DEFENSE ADVISORY COMMITTEE ON MILITARY PERSONNEL TESTING

March 28-29, 2019 Meeting

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

Minutes approved for public release.

Michael Rodriguez June 17, 2019

Dr. Michael Rodriguez, Chair, DACMPT DATE

DEFENSE ADVISORY COMMITTEE
ON
MILITARY PERSONNEL TESTING

Carmel-By-The-Sea, CA
March 28-29, 2019

The meeting of the Defense Advisory Committee on Military Personnel Testing (DACMPT) was held at the Pine Inn, Carmel, CA on March 28-29, 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 of 1972 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 then thanked the distinguished committee members and presenters and informed the audience that one committee member, Dr. Kevin Sweeney, was unable to attend. She also shared the news that Dr. Nancy Tippins, who she said had an amazing reputation, had been approved to join the committee, and that AP was thrilled and honored to have her. Finally, she reported that Ms. Stephanie Miller has returned to her post as Director of AP. 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 to include Joint Advertising Market Research & Studies (JAMRS) Brief (Tab D)

Dr. Sofiya Velgach, Assistant Director, AP, presented the briefing for Ms. Stephanie Miller, Director, AP.

> Dr. Velgach began by summarizing the mission of AP, which is to "develop, review, and analyze policies, resources, and plans for Services' enlisted recruiting and officer commissioning programs." She then presented an organizational chart detailing the structure and programs within AP. A table displayed recruiting outcomes for the Active Components as of February 2019. Regarding active duty components, the Army is slightly behind recruiting goals, having achieved 97% of goal, while all other Services are on track. Among Reserve Components, only the Navy Reserve is behind mission, having reached 92% of goal. Recruiting quality goals include accessing 90% high school degree graduates, with 60% or more in the Armed Forces Qualification Test (AFQT) I-IIIA range, and 4% or less in Category IV. All Services, active and reserve, have met this goal except the active Army where 59.7% of recruits fell in the CAT I-IIIA range.
>
> Dr. Velgach continued by summarizing data provided by JAMRS, which indicate that the youth market is disconnected from today's military, resulting in fewer youth being interested in serving. A lack of familiarity with the military leads youth to rely on stereotypes of military life. Outreach efforts need to be deliberate and sustained. Influencers who are familiar with the military are more likely to support youth service. Outreach to influencers should build awareness and advocacy for service. Dr. Velgach presented statistics demonstrating these points. In 1995, approximately 40% of youth had a parent who served in the military; by 2017 this figure had dropped to 15%. When asked how knowledgeable they are regarding military service, just over half of youth (51%) say not at all knowledgeable. Only 27% of adults age 17-35 can name all five active duty military branches, and 35% do not know there is a difference between an enlisted and officer person. Perceptions of the advantages of military service have dropped steadily over the past decade. In 2004, 85% of youth 16-21 agreed that joining the military would provide money for

college; in 2017 this number had dropped to 59%. Similar results were shown for preparing for a future career, having an attractive lifestyle, and staying in contact with family and friends. At the same time, majorities of 16-24 year-olds think it is likely that someone leaving the military will experience psychological or emotional problems (65%), difficulty readjusting to everyday life (64%), and/or physical injury (57%). Regarding sources of information about the military among 17-35 year-olds, 53% cite the media, 68% personal connections, and 18% Service outreach, with the media providing the most negative impressions. Data on influencer opinions indicate that parents are less positive about military service than grandparents, but parents are also less connected to the military.

As Dr. Velgach presented the status on recruiting goals for FY2019 (slide 5), a committee member asked how the goals were set. Dr. Velgach replied that each Service sets its own goals, for example, the Navy sets goals based on targets it can meet in consideration of retention. She explained that, if retention is high, accession requirements are adjusted downward accordingly. The committee member then asked about the responsiveness of budgeting to force needs. Dr. Velgach responded by clarifying that budgets are established five years in advance such that current recruiting requirements inform budgets in future years. Another committee member asked about the impact of exceeding recruiting goals. Dr. Velgach replied that, in the officer domain, over-execution has stimulated discussions of whether exceeding the target would be beneficial. She explained that unexpectedly high graduation rates can initiate these discussions.

Regarding findings on influencers (slide 12), a committee member asked why mothers, more so than fathers, support enlistment decisions. Dr. Shannon Salyer (Defense Personnel Assessment Center; DPAC) replied that mothers tend to be more supportive, but *only after the decision to enlist has been made*.

2. <u>Career Exploration Program (CEP) Update</u> (Tab E)

Dr. Shannon Salyer, DPAC, presented the briefing.

Dr. Salyer began by presenting Armed Services Vocational Aptitude Battery (ASVAB) CEP numbers and metrics for school years 2013 through 2018. These showed the number of students tested ranged from 670,836 in 2013 to 713,777 in 2018. The percentage of schools tested ranged from 55% in 2018 to 57.2% in 2016. For the first seven months of school year 2019, 559,375 students tested in 10,490 schools. Of these, 662,564 students took the paper-and-pencil (P&P) ASVAB, while 51,213 took the CEP *i*CAT. The number of accessions who used their CEP scores for military entrance ranged from 28,233 in 2014 to 30,257 in 2017. Thus far in 2019, 15,581 students used their CEP ASVAB scores to enlist in the military.

Dr. Salyer continued by presenting five options for eliminating P&P testing in the CEP, along with their drawbacks. She concluded that P&P testing should continue for schools that lack the infrastructure to accommodate computer-based testing. Backup P&P forms need to be identified in the event of test compromise. Steps have been taken to increase use of the *i*CAT in the program, including assigning points of contact (POCs) to address connectivity issues, requesting random access memory (RAM) and server increases, and transitioning testing to the Cloud. Next steps include allowing *i*CAT administration on 12" monitors, expanding browsers to include Safari, and continuing to identify potentially compromised test material.

Dr. Salyer then turned to the results of a CEP needs assessment. Efforts included reviewing past CEP iCAT pilot reports, conducting site visits of CEP ASVAB administrations, and talking with DPAC and Military Entrance Processing Command (MEPCOM) stakeholders. General recommendations included reinforcing rules regarding proctor behavior (e.g., no cell phone use during testing) and instituting a database management system to issue test session IDs. For P&P administrations, one recommendation was to review the instructions given to examinees to reduce redundancy. A variety of actions were suggested regarding

CEP *i*CAT, including eliminating data fields that are not routinely completed by examinees, addressing bandwidth issues, and ensuring help desk support is available when testing is taking place. Additional potential actions include investigating allowing educational support specialists (ESS) to access score reports with a username and password (as opposed to a Common Access Card [CAC]), which could allow for same day testing and test interpretations. Post-test interpretations (PTIs) should be standardized and focus on score reports, the Find Your Interest (FYI) inventory, and career exploration, as well as providing an overview of the array of other information available on the website. In addition, alternative methods of providing PTIs should be explored given the potential for large increases in the number of students participating in the CEP.

