enlisted selection and classification tests and job performance criteria* (TR 1168). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Zaccaro, S. J., Laport, K., & Jose, I. (2012) The attributes of successful leaders: A performance requirements approach. *The Oxford handbook of leadership. Oxford Handbooks Online*. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780195398793.001.0001/oxfordhb-9780195398793-e-1

Trainee and 1st Term Performance (1 of 4 slides)

Hierarchical Trainee and 1st Term Performance Taxonomy Definitions

Performance Category	Performance Dimension	Sub-Dimension	Definition
A. Technical Performance			Performing job tasks proficiently; communicating clearly; making sound decisions; and being alert to safety and security concerns.
	A.1. Task Performance		Being able to perform job-specific and service-wide tasks proficiently
		Job-Specific Proficiency	Being able to perform job-specific tasks at the appropriate skill level.
		General Proficiency	Being able to perform service-wide tasks at the appropriate skill level (e.g., navigation in the Army and Marine Corps).
	A.2. Decision Making, Problem Solving, and Innovation		Making sound, timely decisions, even under pressure; analyzing situations and innovating solutions to problems; resolving conflicts; adapting plans and decisions as situations change.
		Decision Making, Problem Solving, and Innovation	Making sound, timely decisions, even under pressure; analyzing situations and innovating solutions to problems; resolving conflicts; adapting plans and decisions as situations change.
	A.3. Communication		Conveying oral and written information clearly; using appropriate nonverbal communication.
		Oral Communication	Conveying information in a clear, understandable, organized manner when speaking.
		Written Communication	Conveying information in a clear, understandable, organized manner when writing.
		Nonverbal Communication	Using alternative, culturally appropriate methods to interpret and convey meaning when common language is not shared.
	A.4. Safety and Security Consciousness		Following routine safety and security guidelines; and being alert to safety and security threats in non-routine situations.
		Safety and Security Consciousness in Everyday Work	Following safety and security guidelines and instructions, noticing and alerting others to potential hazards in day-to-day work.
		Safety and Security Consciousness during Mission Operations	Being alert to enemy and environmental threats and taking actions that do not place self or others at unwarranted risk.

Trainee and 1st Term Performance (2 of 4 slides)

Hierarchical Trainee and 1st Term Performance Taxonomy Definitions (Continued)

Performance Category	Performance Dimension	Sub-Dimension	Definition
B. Organizational Citizenship & Peer Leadership			Planning and structuring own work, and when in a leadership role, the work of others; taking initiative and persisting in work or training despite difficult conditions; supporting, helping, motivating, and respecting peers; and showing commitment to the organization, the team, and moral/ethical principles.
	B.1. Planning and Structuring Work		Leading peer when given a leadership role; working with team members to plan work; planning and organizing own responsibilities and studying.
		Providing Structure	Leading peers when given a leadership role, giving clear instructions, distributing tasks, and gaining others' cooperation.
		Teamwork	Working with other team members to interpret the mission, set and prioritize team goals, and monitor team performance.
		Self-Management	Managing own responsibilities (e.g., work assignments, gear, equipment, personal finances, family, and personal well-being), and appearing on duty prepared for work. Setting personal work objectives.
		Learning/Training Self-Management	Planning, organizing, and using study time effectively (e.g., setting aside specific times to study; completing assignments on time).
	B.2. Conscientious Initiative		Taking initiative; persisting with extra effort despite obstacles; taking steps to enhance own knowledge and skill.
		Classroom Learning	Being actively engaged in own learning by searching for and obtaining information, taking notes in class, highlighting relevant material, practicing new skills, and participating/contributing during classes.
		Self-Development	Developing or adapting own knowledge and skills by taking courses on own time, volunteering for training and development opportunities offered within the organization; and trying to learn new knowledge and skills on the job from others or through new job assignments.
		Persistence	Persisting with extra effort despite difficult conditions and setbacks, accomplishing goals that are more difficult and challenging than normal completing work on time despite unusually short deadlines, and performing at a level of excellence that is significantly beyond normal expectations.
		Initiative	Taking the initiative to do all that is necessary to accomplish team or organizational objectives encountered, finding additional work to perform when own duties are completed, and volunteering for work assignments.

Trainee and 1st Term Performance (3 of 4 slides)

Hierarchical Trainee and 1st Term Performance Taxonomy Definitions (Continued)

Performance Category	Performance Dimension	Sub-Dimension	Definition
	B.3. Support for Peers		Helping and motivating peers; cooperating with others; being respectful and considerate; accepting individual differences; and modeling core values.
		Helping Peers	Helping others by offering suggestions about their work, showing them how to accomplish difficult tasks, teaching them useful knowledge or skills, directly performing some of their tasks, and providing emotional support for personal problems.
		Cooperating	Cooperating with others by accepting their suggestions, following their lead, being open-minded and adapting to others' ways, and informing others of events or requirements that are likely to affect them.
		Courtesy & Respect	Showing consideration, courtesy, and tact in relations with others.
		Accepting Differences	Showing interest in and respect for people of other backgrounds or cultures by regularly engaging with them in a manner considerate of their norms.
		Motivating	Motivating others by applauding their achievements and successes, cheering them on in times of adversity, showing confidence in their ability to succeed, helping them overcome setbacks, and modelling leadership behavior.
		Serving as a Model	Modeling core values by acting unselfishly, enduring hardships without complaint, treating others well, behaving ethically, and showing confidence and enthusiasm.
	B.4. Organizational Support		Complying with organizational rules; demonstrating selfless service; presenting a positive image of the Service; and demonstrating honesty and integrity.
		Military Presence	Presenting a positive and professional image of self and the military even when off duty, maintaining proper military appearance.
		Selfless Service	Committing to the greater good of the team or group putting organizational welfare ahead of individual goals.
		Support for the Organization	Complying with organizational rules and procedures, encouraging others to comply with organizational rules and procedures, and suggesting procedural, administrative, or organizational improvements.
		Integrity/Moral Courage	Demonstrating honesty and integrity in job-related matters, even when own self-interests might be jeopardized, taking steps to protect the security of military equipment/supplies, and voluntarily reporting thefts, misconduct, and any other violations of military order and discipline.

Trainee and 1st Term Performance (4 of 4 slides)

Hierarchical Trainee and 1st Term Performance Taxonomy Definitions (Continued)

Performance Category	Performance Dimension	Sub-Dimension	Definition
C. Psychosocial Well-Being			Maintaining emotional control in stressful situations; and not engaging in counterproductive work behaviors.
	C.1. Adapting to Stressful Situations		Maintaining emotional control in stressful situations; noticing/monitoring own signs of stress from combat, work and home life and taking positive steps in managing stress reactions.
		Adapting to Stressful Situations	Maintaining emotional control in stressful situations; noticing/monitoring own signs of stress from combat, work and home life and taking positive steps in managing stress reactions.
	C.2. Counterproductive Work Behavior		Not engaging in delinquent behaviors or behaviors that affect the productivity of the organization (e.g., loafing, tardiness); not bullying, harassing, or hurting others; and not engaging in self-destructive behaviors.
		Loafing and Tardiness	Arriving late for work or not showing up; spending work time on personal activities (e.g., surfing the web).
		Abusing Substances and Other Self-Destructive Behavior	Engaging in self-destructive behavior (e.g., alcohol or drug abuse).
		Bullying, Harassing, or Hurting Others	Engaging in deviant behavior directed at others (e.g., physical attacks, verbal abuse, harassment).
		Delinquency	Engaging in deviant behaviors directed at the organization (e.g., theft, sabotage).
D. Physical Performance			Meeting fitness standards and sustaining physical performance over time.
	D.1. Physical Endurance		Sustaining physical performance over long periods of time despite lack of sleep and difficult conditions. Adapting to environmental challenges (e.g., weather, terrain).
		Physical Endurance	Sustaining physical performance over long periods of time despite lack of sleep and difficult conditions. Adapting to environmental challenges (e.g., weather, terrain).
	D.2. Physical Fitness		Meeting military standards for weight, physical fitness, and strength, maintaining own health.
		Physical Fitness	Meeting military standards for weight, physical fitness, and strength, maintaining own health.

Attitudinal Taxonomy

Attitudinal Criterion Domain Taxonomy		
Construct	Definition	Facets
Work Satisfaction	An individual's satisfaction with work	- Whole job satisfaction - Job facet satisfaction - Career satisfaction
Morale	A holistic judgment of one's own morale	
Organizational Commitment	An individual's psychological bond with the organization, as represented by an affective attachment to the organization, internalization of its values and goals, and a behavioral desire to put forth effort to support it.	- Affective - Continuance - Normative
Withdrawal Cognitions/Intentions	Thinking about or intending to quit one's job	- Attrition cognitions - Short-term active duty career continuance intentions - Long-term active duty career continuance intentions - Post-active duty plans - Deployment-attributed change in career intentions
Person-Environment Fit (PE Fit)	Congruence between the individual's abilities, needs, and expectations and characteristics of the organization, job or group.	- Person-Job, Needs-supplies fit - Person-job, Demands-abilities fit - Person-organization fit - Person-team fit

Note. Based primarily on Allen, Knapp, & Owens (2016), Arthur, Bell, Villado, & Doverspike (2006), Cable & Edwards (2004), Cable & Judge (1997), Dawis & Lofquist (1984), Edwards (1996), Greenhaus, Parasuraman & Wormley (1990), Hom (2011), Hom, Lee, Shaw, & Hausknecht (2017), Judge, Cable, Boudreau, & Bretz (1994), Judge & Kammeyer-Mueller (2012), Judge, Weiss, Kammeyer-Mueller, & Husin (2017), Meyer & Allen (1991), Meyer, Kam, Goldenberg, & Bremner (2013), Weiss, Dawis, Lofquist, & England (1967).

Organizational Outcomes

Organizational Outcome Taxonomy		
Outcome Construct	Facet	Example Indicators
Attrition	Delayed entry program (DEP)	- DEP attrition
	Boot Camp	- Attrition from boot camp
	Advanced Training	- Attrition from advanced training
	In-unit	- Attrition in-unit (premature attrition)
	Re-enlistment	- Re-enlistment for 2nd term; propensity to re-enlist
Reprimands	Reprimands	- Articles 15/ reprimands
Experience	Tenure	- Time in grade/rank
		- Time in uniform/Length of service
	Rank	- Rank
Initiative	Awards	- Merit-based awards and commendations
Performance	Advanced Training	- Training school grades
		- Pass/Fail
		- Rank in class
		- Training recycles/Wash-backs
	In-unit	- Supervisor performance ratings/ Enlisted Performance Ratings (EPR in USAF)/ Proficiency marks (PRO marks in USMC)
		- Job knowledge test scores (e.g., USAF Skill/Knowledge Test [SKT]; USAF Promotion Fitness Examination [PFE])

Organizational Outcome Taxonomy (con'd)		
Outcome Construct	Outcome Construct	Outcome Construct
Performance (con'd)	Skill Upgrading	- Skill level attainment (e.g., USAF skill level badges)
	Promotion Potential	- Promotion exam scores
	Physical	- Current physical fitness
	Qualifications	- Rifle/pistol qualification score
		- Other qualifications (swim, NBC, brown belt, Ranger)
	Re-enlistment Eligibility	- Computed Tier Score (re-enlistment eligibility composite based on a number of qualifications)
Productivity	Skilled Tenure	- Qualified man months (QMM - number of months in service at qualified level based on skills test)
	Skilled Tenure	- Months mission ready service (months of service at the highest skill level)
	Quantity of Performance	- Productive capacity (rate of task performance)
Promotion	Rate	− Promotion rate (a deviation score comparing to other service members with the same time in service and in the same job)
	Time	− Promotion time to E-4

Note. Indicators were drawn primarily from Alley, Pacheco, Birkelbach, Schwartz & Weissmuller (2007), Campbell & Knapp (2001), Halper, Goodman, & Alley (2010), Ingerick, Allen, Weaver, Caramagno, & Hooper (2006), Knapp & Campbell, 1993, Mayberry (1990), Sims & Hiatt (2001), and Wathen (2014).

Tab N

Navy ASVAB Validation Program

For

DAC MPT

27 September 2019

Stephen E. Watson, Ph.D.
Director, Navy Selection & Classification

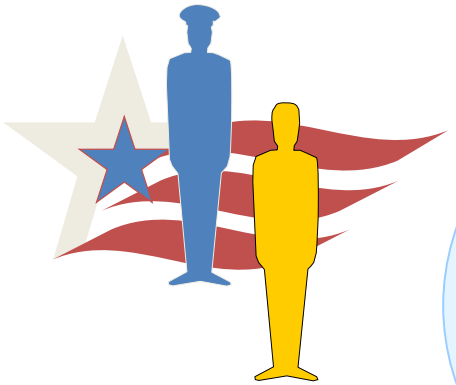
MyNAVYHR
Serving Sailors 24/7

Agenda

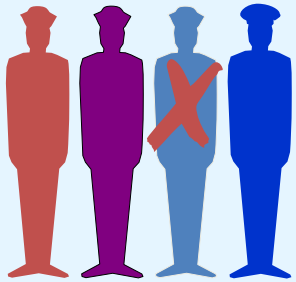
- **Background**
  - Accession Process
  - ASVAB Primer
- **The Business Process**
  - Perform ASVAB Validation Study
  - Review Recommendations (Stakeholders)
  - Review Final Report for Approval
  - Publish Updated ASVAB Standards
  - Update IT Systems with New Standards
- **Process Tools and Artifacts**
  - Predictive Validity Tool
  - Projected Impact Spreadsheets

Background

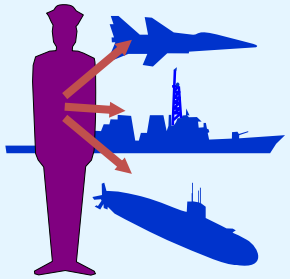
Accession Process



Contact with Recruiter



Selection



Classification

Swear-in/ DEP

- Ability Screens
- Moral, Financial & Educational Screens

- ASVAB
- Medical, Mental & Moral Checks

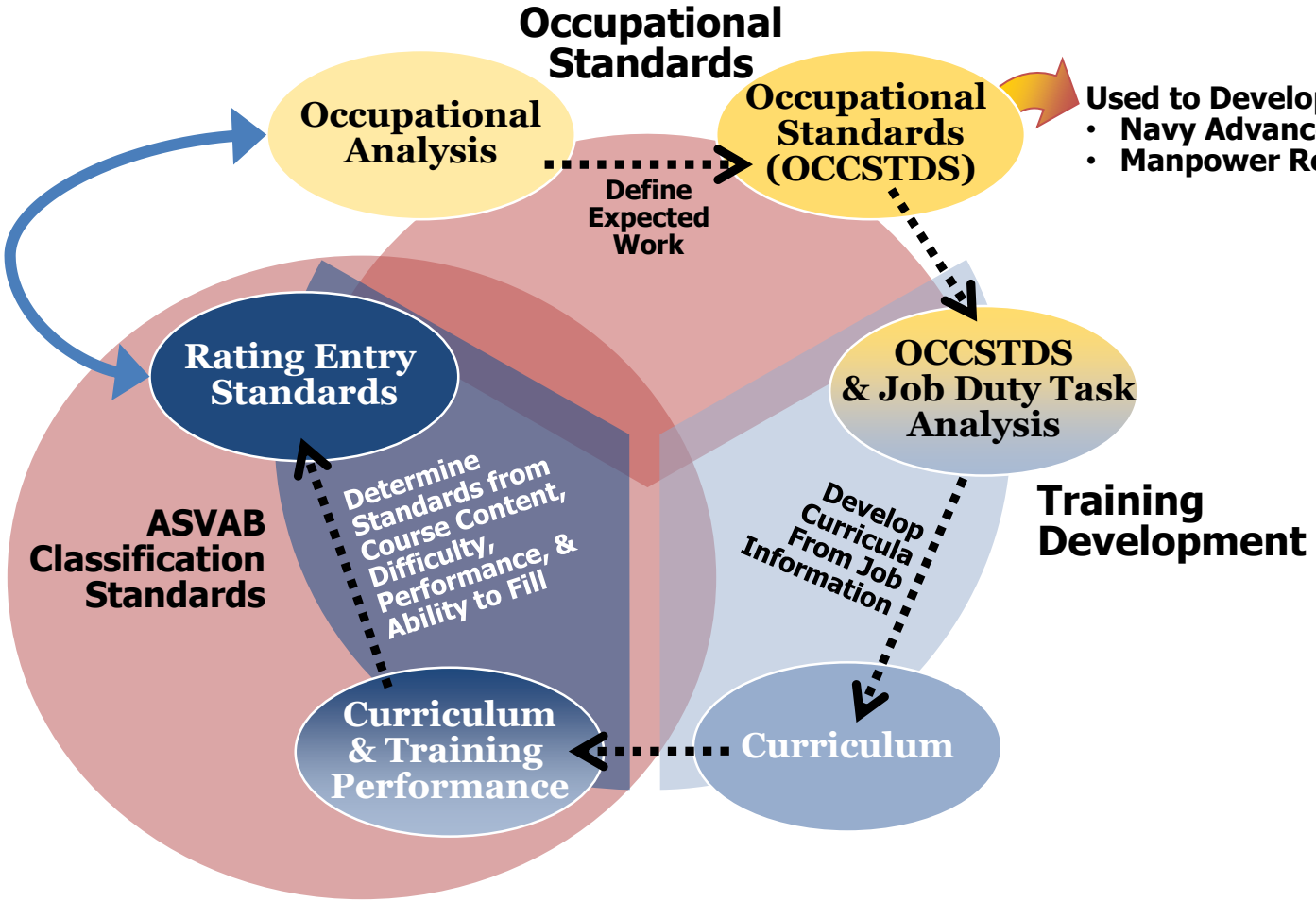
- Job & Career Info
- School Guarantee
- Ship Date