Dr. Salyer then discussed state usage of ASVAB CEP. Four states (TX, IN, AZ, UT) have legislation requiring that the CEP be available statewide. In addition, 16 states have legislation that addresses some segment of the state, either by school district or type, and two others are considering legislation. As of August 2018, 12 states use the ASVAB CEP in some capacity as a career exploration tool, and four others specifically cite the CEP, often as part of graduation or college/career readiness requirements. Dr. Salyer indicated that she supplied a memo to the field regarding the appropriate use of the ASVAB CEP. She has also included guidance in this regard during presentations to state officials and at national conferences. Efforts to monitor state usage of the CEP are ongoing.

Dr. Salyer next discussed efforts to provide PTI proficiency training to standardize the process by which the sessions are conducted. This project involved stakeholders from Accession Policy, Office of People Analytics (OPA), MEPCOM, the Army, Navy, Army National Guard, Air Force Reserves, and the Coast Guard. Among the reasons for embarking on this effort are the increased use of the program, the need to ensure it is standardized across sites, the need to update field personnel on the expanding functionality of the program websites, and the desire to establish more solid contacts with the CEP's national workforce. To become PTI proficient, individuals must be nominated by someone who is already designated as proficient, complete the virtual training modules, be observed conducting a PTI, and submit proof of proficiency. The virtual training addresses ASVAB measurement and data use, interpreting and discussion ASVAB scores, components of the ASVAB CEP, conducting a PTI, and becoming PTI proficient. It includes objectives, learning goals, multimedia content, concept checks, and reflection and application activities. Dr. Salyer then presented some comments received following the first training session, which included high praise and some suggestions for improvement.

For her final topic, Dr. Salyer discussed the recommendations of the CEP Expert Panel convened in 2017 regarding the FYI inventory. These include doing a thorough review of the FYI with the goal of revising outdated items and adding new items to address a broader array of interests and integrating basic interest scales into the CEP. Subsequent research on the FYI revision involved conducting a preliminary expert review of the items, conducting a content analysis of the current FYI Holland Occupational Themes (RIASEC) scales by ethnic and gender groups, and doing in-depth statistical analyses of existing FYI data. The statistical results showed strong internal consistency for the RIASEC scales for both males and females. Males scored higher than females on the Realistic and Investigative domains, while females scored higher on the Social and Artistic domains. There were no significant gender differences on the Enterprising or Conventional scales. Structural analyses suggested that the current FYI items poorly fit the RIASEC model for males. A likely reason is that the item selection procedure that mirrored interrelations among the RIASEC types was not applied to the male sample. A revision of the item pool should use the same procedures with males as was used with females to select items. It is also important to include all racial/ethnic groups in the item selection process, including White students. Dr. Salyer concluded by stating that DPAC is continuing to review the panel's recommendations and discuss the way forward and will keep the committee informed about future efforts.

As Dr. Salyer presented the numbers and metrics for the 2018-2019 school year (Slide 4), a committee member asked how Career Exploration Program (CEP) participation rates fluctuated across the school year. Dr. Salyer replied that participation was extremely heavy at the beginning of the year and the end of the year, which portends a likely increase in the yearly total from 2018 to 2019.

When Dr. Salyer presented recommendations on eliminating the P&P version of the CEP (slide 8), a committee member asked if eliminating the P&P form was still an objective. Dr. Salyer replied that eliminating the P&P CEP was still the goal, but she said significant obstacles (e.g., limited bandwidth at schools) persisted.

Regarding the development of backup P&P forms for use in the event of a compromise (slide 9), a committee member asked if there was only one CEP form. Dr. Salyer said yes, but that a second form was being developed. She then described how increased use of the CEP *i*CAT (Internet CAT [Computer Adaptive Test] ASVAB) was resulting in more frequent connectivity problems and server crashes, which prompted Dr. Velgach to explain that technical issues were one reason for delaying the transition from P&P. Another committee member asked about the timeline for transitioning to the Cloud. Dr. Mary Pommerich (DPAC) said the topic would be covered in the milestone briefing.

When Dr. Salyer said many schools were purchasing smaller (e.g., 12") monitors (slide 10), a committee member asked if those were real monitors as opposed to tablets. Dr. Salyer said they were and noted that similarly-sized monitors had been used to conduct the original *i*CAT research. Dr. Dan Segall (DPAC) added that his team had reviewed prior compatibility studies investigating DOS and early Windows machines that used 10" and 11" monitors.

On the recommendations for P&P and *i*CAT administrations (slide 13), Dr. Salyer noted an issue with documenting session numbers; she said session information is recorded differently on the P&P and *i*CAT. A committee member asked if session documentation included location, date, and time, and Dr. Salyer said it did.

Regarding recommendations related to *i*CAT administration (slide 14), a committee member asked if the recommendation to address issues regarding screen resolution was specific to administration in schools. Dr. Salyer said it was, due to the lack of uniformity of computer labs across schools. She added that, even within a school, screen resolutions get changed, and test administrators (TAs) must check for that.

As Dr. Salyer discussed the state of CEP usage across states (slides 17-18), a committee member commented that one of the states requiring state-wide CEP availability (i.e., Texas) was quite large. Dr. Salyer replied that she has only eight people in Texas, which she said made it tough to avoid the appearance that the program is either targeting or avoiding certain areas and schools. She commented, however, that laws tend to change, which made it difficult to plan long term. The committee member stated that higher usage trends fundamentally alter the playing field for the program, and Dr. Salyer agreed.

A committee member commented that most states function similarly in how they roll out Federally-mandated tests; that is, they train assessment coordinators in test administration to meet Federal requirements for maintaining school accountability. Dr. Salyer replied that most of the CEP modules tie into a focus on college readiness. She added that the CEP is useful in this domain due to its broad coverage of career areas. The committee member then noted that the program could benefit from the interest and the associated increase in available TAs. Dr. Salyer said the program is already attempting to make use of those resources.

A committee member asked about the yield of CEP administration in Puerto Rico, to which Dr. Salyer replied that the language barrier reduces the number of recruits that might otherwise be accessed. Dr. Velgach referred to continuing requests that the ASVAB be provided in Spanish but noted that one issue with doing so would be the subsequent demand that the test be made available in other languages as well. She explained that military training and performance are in English and said this provides a practical reason to test only in English.

On the use of the ASVAB CEP by states for school accountability purposes (slide 19), Dr. Salyer explained that the program provides a very explicit statement of the test's purpose in a memorandum that addresses the program's use in relation to Every Student Succeeds Act (ESSA) requirements. She said the memorandum states that the CEP is not appropriate for any use other than for career exploration. Dr. Velgach added that AP has been concerned with how CEP data might be used in the future. Dr. Salyer gave the example that Tennessee wants to collect CEP data from schools to assess work force readiness and determine if ASVAB is predictive of critical outcomes for the State.

As Dr. Salyer addressed the PTI proficiency training (starting at slide 21), she referred the committee to the PTI Proficiency Training binder and described the Evaluation Metric tab. A committee member commented that the metrics cover more than just interpretation but also the intended use of the program. Dr. Salyer agreed and explained that the intent of the training is not to be prescriptive about how to use the program, but to educate users on the many ways the program *can* be used. A committee member asked if the training is designed to help TAs communicate with parents and other influencers, and Dr. Salyer said it is. After Dr. Salyer described the remaining content of the binder, a committee member said the training is an amazing tool and should be used in schools because it addresses topics such as how to work with parents and students (e.g., what to do, and what not to do). Dr. Salyer replied that many people had provided input into the training's development.

Drs. Velgach and Salyer then discussed the ability of the training to (a) reduce the strain on the MEPCOM budget and (b) allow students to interact with Recruiters without feeling pressure to join the military. Dr. Salyer also reported how one ESS in Tennessee had asked if she could give all her teachers access to the virtual training tool and then work with them to get them involved. A committee member commented positively on the value of having a training program to support CEP implementation. Dr. Velgach added that training also adds the ability for participations to interact and build needed relationships, specifically, the "homework" portion was designed to foster relationships and collaboration among ESSs and between ESSs and recruiters.

As Dr. Salyer presented the results of the FYI research (slide 32), a committee member asked if gender differences impacted any group negatively. Dr. Salyer explained that they did not because the scores are used for career guidance not for selection or classification. Additionally, FYI scores are presented in both ways: as gender normed and combined. She suggested, however, that gender fluidity may become an issue in future reporting of results by gender. Another committee member asked if Dr. Salyer planned to continue the use of the RAISEC model even though the FYI items seemed to fit that model poorly. Dr. Salyer explained that she did and explained how items could be modified to better fit the model by making them more realistic while, at the same time, maintaining relationships with adjacent constructs. Dr. Pommerich commented on the difficulty of measuring interests due to how much they can fluctuate. She said DPAC is collaborating with the Army Research Institute for the Behavioral and Social Sciences (ARI) in its development of the Adaptive Vocational Interest Diagnostic (AVID).

A committee member asked if Dr. Salyer planned to incorporate military contexts into the FYI. Dr. Salyer responded that the Services are developing separate inventories that meet that need. She also said the military is working on a measure for use by Service members who are transitioning out of the military, and DPAC will be able to see how the FYI norms to the military populations.

At the end of the briefing, a committee member asked about the plan for developing career counselors, and Dr. Salyer replied that the PTI proficiency training can be a stepping stone toward obtaining the Certified Career Counselor Credential offered by the National Career Development Association.

3. **Milestones and Project Schedules – (Tab F)**

Dr. Mary Pommerich, Deputy Director, 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 CEP, ASVAB and Enlistment Testing Program (ETP) revision, the Air Force Compatibility Test, and the Defense Language Aptitude Battery (DLAB).

- New 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 May 2020.
- Developing New CAT Item Pool for the CEP. The objective of this project is to build a CAT pool from 20 B, 21 A&B, and 22 A&B for implementation of the *i*CAT in the CEP. The new pools will be implemented in the Spring 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 September 2019.
- Automated Generation of GS items. The objective of this effort is to develop procedures for automating General Science (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 Measures.
    - 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 WPA; and Army, Air Force, and Navy interest inventories. The project is ongoing.
- Air Force Compatibility Assessment (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.
- Web/Cloud Delivery of Special Tests. The objective of this effort is to transition delivery of special tests from a Windows-based platform to a web-based and/or cloud platform. The anticipated completion date is December 2021.
- Expanding test availability by moving to the Cloud. The objective of this project is to examine the feasibility of moving all test delivery to the Cloud. Project completion is scheduled for December 2021.

As Dr. Pommerich explained the seeding of AR and MK items developed through automated item generation (AIG), a committee member asked what she meant when referring to "enemy items" in this context. Dr. Pommerich explained that an item generated from an existing item (i.e., a clone) is considered an enemy of the existing item because their commonalities should prohibit administration of both items to the same person.

As Dr. Pommerich briefed the ASVAB technical bulletins (slide 9), a committee member asked when new CEP versions would be available and how many CEP P&P versions were being developed. Dr. Pommerich replied that when Forms 11-15 are available for use at Military Entrance Processing Stations (MEPS), Forms 5-9 can be repurposed for CEP use. She added that DPAC will no longer be developing new P&P forms. However, she said DPAC was working with AP to repurpose a retired enlistment form into two P&P backup forms to be used if the current CEP form is compromised. Dr. Pommerich also reported that DPAC was working to reduce connectivity problems experienced at schools. A committee member then asked about the prospects for increasing *i*CAT usage. Dr. Pommerich said, "maybe when we get to the Cloud." Dr. Segall then noted three issues with iCAT delivery: (a) testing capacity is limited due to information technology constraints, (b) computer and connectivity resource limitations at schools, and (c) form availability. He said there should be significant improvement in *i*CAT administration over the next five years, but that it will not happen overnight, which is the reason for the near-term backup plan. Dr. Velgach clarified that DPAC is evaluating code to make sure it is effective and exploring ways to increase capacity (e.g., increasing the number of application servers available). She also mentioned that device flexibility and bandwidth are particularly difficult challenges. Mr. Paul Aswell (US Army G-1) asked if the device evaluation effort could result in increased CEP *i*CAT administration, but a committee member replied that the issues associated with increased CEP *i*CAT usage went beyond device expansion.

As Dr. Pommerich briefed slides 22-25, she asked Dr. Cristina Kirkendall (ARI) about the status of the Work Preferences Assessment (WPA), and Dr. Kirkendall explained that ARI is still collecting data at MEPS, but that they hope to have something by May 2019. Dr. Greg Manley (DPAC) identified "SDI" on slide 23 as the Self-Description Inventory, a personality inventory developed by Dr. Manley based on the original USAF work pioneered by Tupes and Christal.