- Swear-In
- Delayed Entry Program
- Ship to RTC

NETC N55 Develops Policy on Selection and Classification Standards

DRAFT // PRE-DECISIONAL

Occupational, Training, & ASVAB Standards



Rating Entry Standards Ultimately Meet Occupational Standards

ASVAB Use in Selection & Classification

- **Armed Services Vocational Aptitude Battery (ASVAB)**
- **Consists of 9 sub-tests used by all military services for enlisted personnel**
  - Use in Selection:
    - Armed Forces Qualification Test (AFQT)
    - Measure of General Intelligence
    - 2VE+MK+AR
    - VE itself a combination of PC & WK scores
  - Use in Classification:
    - Navy Classification Composites
    - Measures of Specific Intelligence
    - Ratings use Specific Composites

ASVAB & Special Tests

TEST	CONTENT	
General Science (GS)	Biological and physical sciences	ASVAB
Arithmetic Reasoning (AR)	Arithmetic word problems	
Word Knowledge (WK)	Synonyms/meaning of words in context	
Paragraph Comprehension (PC)	Written passages	
Mathematics Knowledge (MK)	Algebra, geometry, fractions, decimals, exponents	
Electronic Information (EI)	Electrical Principles and electronics	
Auto and Shop Information (AS)	Automotive, tool, shop, practices	
Mechanical Comprehension (MC)	Mechanical and physical principles	
Assembling Objects (AO)	Patterns and connection point recognition	
Coding Speed (CS)	Perceptual Speed/Accuracy	Special Classification
Defense Language Aptitude Battery (DLAB)	Aptitude for learning languages	
Navy Advanced Placement Test (NAPT)	Advanced Physics, Mathematics, and Chemistry	
Mental Counters (MCt)	Working memory test (currently under development)	
Cyber Test	Knowledge in cyber field	

DRAFT // PRE-DECISIONAL

Navy ASVAB Composites

Composite	Ratings Using This Composite	Occupational Group Name	Composite Name
VE+MK	CTR LN PS RP YN	Administration	A1
VE+MK+CS	LN OS PS RP YN	Administration	A2
VE+AR	AZ CS EOD IS LS MC NC ND QM S/PACT SB SH SO	Specialized	S1
VE+MK+GS	AG CTI HM HM/ATF HMDA CTT	Specialized	S2
WK+AR	HM/ATF MA	Specialized	S3
VE+AR+MK+GS	HM/ATF	Specialized	S4
AR+MC+AS	BU CM EO SW	Mechanical	M1
MK+AS+AO	BM AO PR	Mechanical	M2
VE+AR+MK+AS	ABE ABF ABH AD AIRC AIRR AM AME A/PACT AO AS AW BM DC E/PACT EN GSM HT MM MR PR	Operations	O1
AR+2MK+GS	CTN EA IT ITS OS	Operations	O2
VE+AR+MK+AO	AME BU EN GSM MM MN	Operations	O3
VE+MK+MC+CS	SO AC	Operations	O4
VE+AR+MK+MC	AC AD AE AIRC AIRR AM AS AT AV AW CSS CTN DC E/PACT EM ETS FTS GSE HT LSS MMS MN MR MT NF SECF YNS	Technical	T1
AR+MK+EI+GS	AE AECF AT AV CE CSS CTM CTT EM ET ETS FC FT GM GSE IC IT ITS LSS MT NF SECF STG UT YNS	Technical	T2
GS+MC+EI	SO EOD	Technical	T3
DLAB (Special Test)	CTI	Language Aptitude	
CT (Special Test)	CTN	Cyber Knowledge	
NAPT (Special Test)	NF	Advanced Programs Test	

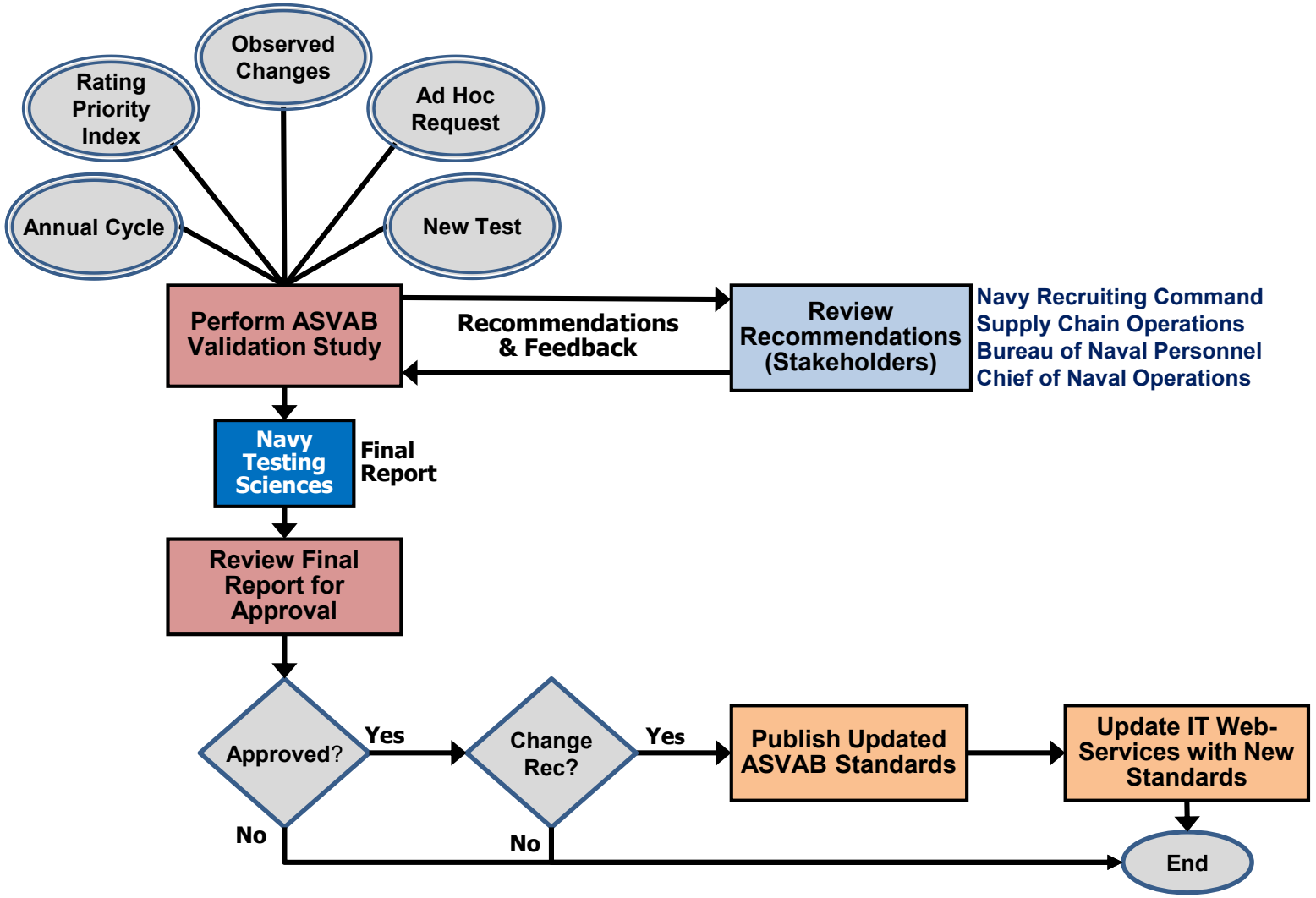
ASVAB Institutional Controls

- **OUSD(PR):** Office of the Under Secretary of Defense for Personnel & Readiness, Accession Policy (AP) Directorate, sets ASVAB policy
- **DPAC:** Defense Personnel Assessment Center is the Executive Agent (EA) for ASVAB Development and Maintenance
- **MEPCOM:** Military Entrance Processing Command is charged with ASVAB administration, funding infrastructure and support
- **SERVICES:** Navy
  - Charged with validating their ASVAB classification composites
  - Establishing and maintaining efficient cutscores
  - For Navy, both are directed and executed by NETC N55