A committee member asked how the AFCA (slide 26) would be used on the platform. Dr. Velgach said it would be used in an exploratory manner to determine its proper use. She said there are a variety of possible uses, ranging from selection to identifying individuals who would benefit from coaching or mentoring. She also said the Sexual Assault Prevention and Response (SAPR) office is interested in using the AFCA to decrease the rate of sexual assault. The committee member expressed concern about putting the AFCA on the platform without sufficient validation evidence and an approved purpose; s/he said that might invite inappropriate use, especially selection. Dr. Pommerich clarified that, though it would be on the platform, it would be "turned off" to restrict access. Dr. Velgach added that it would not be administered until all approvals have been granted.

A committee member returned to the topic of interest inventories and inquired about linkages among the AVID, Air Force Interest Inventory (AF-WIN), and Job Opportunities in the Navy (JOIN) personalized career interest assessment. Dr. Pommerich explained that the Services owned those inventories. Dr. Velgach explained that the JOIN and AF-WIN use a similar model for organizing occupations, but the AVID uses its own model. The committee member asked if the inventories were tailored based on the needs of the individual Services, and Dr. Velgach said they were. The committee member then questioned the applicability of the AFCA across Services, at which point Dr. Velgach said that the plan is to assess applicability across DoD. Dr. Segall explained that the AFCA was designed to measure AF Core Values, but that core values should apply to the other Services as well. Dr. Velgach then clarified that the AFCA is an integrity test that looks at abusive behaviors, calling it a type of moral character assessment.

As Dr. Pommerich elaborated on the difficulty of transitioning to the Cloud (slides 29-32), a committee member asked if DPAC had started to test delivery functionality. Dr. Segall replied affirmatively and added that the Device Study is administering over the Web using a prototype architecture. He said the main obstacle has been obtaining the authority to operate on the Cloud, which has required DPAC to focus on developing security-related protocols and controls. He then explained that in the interim, DPAC is moving off Windows to an Internet system delivered through Defense Manpower Data Center (DMDC), which has raised resource concerns. He added, however, that programming the *i*CAT for web delivery now would ease the eventual move to the Cloud. Dr. Pommerich said the date for moving to the Cloud would probably slide because MEPCOM, which is also making the transition, has experienced its own delays.

A committee member asked if moving to the Cloud was driven by the Force of the Future (FOTF), now called Testing Modernization, initiative. Dr. Segall said it was, and that the FOTF was an initiative to gain additional funding over a five-year period to modernize aspects of testing administration. He said things were moving slowly because DPAC is on the cutting edge of the effort. The committee member asked if deployment to the Cloud would reduce DPAC's in-house control and agility to modify tests on-the-fly. S/he also asked if there was a technology partner maintaining the Cloud. Dr. Segall replied that Human Resources Research Organization (HumRRO) is the prime contractor, but that Northrup Grumman is the sub-contractor who is overseeing technical aspects of the transition. He then said moving to the Cloud would, in fact, increase DPAC's agility in several areas. First, he said DPAC would be able to add servers through a drop-down menu to accommodate greater testing loads. Second, he said new technology would be easier to integrate. Third, he said the current constraints on DPAC, due to it being only one of many DMDC clients, would be eased in that they would be able to release new packages directly without having to rely on DMDC. Dr. Pommerich added that DPAC would have an increased capability to access diagnostic information upon demand. Dr. Segall clarified that DMDC currently updates the system during testing hours, which interrupts service and interferes particularly with applicants who have travelled to test; he said these applicants have to go home without being able to complete their session. Concluding the discussion, Dr. Pommerich said DPAC would benefit from having more control, but that the increased responsibility was a little daunting. She said contractor support would be even more critical in the future than it is now.

4. <u>ASVAB Evaluation Plan</u> (Tab G)

Dr. Mary Pommerich, Deputy Director, DPAC, presented the briefing.

> Dr. Pommerich began by providing an overview of her presentation, the goal of which is to provide background and updates on the status and plans for the evaluation ASVAB tests. The overall focus is on determining which tests should be administered as part of the ASVAB and which tests should be on the ASVAB platform. This involves (a) continuing efforts to evaluate and resolve (as needed) issues and concerns pertaining to the new tests of interest (i.e., TAPAS, Cyber Test, MCt), (b) continuing efforts to evaluate the tests currently in the ASVAB, (c) completing efforts to apply an arguments-based approach to validation of the ASVAB, and (d) reviewing and updating the psychometric checklist for the purpose of evaluating tests to be administered as part of the ASVAB. This will involve having the Services or proponents complete the updated psychometric checklist for new tests of interest and documenting all new information since the checklist was last completed. Stakeholders will develop a shared vision that defines the purpose and general makeup of the next generation ASVAB. It will also be necessary to establish a systematic process to follow for evaluating potential changes and making decisions regarding tests to be included in the battery. Logistical questions will need to be addressed with stakeholders, including the feasibility of lengthening the ASVAB and/or dropping existing tests. Stakeholders will have to summarize the impact of potential modifications to the battery and identify resources to support a revised battery. After all information is compiled, discussions will ensue about potential changes to the contents of the ASVAB and tests administered in the ETP.
>
> The Services and DPAC are continuing to study new tests of interest with an eye toward use with the next generation ASVAB. These include the Cyber Test, MCt, TAPAS, and Abstract Reasoning. Total testing time across the ASVAB and special tests, as well as potentially dated content, continue to be a concern. Therefore, there is a strong interest in assessing how the ASVAB might be modified to accommodate new tests. Potential ways to do so include dropping, combining, or shortening existing tests, or merging new and existing tests. Although research has been ongoing to evaluate new tests of interest, a similar level of scrutiny has not been applied to those already included in the ASVAB. A comprehensive assessment of the current tests will give insight into their utility, quality, and potential modifiability.
>
> DPAC has initiated an extensive plan to evaluate the current ASVAB tests to determine their desirability or expendability.

1. The first step is to trace the history of each of the tests currently in the battery to document where they came from and why they were originally included. Information regarding Assembling Objects (AO) and Paragraph Comprehension (PC) has been found, while the others are still in progress. PC was included to increase literacy requirements in the AFQT in response to findings that recruits had difficulty reading the instructional materials in their training courses. AO was one of the best of the nine measures that were part of the Enhanced Computer-Administered Testing project.
2. Step 2 is to complete psychometric checklists for the current tests and evaluate the psychometric value and limitations of each. Dr. Pommerich then presented the results of this process for PC and AO.
3. Step 3 is to evaluate the usefulness and appropriateness of existing tests with regard to the current population. This involves tracking test score trends over years 1984-2019, evaluating what fraction of the population possesses the knowledge/skill assessed by the test, evaluating the overlap between latent ability and score information for the current test population, and conducting pseudo-standard settings and evaluating the percent in each category over time for the technical tests. Dr. Pommerich indicated that the data needed to conduct these analyses have been partially located, the analysis programs are in place, and planning for the standard settings for the technical tests is in progress.
4. Step 4 is to identify estimated yearly development costs for each test. This includes identifying per-item development costs, the desired replacement schedule, the number of items needed per year per test, and the total per-year cost. This has been completed.
5. Step 5 is to evaluate the overall ease or difficulty in developing good items for each test, including the finiteness of each of the domains (more finite = more difficulty), the feasibility of using AIG (less feasible = less ease), and the item retention rates (less retention = less ease). This work has been partially completed.
6. Step 6 is to evaluate how likely the content of each test is to stand the test of time, including (a) how relevant the content is to today’s applicant population, (b) how prone the content is to obsolescence, (c) the degree to which content is in need of frequent updating to stay current, and (d) the extent to which it is difficult to keep up with new technology or changes in technology. Rating scales are being developed for each task.

- Step 7 involves evaluating the efficiency of content coverage based on prior research. This includes identifying redundancies in content coverage across tests as well as gaps in coverage and potentially unnecessary content coverage. This will require a literature review.
- The goal of Step 8 is to evaluate the vulnerability of item content and item pools to compromise by identifying the features of tests and item pools that could leave them susceptible to compromise and examining previous incidents of compromise and the tests that were breached. Information gathering and review of prior compromise history in the ETP, Armed Forces Classification Testing program, and the CEP is underway.
- Step 9 focuses on the vulnerability of each test to other unwanted effects, such as coaching, practice, hardware and mode effects, and local dependence. This will involve reviewing prior findings for the ASVAB and developing rating scales to summarize vulnerability across the various factors.
- Step 10 focuses on the efficiency of each test regarding testing time allotted and testing time used. This can be accomplished by summarizing the total testing time allotted to the CAT-ASVAB, the observed testing times for applicants overall and per test, and the allocated versus actual time spent per item and per test.
- The final step in the initial evaluation is to synthesize the findings across all criteria and summarize the desirability or expendability of each test. This will involve identifying a way to concisely summarize the results over all steps and a way to aggregate the findings and compute an overall rating. Dr. Pommerich indicated that suggestions for accomplishing this goal are appreciated.

Dr. Pommerich concluded by outlining future steps, which include evaluating the feasibility and/or psychometric impact of (a) shortening various tests, (b) shortening AR and/or MK and computing a math composite score, (c) combining AR and MK into a single test, (d) combining Electronics Information (EI) and the Cyber Test into a single test, and (e) dropping Auto Information (AI), Shop Information (SI), AO, EI, Mechanical Comprehension (MC), General Science (GS), or WK.

As Dr. Pommerich briefed progress on the Next Generation ASVAB (slides 4-6), a committee member asked if the Coding Speed (CS) test would be retained. Dr. Pommerich explained that CS was a special test used by the Navy. Dr. Segall added that the test had been dropped from the battery but was still on the platform and would be transitioned to the Cloud along with the other tests. The committee member then asked for a list of tests that are (a) on the platform and (b) in the battery. Dr. Pommerich said DPAC could provide that information.

Regarding average testing times for the various tests (slide 7), a committee member sought further clarification on which tests were in the battery. After Dr. Pommerich responded, the committee member asked if the times shown for those tests were average testing times, and Dr. Pommerich said they were average times for the special tests and total testing times for the ASVAB subtests. A committee member noted the extensive time required by AR (i.e., 39 minutes total).

As Dr. Pommerich briefed Step 2, evaluating usefulness and appropriateness of the tests (slide 14), a committee member asked how DPAC would determine the meaning of “proficiency” and “non-proficiency.” Dr. Segall responded by using the Auto Shop (AS) test as an example. He said that 40 years ago, a greater percentage of students were exposed to the knowledge required to do well on that test. He said the challenge would be to set a proficiency level that identifies those who have had the opportunity to acquire specialized knowledge in the area. He noted that this group has shrunk, and normal test development and statistics would not provide an accurate reading of whether the test should still be given. He said, that is, the low number of proficient scorers might be interpreted by decision-makers as indicating the test should no longer be administered. The committee member then asked what DPAC planned to do with such scoring trends. Dr. Segall replied that, once cut scores have been determined, DPAC can look at 30 years of data to see what percentages would have met the cut (e.g., 60% 30 years ago, 50% 20 years ago, and 10% today). He said they could then look at performance in relation to specialized knowledge versus nonspecialized knowledge to analyze classification efficacy.

On selecting participants for the standard-setting activity, a committee member suggested that participants would need to know the content as well as something about the relevant characteristics of the targeted population. Dr. Segall said participants would include some who have specialized knowledge and some who do not, because including both groups would help identify the cut point. The committee member then asked if Dr. Segall was talking about panelists, and Dr. Segall replied that he was. The committee member next asked if the task could be executed empirically, but Dr. Segall said DPAC currently did not have the requisite information. Dr. Tonia Heffner (ARI) asked what DPAC planned to do with the information; she said if only 10% have specialized knowledge, that might be a reason to drop the test, but she also said it could be that the 10% are who the Services need to identify. Dr. Segall replied that DPAC planned to use the results to determine whether the test should be given to everyone or only to those who show an interest in relevant jobs. Dr. Steve Watson (Navy Selection and Classification Policy) reiterated that the test may be very valuable but only as a special, complementary test used to place certain people in certain jobs. Dr. Heffner noted, however, that placement depends not only on proficiency, but also applicant preference for the targeted jobs. She said an applicant might know a lot about mechanics but may want to join the Infantry. Dr. Pommerich said that situation illustrated the difficulty of determining both value and usage, especially across the Services.