The Business Process

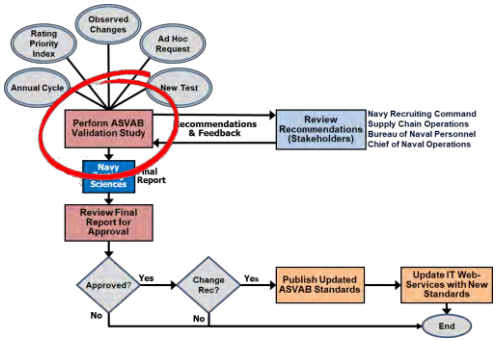
High-Level Business Process Diagram

Process Name: Develop & Validate ASVAB Selection and Rating Entry Standards



ASVAB Validation Study Frequency & Type

- **Frequency**
  - Annually for every rating
  - Schedule priority for a rating may change based on:
    - Observed predictive validity changes
    - Ad hoc requests from stakeholders
    - Observations from the Rating Priority Index
    - New tests
- **Study Types**
  - Automated : Based on analysis of historical training data
  - Deep Dive : When additional input from schoolhouse is required



"Automated" Study

- **Process Accession Data using Training Outcomes**
- **Correlate Test Scores to Training Success**
  - For both *operational and alternative* ASVAB composites
  - Correct for Range Restriction & Calculate Adverse Impact (AI)
- **Identify Candidate Alternative Composite(s)**
- **Evaluate Key Metrics:**
  - Qualification Rate & Adverse Impact using historical data
  - Training Success using regression-based model
  - Cross-validate with latest (current year) data
- **Test Candidate Alternative Composite/Cutscore Sets**
  - Full-year whole-Navy assignment simulation
  - Compare simulation results against operational standards

Process Tools

Predictive Validity Tool

- **Purpose:** Calculates correlation estimates and population-level group-score differences of ASVAB composites/special tests against training success metrics, for each rating
- **Use:** Select the best composites for
  - Minimizing academic setback and failure rates
  - Reducing adverse impact on gender & race/ethnicity

Predictive Validity Tool

Predictive Validity Estimates and Population-Level Group-Score Differences of ASVAB Composites and Special Tests (DLAB, NAPT) by Rating										
		Predictive Validity					Standardized Mean Difference			
Rating	ASVAB Composite/Special Test	*r* MRR+DC	*r* MRR	*r* PAY97 MRR+DC	*r* PAY97 MRR	*r* Obs	*d* F-M	*d* AA-W	*d* H-W
AC	Nuclear (VE + AR + MK + MC)	-.60	-.48	-.89	-.71	-.32	-.59	-.90	-.44
AC	Engineer (VE + AR + MK + AS)	-.58	-.46	-.88	-.70	-.31	-.69	-1.00	-.53
AC	Exp AO2 (VE + AR + MK + AO)	-.57	-.46	-.88	-.70	-.30	-.35	-.74	-.29
AC	Exp CS2 (VE + MK + MC + CS)	-.57	-.45	-.87	-.70	-.29	-.39	-.76	-.34
AC	General Tech (VE + AR)	-.57	-.45	-.87	-.69	-.29	-.47	-.77	-.41
AC	Electronics (AR + MK + EI + GS)	-.56	-.45	-.87	-.69	-.29	-.68	-.93	-.46
AC	Basic Elec (AR + 2MK + GS)	-.56	-.44	-.86	-.69	-.28	-.40	-.67	-.28
AC	Administration (VE + MK)	-.55	-.44	-.86	-.69	-.27	-.26	-.60	-.36
AC	Corpsman (VE + MK + GS)	-.54	-.43	-.86	-.68	-.26	-.44	-.83	-.47
AC	Security (AR + WK)	-.53	-.43	-.85	-.67	-.26	-.47	-.76	-.44
AC	Mechanical (AR + MC + AS)	-.51	-.41	-.77	-.62	-.25	-.95	-1.19	-.59
AC	Exp CS1 (VE + MK + CS)	-.50	-.40	-.84	-.67	-.20	-.11	-.46	-.19
AC	Exp AO1 (MK + AS + AO)	-.49	-.39	-.80	-.63	-.24	-.63	-1.02	-.42
AC	EOD-SEAL (GS + MC + EI)	-.48	-.38	-.77	-.62	-.22	-.87	-1.14	-.62
AD	Electronics (AR + MK + EI + GS)	-.45	-.24	-.74	-.39	-.15	-.68	-.93	-.46
AD	Engineer (VE + AR + MK + AS)	-.45	-.23	-.74	-.39	-.16	-.69	-1.00	-.53
AD	EOD-SEAL (GS + MC + EI)	-.43	-.23	-.70	-.37	-.15	-.87	-1.14	-.62
AD	Nuclear (VE + AR + MK + MC)	-.43	-.23	-.73	-.39	-.13	-.59	-.90	-.44
AD	Corpsman (VE + MK + GS)	-.42	-.22	-.71	-.38	-.13	-.44	-.83	-.47
AD	Mechanical (AR + MC + AS)	-.41	-.22	-.68	-.36	-.14	-.95	-1.19	-.59
AD	Administration (VE + MK)	-.39	-.21	-.70	-.37	-.11	-.26	-.60	-.36
AD	Exp AO1 (MK + AS + AO)	-.39	-.21	-.69	-.37	-.13	-.63	-1.02	-.42
AD	Basic Elec (AR + 2MK + GS)	-.39	-.21	-.70	-.37	-.10	-.40	-.67	-.28
AD	Exp AO2 (VE + AR + MK + AO)	-.38	-.20	-.71	-.37	-.10	-.35	-.74	-.29
AD	General Tech (VE + AR)	-.38	-.20	-.70	-.37	-.10	-.47	-.77	-.41
AD	Exp CS2 (VE + MK + MC + CS)	-.37	-.20	-.70	-.37	-.09	-.39	-.76	-.34
AD	Security (AR + WK)	-.36	-.19	-.67	-.36	-.08	-.47	-.76	-.44
AD	Exp CS1 (VE + MK + CS)	-.29	-.15	-.64	-.34	-.06	-.11	-.46	-.19
AECF	Exp AO2 (VE + AR + MK +AO)	-.48	-.31	-.82	-.53	-.20	-.35	-.74	-.29
AECF	Nuclear (VE + AR + MK + MC)	-.48	-.31	-.82	-.53	-.20	-.59	-.90	-.44
AECF	Engineer (VE + AR + MK + AS)	-.47	-.30	-.81	-.52	-.19	-.69	-1.00	-.53

Projected Impact Spreadsheets

- **Purpose:** Calculates projected impact of alternative cutscores on
  - Qualification Rate
    - Total
    - Adverse impact ratios by gender & race/ethnicity
  - Setback/Attrition Rates
    - Total
    - By gender, race/ethnicity