As Dr. Pommerich explained the process of evaluating the difficulty of developing good test items (slide 16), a committee member asked if there was a point on the finiteness scale at which DPAC would be concerned. Dr. Pommerich replied, "a four or a five." Another committee member asked about the rationale behind the scale. Dr. Pommerich replied that, for WK, for example, all the words are taken from a corpus. Dr. Segall explained that WK is a vocabulary test and the number of words available for testing is finite. He said the number of words in common usage is in the thousands, but that they write thousands of items each year. He said, to avoid using obscure words, they must use words that have already been used. The committee member asked what grade levels the WK items targeted, and Dr. Pommerich said third grade through college, but that the focus was at the high school level. Dr. Segall explained that, because they use item pools in an adaptive environment, most of the items administered were at the middle school and high school levels. Dr. Pommerich said the distribution revealed an inverse U-shape pattern. Another committee member asked if redundancy among subtests, that is, the degree to which subtests are inter-correlated, might be an elimination criterion. Dr. Pommerich said it was, and that slide 20 addressed that factor. Dr. Segall then stated that examining incremental validity was also a critical and related decision factor. The committee member suggested that tests such as Auto Shop and Mechanical Comprehension might be interrelated and that tests that have low finiteness ratings might cover similar constructs as those with more infinite item sets. Dr. Pommerich replied that DPAC has conducted factor analyses that showed some tests cluster together, for example, GS clusters with the verbal tests. A committee member noted that "they all have whispers of *g*."

When Dr. Pommerich described future steps (slide 25), a committee member asked if efficiency could be gained by skipping the remaining steps after a test has failed a step. Dr. Pommerich replied that the Services would want a full evaluation of each test. The committee member then asked whether positive results in early steps might cause DPAC to skip to the latter steps, but Dr. Pommerich said they were planning to conduct all the steps, regardless of the results. Another committee member agreed that a comprehensive evaluation would be best, at which point it was suggested that some solutions, such as combining AR and MK, might be low hanging fruit.

A committee member observed that some steps had clearer metrics than others. Dr. Pommerich said one of their tasks was to identify the best metrics for each criterion. The committee member suggested the application of a rubric. Dr. Heffner then asked Dr. Pommerich which tests could be added to the battery, and Dr. Pommerich replied that possible candidates for addition were the tests shown on slide 7 (Cyber Test, MCt, TAPAS, and Abstract Reasoning). She said all those tests would be evaluated as part creating the Next Generation ASVAB. Dr. Heffner then said she was thinking of other measures, such as systems thinking, and asked if there were other possibilities. Dr. Pommerich replied that they were not independently considering constructs or measures that the Services were not already researching. Dr. Velgach explained that a Service must have already completed the psychometric checklist for a test before it could be included in this evaluation. Dr. Heffner replied that certain tests might be worthwhile to include if they cover new constructs not represented in the current tests. Another committee member then asked about the interest inventories, but Dr. Pommerich said the interest measures were not being included because they are administered external to the ASVAB platform. She said that the Air Force and Navy administer their interest inventories at the Recruiting Commands. She also told the Service representatives that, if they have ideas for measures, they should bring those ideas to the Manpower Accession Policy Working Group (MAPWG) to get them in the flow. Dr. Watson replied that the Navy has several ideas, such as emotional intelligence, and asked if there was a disconnect between DPAC and the Services regarding what the Services are researching. Dr. Velgach suggested that the Services might bring their ideas to the MAPWG earlier. Dr. Segall said that was a possible solution but that DPAC does not want to know so early that they become distracted by tests that will fall out at some point. Dr. Watson proposed that it could be just a synopsis of what the Services were working on to increase DPAC's awareness of what might be coming down the road. Dr. Pommerich suggested Service snapshots of what they are working on, limited to five minutes.

At the end of the discussion, the committee members said they really liked the work DPAC was doing.

5. CAT-ASVAB New Forms Update (Tab H)

Dr. Matthew Trippe, HumRRO, presented the briefing.

> Dr. Trippe began the presentation by explaining that the goal of this project is to develop ASVAB forms on a more aggressive schedule. It will begin with Forms 11-15, which will be assembled from experimental items administered under old and new item seeding configurations. These will replace operational forms and P*i*CAT and will include one additional form over the original goal of four forms. Forms 11-15 will be assembled from ten experimental item series, or sets of 100 experimental items, per test. Each experimental item is reviewed for psychometric (e.g., model fit, information) and content quality. Items that survive the review process are moved on to form assembly. Dr. Trippe then presented a table showing the form development steps and status.
>
> Next, Dr. Trippe turned to enemy item identification. Local dependence (LD) analysis was conducted prior to assembly of Forms 5-9, and the results suggested that MK and MC are susceptible to LD. Mitigating LD requires identifying item enemy groups, including items likely to: (a) trigger LD if administered to the same person and (b) include two or more items that measure similar or highly related content. Before assembling Forms 5-9, DPAC developed a content framework for identifying enemy groups. The process involved evaluating 700+ items for match with enemy groups and resulted in 95 MC and 155 MK content areas. HumRRO developed a procedure to optimize human judgment and qualitative roles. For MC and MK, Method 1 involved two humans independently linking each item to the DPAC-defined categories and identifying new categories as necessary. Disagreements were resolved by a third rater. Method 2 involved unsupervised classification using text analysis of items and supervised classification analysis based on the existing DPAC content framework developed during Form 5-9 construction. The outcome being predicted is enemy group label. Method 3 comprised sparse data local dependence analysis and addresses the LD concerns directly through Q_3. It will be possible for seed series to be included in future form assembly efforts. For all other tests, unsupervised classification using text analysis of items and human review of "heatmap" hot spots were conducted.
>
> For MK, prediction was generally good at higher order group levels (e.g., angles). Prediction was not possible at the subgroup level (e.g., angles complementary, angles obtuse). Human judgment cannot be replaced, but a complicated and tedious task can be simplified and accelerated with model-based tools, such as a group assignment probability matrix and heatmaps. For MC, prediction was generally good. Much of the information in MC is stored in images/artwork, which at this point is difficult to quantify or tokenize for this type of analysis. Many group membership assignments are based on similarity to existing items in the group rather than an entirely discrete concept. There is more conceptual overlap between groups in MC than MK.