Projected Impact Spreadsheets

	A	B	C	G	H	I	P	AE	AH	AN	AO	AP	AQ	AT
1				Overall QR			Gender	Race/Ethnicity		Overall Academic Setback-Failure Rate			Acad Setback-F	
2		Qualifying Composite					AIR F-M	B-W AIR	H-W AIR				Female	Male
3	Rating	Name	Cutscore	Estimate	CI L 95%	CI U 95%	Estimate	Estimate	Estimate	Estimate	CI L 95%	CI U 95%	Estimate	Estimate
4	CTR	AR+VE	109	48.2%	48.0%	48.4%	0.64	0.46	0.72	4.38%	4.3%	4.5%	4.5%	4.4%
5	CTR	AR+VE	103	71.2%	71.0%	71.4%	0.78	0.67	0.85	10.9%	10.8%	11.1%	11.6%	10.8%
6	CTR	AR+VE	104	67.2%	67.1%	67.4%	0.76	0.63	0.83	9.7%	9.5%	9.8%	10.1%	9.6%
7	CTR	AR+VE	105	63.3%	63.1%	63.5%	0.73	0.60	0.80	8.5%	8.3%	8.6%	8.8%	8.4%
8	CTR	AR+VE	106	59.4%	59.3%	59.6%	0.71	0.56	0.78	7.3%	7.1%	7.4%	7.5%	7.2%
9	CTR	AR+VE	107	55.6%	55.4%	55.8%	0.69	0.52	0.76	6.2%	6.1%	6.4%	6.4%	6.2%
10	CTR	AR+VE	108	51.8%	51.6%	52.0%	0.67	0.49	0.74	5.3%	5.1%	5.4%	5.4%	5.2%
11	CTR	AR+VE	109	48.2%	48.0%	48.4%	0.64	0.46	0.72	4.38%	4.3%	4.5%	4.5%	4.4%
12	CTR	AR+VE	110	44.7%	44.5%	44.9%	0.62	0.43	0.69	3.6%	3.5%	3.7%	3.8%	3.6%
13	CTR	AR+VE	111	41.3%	41.1%	41.5%	0.60	0.40	0.67	3.0%	2.8%	3.1%	3.1%	2.9%
14	CTR	AR+VE	112	38.1%	37.9%	38.3%	0.58	0.38	0.65	2.4%	2.3%	2.5%	2.4%	2.4%
15	CTR	AR+VE	113	35.0%	34.8%	35.2%	0.56	0.36	0.64	1.9%	1.8%	2.0%	2.0%	1.9%
16	CTR	AR+VE	114	32.0%	31.9%	32.2%	0.54	0.34	0.61	1.5%	1.4%	1.6%	1.5%	1.5%
17	CTR	AR+VE	115	29.2%	29.0%	29.4%	0.52	0.31	0.59	1.2%	1.1%	1.3%	1.3%	1.2%
35	CTR	MK+VE	103	82.8%	82.7%	83.0%	0.94	0.86	0.92	8.6%	8.5%	8.7%	10.2%	8.1%
36	CTR	MK+VE	104	79.0%	78.8%	79.1%	0.93	0.82	0.90	6.7%	6.6%	6.8%	8.0%	6.3
37	CTR	MK+VE	105	74.7%	74.6%	74.9%	0.92	0.78	0.88	5.0%	4.9%	5.1%	6.0%	4
38	CTR	MK+VE	106	70.2%	70.1%	70.4%	0.90	0.75	0.86	3.6%	3.5%	3.7%	4.4%	
39	CTR	MK+VE	107	65.6%	65.5%	65.8%	0.88	0.70	0.84	2.5%	2.4%	2.5%	3.1%	
40	CTR	MK+VE	108	61.0%	60.8%	61.2%	0.86	0.66	0.82	1.6%	1.6%	1.7%	2.1%	
41	CTR	MK+VE	109	56.4%	56.2%	56.6%	0.83	0.63	0.79	1.1%	1.0%	1.1%	1.4%	
42	CTR	MK+VE	110	52.0%	51.8%	52.2%	0.81	0.59	0.77	0.7%	0.6%	0.7%	0.9%	
43	CTR	MK+VE	111	47.7%	47.5%	47.9%	0.78	0.55	0.74	0.4%	0.4%	0.4%		
44	CTR	MK+VE	112	43.6%	43.4%	43.8%	0.76	0.51	0.72	0.2%	0.2%			
45	CTR	MK+VE	113	39.7%	39.5%	39.9%	0.73	0.48	0.69	0.1%				
46	CTR	MK+VE	114	36.1%	35.9%									

Example Applications

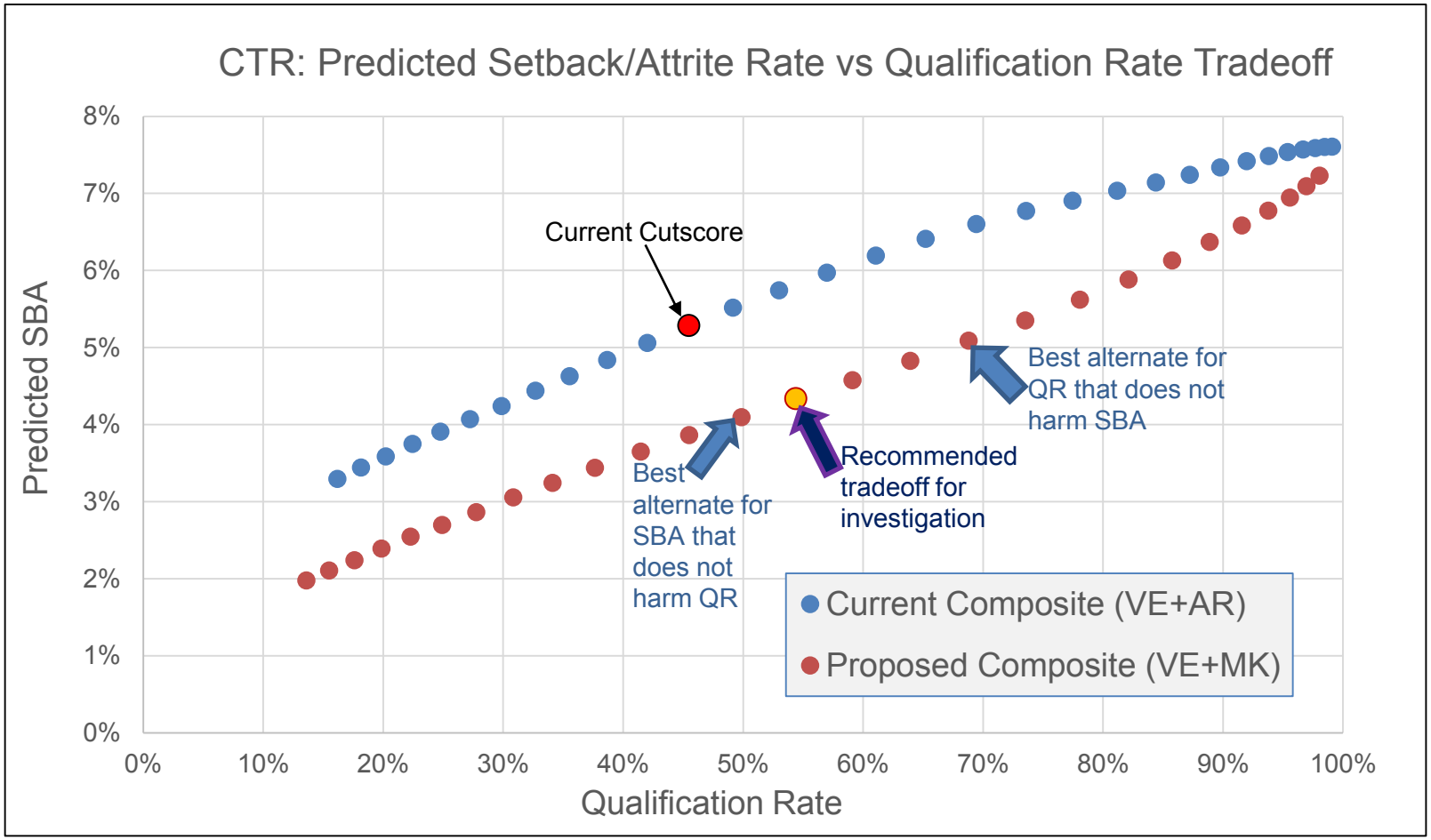
Composite Validity for Selected Ratings

Common observation from analysis was that operational composites did not always offer the highest predictive validity

Rating	ASVAB Composite	Validity
CTR	MK + VE	0.46
CTR	AO + AR + MK + VE	0.45
CTR	AR + GS + 2MK	0.45
CTR	AR + MC + MK + VE	0.43
CTR	GS + MK + VE	0.42
CTR	AR + EI + GS + MK	0.41
CTR	AR + AS + MK + VE	0.41
CTR	AO + AS + MK	0.40
CTR	CS + MK + VE	0.40
CTR	CS + MC + MK + VE	0.39
CTR	AR + VE	0.38
CTR	AR + WK	0.36
CTR	EI + GS + MC	0.34
CTR	AR + AS + MC	0.32