> Future work will continue to improve the process for MC. Dr. Trippe then presented a graphic showing a tool to facilitate MK item enemy review and a sample heatmap showing results for GS.
>
> Turning to form assembly, Dr. Trippe noted that CAT administration is based on forms from which a potentially unique set of items is administered to each examinee. Therefore, forms need to contain items from the full range of content and difficulty spectrums and they need to contain sufficient information/score precision across the full range of ability. The goals of form assembly are to (a) assign each item to one of five forms for each test, (b) maximize conditional precision levels for each form, (c) constrain conditional probability levels to be comparable across forms, (d) account for enemy items and distribute them evenly across pools, and (e) account for content taxonomies where applicable (e.g., GS, AO).
>
> Forms were assembled algorithmically to optimize the stated goals. This involved assembling the main analytic functions in Fortran as a dynamic-link library and developing an R package to wrap Fortran functions and to implement CAT analyses. This represents a "best of both worlds" approach by combining the speed of Fortran and the flexibility of R which facilitates changes to problem configurations, analysis of results, and promotes quality control. Test information was computed using CAT simulations. The entire item pool was partitioned into four or five candidate forms. Item exposure parameters were calculated using preliminary CAT simulations. A large sample of scored responses was generated using another round of CAT simulations and approximate test information was based on a sample of scored responses. Items that were not administered in the second round of CAT simulations were trimmed. Information was compared to the original P&P ASVAB, current operational Forms 5-9, and observed theta density. The items available for form assembly were expanded using unused or unassigned items from the Form 5-9 assembly and additional item series (89800 and 89900). Dr. Trippe then displayed a graphic showing MK simulation results. He indicated that MK information is not well aligned with existing operational forms or observed applicant ability. Including previously unassigned items mitigates this issue in the middle range of ability. Including items from additional series provides more information where needed (low to middle theta). For all other tests, information alignment is comparable to existing forms. Including additional series will help maintain information with the new goal of five new forms. They psychometric team must coordinate with the item development teams regarding information alignment with the observed ability distribution.
>
> Dr. Trippe concluded by summarizing the remaining project tasks. For the 89800 and 89900 series technical tests seed items, tasks include data cleaning, calibration, rescaling, initial screening, and enemy identification. For preliminary form assembly, additional item enemies in tests other than MC and MK need to be identified and the score information functions evaluated. Following final form assembly, equating and equating analysis and evaluation will need to be completed. Dr. Trippe noted that the team recently completed equating on Form 10 and is on top of the learning curve for equating Forms 11-15.

As Dr. Trippe recapped the completed steps of Form 11-15 development (slide 4), a committee member asked if each of the ten series had 100 items. Dr. Trippe said, yes, that there were 1,000 experimental items developed, but he clarified that not all of them survived. The committee member then asked if a series represented a set of items studied in field test conditions, and if each series, or set, was studied independently. Dr. Trippe replied that some series were studied together; that is, a seed version typically contains two series for technical tests and up to eight series for the WK test. The committee member said it sounded like seed versions and series were not interchangeable – that seed versions were composed of multiple series. S/he then asked if the numbers shown in the column headers on slide 4 were series numbers, and Dr. Pommerich replied that they were. Another committee member asked why the last two tests – the AO Connections and Puzzles tests – did not have as many series. Dr. Trippe replied that his team was not working on the AO test, and Dr. Pommerich said the rows should probably be removed from the table.

Regarding the results of enemy item identification (slide 9), a committee member asked Dr. Trippe to describe what “good prediction” meant. Dr. Trippe replied that it meant an item could be narrowed down to one or two groups (e.g., p3 and p5) in the table. The committee member then asked what it meant if an item landed in the bottom row of the table. Dr. Trippe said it meant the probability that the item belonged to the group was 1.00. To clarify how that information helped identify enemy items, Dr. Trippe said each item assigned to a group is an enemy of the other items assigned to the group. He said, thus, that items in the same group would need to be assigned to different forms. Another committee member said this likely presented implications for content coverage and asked if the process had been employed previously. Dr. Pommerich replied that DPAC had used this process for Forms 5-9, which she said was when they developed the model being applied today.

As Dr. Trippe explained the use of the heat map shown on slide 10, a committee member clarified that, even though the analysis was very mechanical, human judgement was still required. Dr. Trippe agreed and explained that his team did not have enough data to make predictions at a sufficiently fine-grain level. He said, however, that even if they had enough data, he might not trust it enough yet. The committee member stated that the process appeared to accelerate the process, and Dr. Trippe agreed, saying that it helped his team know where to target their resources.

Next, a committee member expressed concern over using the terms, “form” and “pool,” interchangeably. S/he suggested that one large pool or, alternatively, several smaller pools, could support the requirement of ensuring an item appeared only once on a given administration. Dr. Trippe replied that the five forms, if viewed as being stacked on top of each other, would essentially be one large pool, or five smaller pools if considered separately. Dr. Segall then said the term “pool” derived from the concept of operations that governed the construction of P&P forms, which in part were used to enable implementation of a retest policy. The committee member said that could also be achieved by having different pools or recording which items were used and blocking those items from being used again. Dr. Segall said that approach gets complicated and explained that the algorithm they currently use is “fairly greedy” to ensure that an examinee gets a parallel pool the second time s/he takes a test. The committee member then mentioned that the Graduate Record Exam (GRE) had employed the methodology s/he was recommending, but for security purposes, and that it worked well. Another committee member replied, however, that the model DPAC was using appeared to be pretty functional. That committee member said s/he had seen recent work on automated form assembly from pools, creating tests on the fly, but that s/he did not know whether that method was more efficient. Another committee member commented on the danger of having one large pool that may become compromised. S/he said DPAC’s approach better mitigates that situation.

A committee member asked Dr. Trippe to return to slide 4, which showed the complete list of item series, and to explain the inclusion of the last two series shown in red. Dr. Trippe said those series were added upon the realization that they needed more items in the preliminary simulation analysis stage. He added that development of those series lagged behind the others because they were added in late. He said they were trying to bring all the series to the same stage, that is, ready to be fed into the simulation. Dr. Segall took responsibility for the fact that the last two series were added late, saying that he decided to add them to develop an additional form to examine how a previously retired pool and a new pool intended for use with the Pending Internet-delivered Computer Adaptive Test (P*i*CAT) would perform. The committee member then asked about the status of the two pools, and Dr. Trippe said they were still evaluating them psychometrically. He also explained, however, that the AFQT and GS items were closer to being ready than the items from other tests. The committee member questioned how the forms could be used if they were not fully developed. Dr. Trippe reiterated that the AFQT and GS components were farther along than the others and said the preliminary simulation results he would show next would provide a good idea of what the information would look like. He also said there is a final process in which everything would be buttoned down. Referring to slide 14, he explained that item parameters already existed for AFQT and GS, which allowed those items to be loaded. He said the simulation results do not account for what is unknown, and so some of the items would be lost.