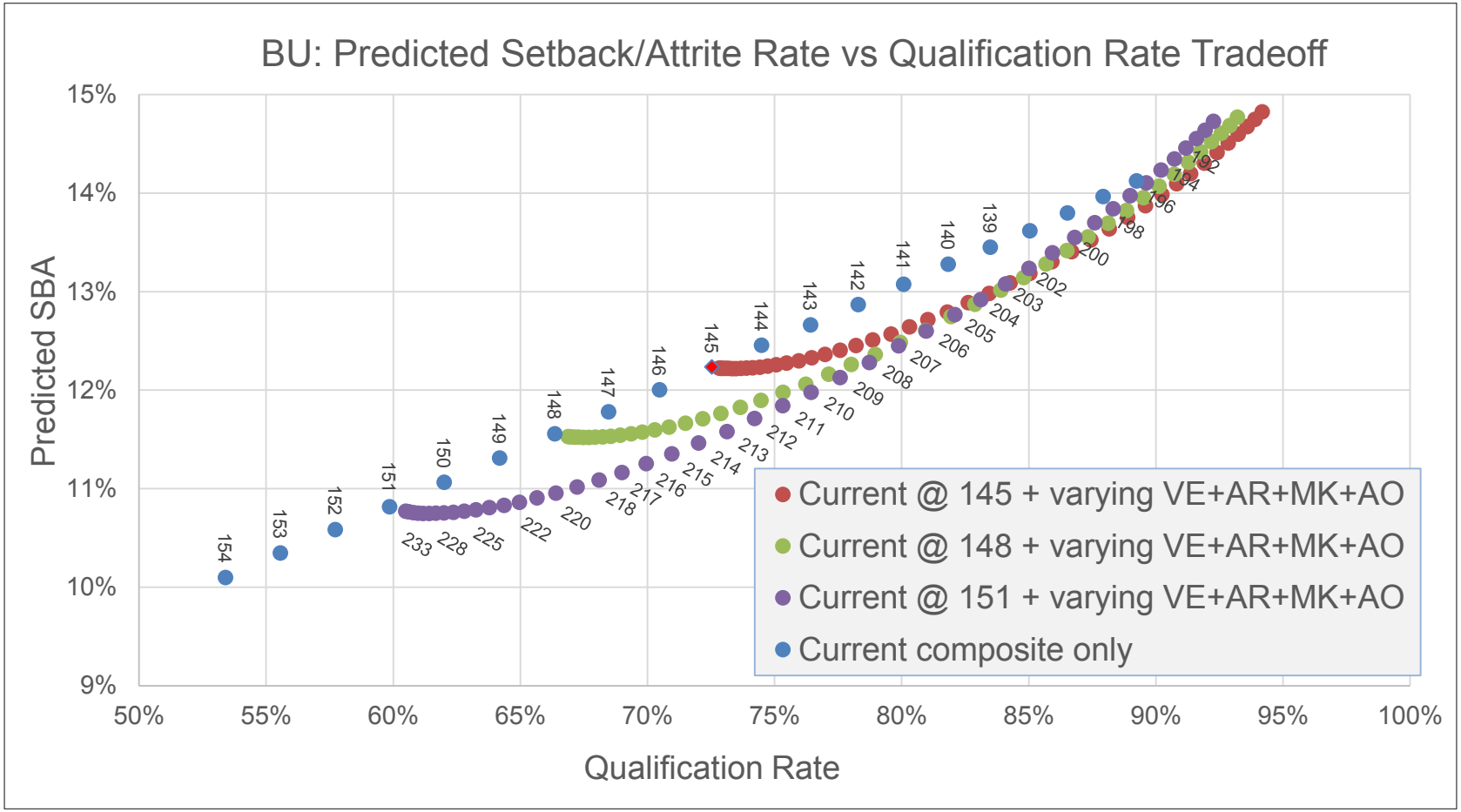
Recommend Replace AR + VE with MK + VE

CTR Composite/Cutscore Options



Recommend use of "VE+MK => 109" Composite/Cutscore

BU Composite/Cutscore Options



Recommend Add "VE+AR+MK+AO" As Alternate Standard

ASVAB Validation Study – Outcomes

Simulation Results: Current vs. Proposed Composites/Cutscores

Navy-Wide Impact	Baseline (FY19)	Proposed (FY19)
Predicted FPPS	85.95%	85.99%
Number Req Waiver	399	390

	CTR		BU	
	Baseline	Proposed	Baseline	Proposed
	AR+VE>=109	**VE+MK>=109**	**AR+MC+AS>=145**	**AR+MC+AS>=145 or VE+AR+MK+AO>=209**
Qual-Rate	42%	48%	65%	71%
		+2134		+2052
SBA%	6.1 %	5.8 %	16.5 %	16.0 %
AIR: F/M	0.64	0.83	0.60	0.84
AIR: AA/W	0.46	0.63	0.53	0.73

Proposed provide higher qualification rates and improved adverse impact ratios without increasing training setback rates

Questions?

Tab O

OPA

OFFICE OF PEOPLE ANALYTICS

Standard Setting Study for ASVAB Technical Tests

DACMPT Meeting

09.27.2019 | Philadelphia, PA

Tia Fechter

Defense Personnel Assessment Center

BACKGROUND

- Early 2015, DPAC (Mary Pommerich) generated a 23-task list—*Plans for Evaluating Current ASVAB Tests*—to guide a holistic evaluation of the ASVAB with respect to its relevancy for this and future generations.

- One task involves evaluating the usefulness/appropriateness of existing tests with the current population.

BACKGROUND

Subtasks include:

- Tracking test scores over years (1984–2014)
- Evaluating what fraction of the population possesses the knowledge/skills assessed by the test, over time.
  - The context for this task is criterion-referenced
  - Criterion-referenced: a measure of performance against a fixed set of predetermined learning standards (i.e., criteria, knowledge, skills)
  - ASVAB subtests do not have cut scores on a test score scale that establish a demarcation that categorizes examinees into those with significant exposure to the content the subtest measures and those who don't*
  - Implementing a standard setting would establish these cuts and allow for the calculation of the proportion of the population that could be considered "significantly exposed"

BACKGROUND

Subtasks, cont'd.

- Cut scores will help us determine the proportion of applicants significantly exposed to the content of interest by subtest and by year (1984–2018) and thereby will enable us to evaluate the following:
  - Trends in what proportion of applicants are "significantly exposed" across time
  - Continued usefulness of subtests for making classification decisions with the current population of youth

SUBTESTS OF INTEREST

ASVAB Science and Technical Tests

- Automotive Information (AI)
- Shop Information (SI)
- Electronics Information (EI)
- Mechanical Comprehension (MC)
- General Science (GS)

Special Tests

- Cyber Test

SUBTESTS OF INTEREST

- ASVAB Science and Technical Tests scores and the Cyber Test scores are used in composite scores for classifying military personnel into various military occupational specialties (MOS; e.g., Air Traffic Controller)
  - Performance on ASVAB subtests are weighted, as appropriate, to match the skills and abilities required for successful performance in training schools for the respective MOS.
  - Each of the armed services is responsible for validating their composite scores.

SUBTESTS OF INTEREST

- Beyond automotive, electronics, mechanical, and shop specialties?
  - Since 1968, the ASVAB has included tests for AI, EI, MC, and SI (and GS).
  - Since 1968, possible MOSs have expanded to include fields such as cyber security.
  - DPAC would like to determine whether AI, EI, MC, and SI are still dominant technical areas that high school students are exposed to.
  - DPAC would like to determine whether AI, EI, MC, and SI are technical areas that represent the bulk of current-day vocational interests and needs.

SUBTESTS OF INTEREST

- Beyond automotive, electronics, mechanical, and shop specialties?
  - DPAC would like to determine whether areas such as computer science may be more prevalent and relevant.
  - DPAC would like to use GS as a baseline subtest for the technical tests and Cyber Test comparisons.

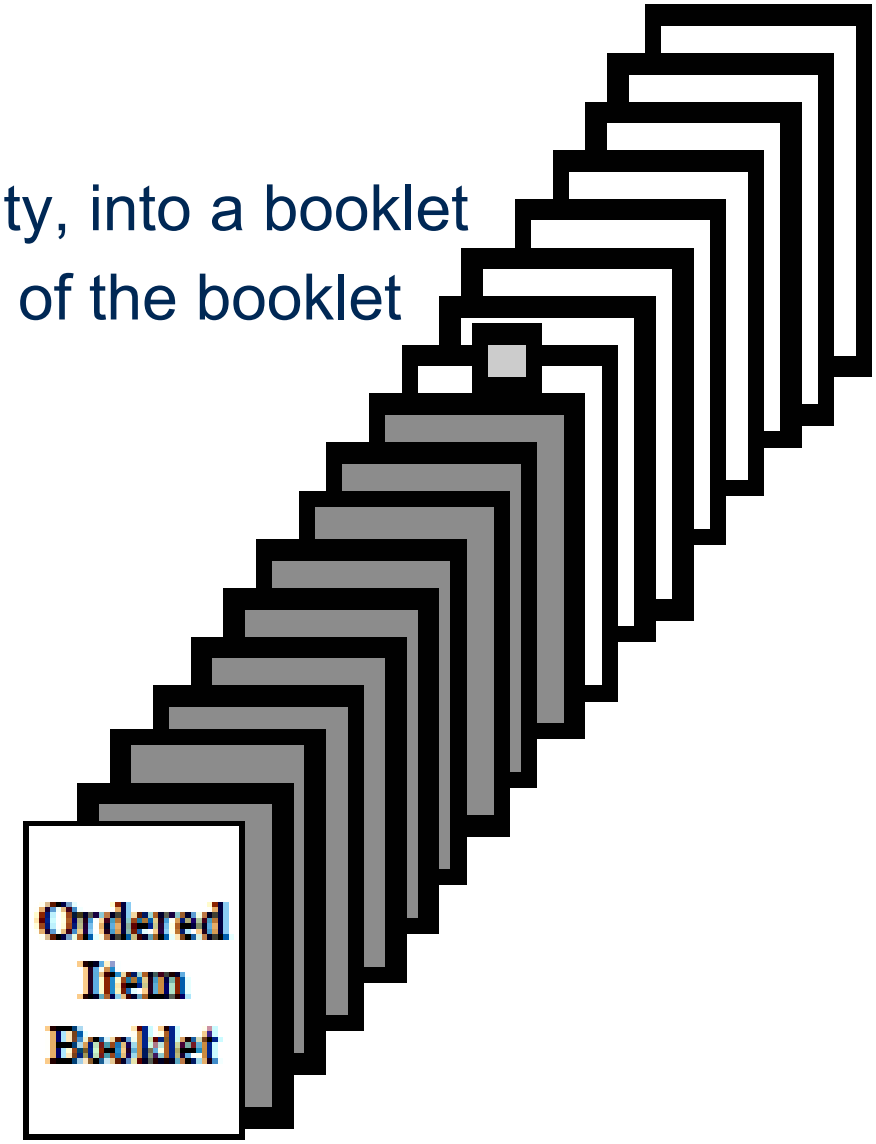
PROPOSED METHODOLOGY OVERVIEW

1. Bookmark Method
2. Assessment Targets
3. Logistics
    i. Panel of Experts
    ii. Meeting Location
4. Process

Bookmark Method

BOOKMARK

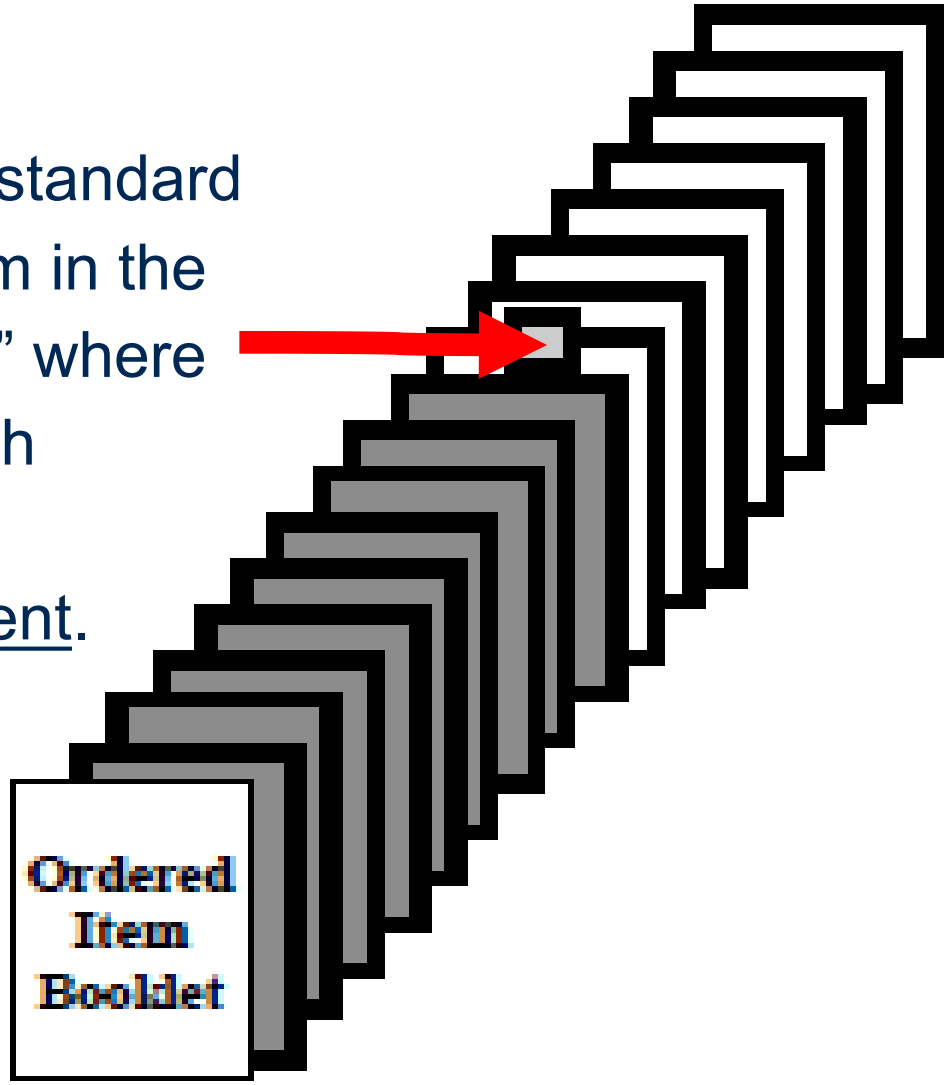
- Order test items, based on difficulty, into a booklet
  - One item appears on each page of the booklet
  - Easiest items placed first
  - Hardest items placed last
  - Items increase in difficulty



OPA
OFFICE OF PEOPLE ANALYTICS

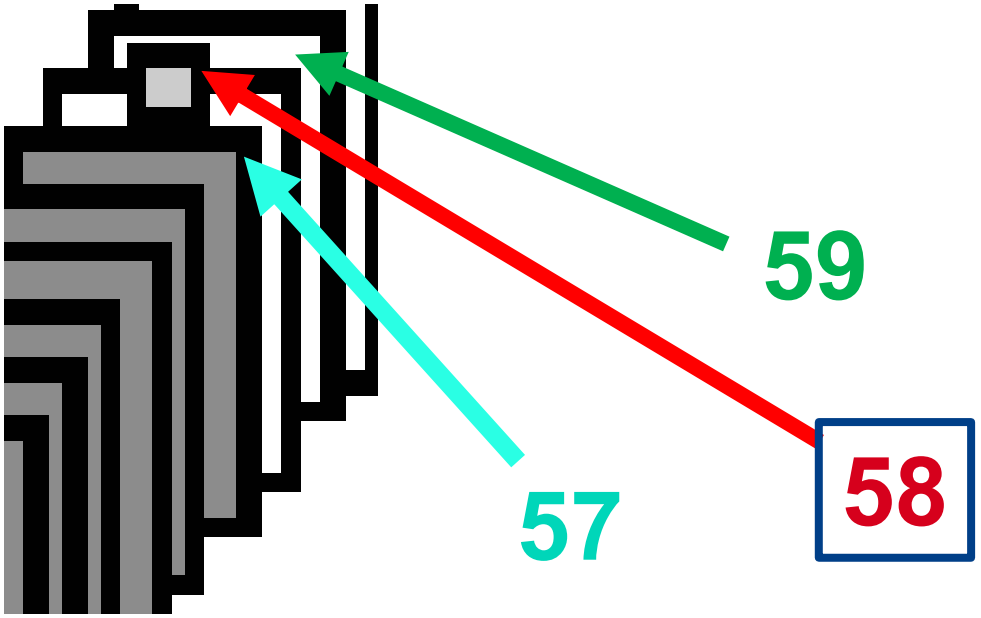
BOOKMARK

- Those invited to participate as standard setting panelists select a test item in the booklet that represents the “spot” where applicants have mastered enough of the content to be considered significantly exposed to the content.

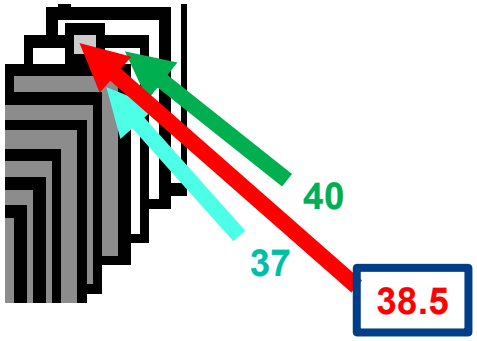


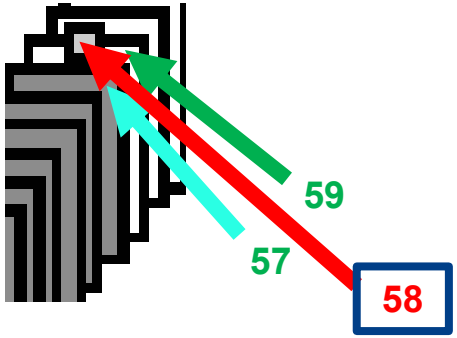
BOOKMARK

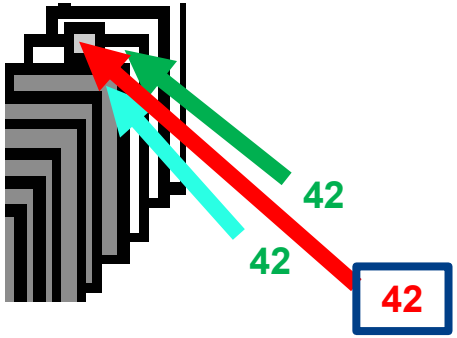
- Panelists place bookmarks (representing the midpoint between two items where the bookmark “sits”).
- This midpoint is averaged across panelists to determine the cut score.

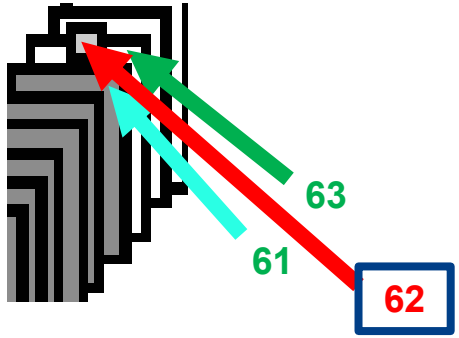


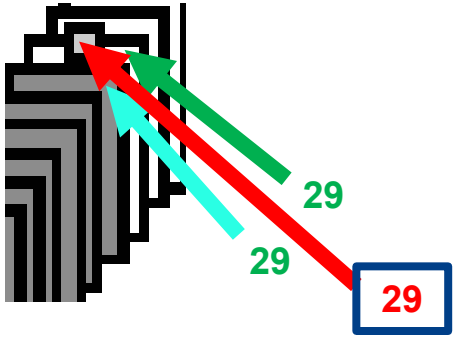
BOOKMARK

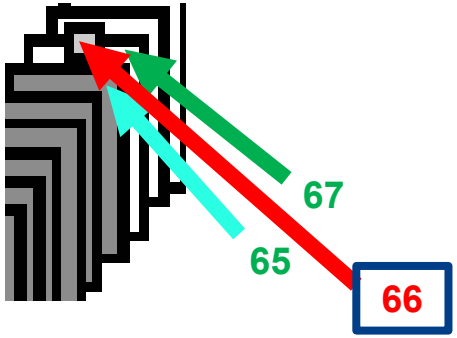


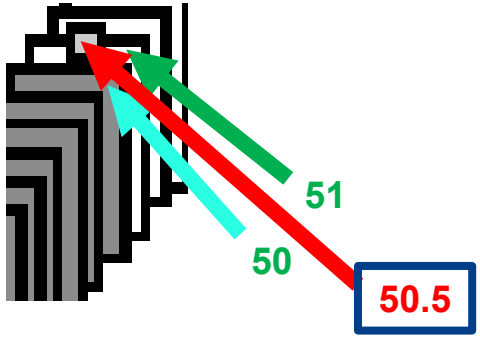


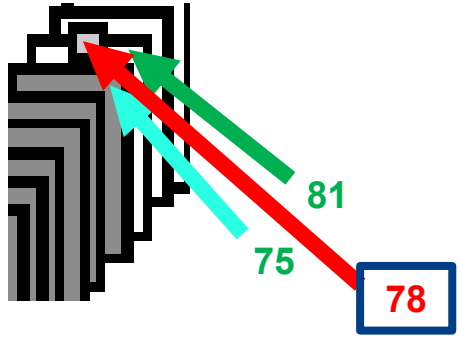












Cut Score = 53

Assessment Targets

ASVAB SUBTESTS UNDER EXPLORATION

—CONSTRUCT DEFINITIONS—

- **Automotive Information (AI)**
  - Knowledge of automobile technology
  - Items are designed to measure an examinee's basic knowledge, procedures, and principles, including automotive repair

- **Electronics Information (EI)**
  - Knowledge of electricity and electronics
  - Items are designed to assess an examinee's aptitude for understanding electrical currents, circuits, devices, and systems

ASVAB SUBTESTS UNDER EXPLORATION

—CONSTRUCT DEFINITIONS—

- **Shop Information (SI)**
  - Knowledge of tools and shop terminology and practices
  - Items are designed to measure an examinee's basic knowledge of general shop practices and building construction

- **Mechanical Comprehension (MC)**
  - Knowledge of mechanical and physical principles
  - Items are designed to assess an examinee's aptitude for principles of mechanical devices, structural support, and basic properties of certain materials. Problems cover the principles of gears, pulleys, levers, etc., as well as force and fluid dynamics. Items require general reasoning skills and the manipulation of spatial concepts

ASVAB SUBTESTS UNDER EXPLORATION
—CONSTRUCT DEFINITIONS—

- **General Science (GS)**
  - Knowledge of physical and biological sciences
  - Items cover three content areas: Life Science, Physical Science, and Earth/Space Science. The items are designed to measure the examinee's ability to recognize, apply, and analyze scientific principles, including the facts, concepts, theories, and laws of science.

- **Cyber Test**
  - Knowledge of information and communications technology
  - Items are designed to assess examinee's aptitude for networking and telecommunications, computer operations, security and compliance, and software programming

Logistics

SELECT A PANEL OF EXPERTS

Two of each, per subtest:

- Military training personnel
- High school vocational teachers
- Post-secondary vocational/community college instructors
- Members of associated business community

  - Panelists should be considered subject matter experts in the content the subtest measures
  - Panelists should understand the variation of knowledge and skills of the youth population as it relates to the content area

SECURE A MEETING LOCATION

- The standard setting process is best conducted with all panelists gathered at the same location, typically a hotel conference center*

- Expected time commitment: 3 days for each subtest

Process

STANDARD SETTING MEETING AGENDA

- Orientation to standard setting
- Have panelists take example test
- Self-score the test
- Implement Bookmark Method
  - Two rounds
  - Feedback between rounds
    - Distribution of cuts across panelists
    - Impact data including % classified as “significantly exposed”
    - NAEP High School Transcript Study longitudinal results
- Calculate cut scores
- Develop a definition of “significant exposure” for each subtest
- Evaluate standard setting process

OPA
OFFICE OF PEOPLE ANALYTICS

DACMPT ROLE

- Offer technical advisement on DPAC's proposed methodology
  - Choice of method: Bookmark
  - Number of panelists/subtest: 8
    - given low stakes nature for using results
  - General meeting process
    - Outline/tasks (Agenda)
    - Feedback offered (Agenda)
    - Number of rounds: 2
    - Number of cuts: Would 2 be better?
    - Amount of time: 2 ½ days for meeting

- Offer technical advisement on whether the outlined approach will sufficiently answer our questions (DPAC Goals)

Tab P

OPA

OFFICE OF PEOPLE ANALYTICS

Future Topics

Daniel O. Segall

Briefing presented at a meeting of the Defense Advisory Committee on Military Personnel Testing, 26-27 September 2019

Future Topics

- ASVAB Resources
- ASVAB Development
  - Pool Development
  - Evaluating/Refining Item & Test Development Procedures
  - Item writing guidelines and tools
- Adverse Impact
- PiCAT/Vtest Updates
- APT
- TAPAS Evaluation
- Test Security/Compromise
- ASVAB Validity
  - Improving the Validation Process and a review of the Service validity studies
  - ASVAB Validity Framework
  - Criterion Domain / Performance Metrics
- Career Exploration Program Updates
  - Web Site
  - iCAT Expansion
- Adding New Cognitive Tests
  - Cyber
  - Working Memory
  - Abstract Reasoning (including Adverse Impact)
- Adding New Non-cognitive Measures
  - Personality and Interest Measures
  - AVID
- Automatic Item Generation
- Web and Cloud efforts
- Device Evaluation and Expansion
- ASVAB Evaluation
  - Standard Setting Study
  - Other evaluation efforts

OPA
OFFICE OF PEOPLE ANALYTICS