As Dr. Trippe presented slide 15, a committee member sought clarification on whether each line represented a form, as opposed to the pool. Dr. Trippe explained that the lines represented the score information, not the test information, and, thus, a "form" as it is administered to a simulated applicant. He went on to say that the information is at the candidate level, not at the pool level. The committee member agreed it referred to a form, not the pool, but asked if Dr. Trippe wanted the information curve to peak at the theta density. Dr. Trippe said he did, but it was not the case. He said the MK test was a bit unusual in the sense that the information is not that well aligned with the existing operational forms. He said it shifted to the right but that including additional series helped and this was as good as they could get. He said they have a lot of information at the top end and would like more at the middle and bottom, but that they cannot do anything about it right now. He added that they would coordinate with the item development team about it, because that team needs this information, in conjunction with the item-level feedback and difficulty estimates, so they can focus the target where people are. He said the current outcome is currently higher than where it needs to be.

Another committee member asked if the problem might be a function of the theta density being obtained from the average of the information functions, such that theta does not apply to any specific test. Dr. Trippe said, no, that these are simulated applicants and are MK-specific. The committee member then commented that there is more information at the higher theta, but it is relatively uniform from the peak of the theta. S/he also pointed out that, across the theta scale, these were providing more information than the P&P series. Another committee member asked if the stopping rule was the number of items, and Dr. Trippe said it was. He also said they could make these more alike, but it is not a problem to have more information. Dr. Segall said they should not get rid of it. The committee member asked if more items are administered at higher ability levels. Dr. Pommerich replied that there are more highly discriminating hard items than highly discriminating moderate items, which makes sense for math. She also said, however, that the effect is more pronounced than it has been in the past. Dr. Segall said it is an open question as to whether it can be fixed through better targeting by item writers. He said it could be that this construct is just difficult, relatively speaking, for this population. Dr. Pommerich recalled that HumRRO had tried to target moderate and easy items when they developed the Cyber Test, but that they were not very successful. Dr. Trippe added that the outcome for the Cyber Test was even more pronounced.

When Dr. Trippe mentioned the learning that occurred from equating Form 10 (slide 19), a committee member asked if the forms currently being examined predated the auto-generation of items. Dr. Trippe said there was some overlap. The committee member next asked about the effect of having automated generation fully in place. Dr. Trippe said it would accelerate the process. The committee member also asked if sibling items would necessarily be enemies, and Dr. Pommerich said they would. She also said they did not know whether they would be able to use AIG to develop math items and that, for WK, it remained to be seen how Dr. Isaac Bejar's work would perform.

6. Mental Counters – Rapid Guessing Behavior (Tab I)

Dr. Ping Yin, HumRRO, presented the briefing.

Dr. Yin began by explaining that MCt is a test of working memory originally developed by the Navy and studied as part of the Enhanced Computer-Administered Test (ECAT) battery evaluation. It includes 32 items that are currently administered to Navy applicants on the CAT-ASVAB platform. MCt (a) measures a unique domain not represented on the ASVAB; (b) demonstrates evidence of incremental and predictive validity (short-term/working memory), classification efficiency, and excellent reliability; (c) shows no adverse impact for gender and only a small practice effect; and (d) is an excellent candidate for AIG. Dr. Yin then provided a brief demonstration of the MCt.

Dr. Yin continued by showing two graphs of the distribution of Version 2.0 and 3.0 scores, both of which showed moderately good distributions except for a floor effect, where nearly 9% of examinees had a score of 0 on the first version, and about 4.5% on the second. Another graph displayed results from administrations of version 3.0 from 2015-1018, each of which demonstrated a significant floor effect. Possible explanations for this result include examinees not understanding the instructions, a lack of motivation, the test being too difficult, fatigue or frustration, or some combination of these. The operational definition of "not trying" at the test level is an examinee spending less time over all items compared to those who "try." At the item level it is evidenced by an observable pattern of spending less time on items as item number increases (sequential order effect), and is independent of item design (i.e., the number of adjustments and delay).

The first question guiding the research is whether it is feasible to identify examinees who are "not trying" using response time distributions. A second question is whether it is feasible to identify examinees who are "not trying" using an index at the test level. The impact on the floor effect and the correlation between the ASVAB (subtests and AFQT) and MCt total scores are examined before and after excluding examinees identified as "not trying." A third question is whether it is feasible to identify examinees who are not trying by examining the item-level sequential effect to determine if there is an observable pattern of the sequential effect. Dr. Yin continued by reviewing some previous research aimed at identifying unmotivated examinees. Schnipke (1995) noted that the response time (RT) distribution for incorrect answers often had a sharp spike during the first few seconds, representing rapid guessing behavior, and the RT distribution for correct answer had a broader distribution with a smaller peak, representing solution behavior. Lee & Jia (2014) noted that for multiple-choice items, the conditional p-value associated with rapid guessing is expected to be near the chance level. The response-time effort (RTE) index was developed to identify examinees who are engaged in solution behaviors (SB) over all items. Wise and Ma (2012) defined the threshold for an item as a percentage of the average response time (e.g., 10%). The RTE index is obtained by aggregating SB values over all items.

For the present research, the analyses focused on RT and RT/accuracy distributions. A test-level aggregated index was used to examine the floor effect after removal. The item-level sequential/order effect was examined by analyzing item RT (ordered by item number) for two groups to determine if there is an observable pattern; those who are at the floor and those who are above the floor. Item difficulty for those with a raw score of 1 and those with a raw score greater than 1 was examined to determine how likely it is to answer an item correctly by guessing.

The analysis of RT distributions for all items suggested that for most items the 75th percentile of RT is less than 10 seconds, which indicates most examinees spend less than 10 seconds on most items. This finding is consistent with the definition of working memory. Dr. Yin then showed a graph plotting RT and accuracy. This showed that the RT distribution is not bimodal; it is positively skewed with only one mode. For both example items shown, the RT peaks at about 5 seconds. The conditional p-value also peaks at about 5 seconds, meaning the accuracy is highest when RT is about 5 seconds. The conditional p-value declines after the first peak, which means spending more time does not lead to more accurate answers.

The analyses of the RTE index is obtained by aggregating SB values over all items. The mean and median threshold definitions are set at 10% through 90%, yielding 9 variations. Because the MCt is a working memory test, RT is very short, and an examinee can spend less time on an item but still get it correct. Spending more time does not always lead to increased accuracy. Therefore, a modified RTE was used in which SB = 0 if the RT is less than the threshold and the response is incorrect. There are a total of 9+9 or 18 variations of the modified RTE. Results are presented for a threshold of 80% of median RT with RTE greater $\geq$ 0.85. For raw score (RS = 0 and RS > 0), the value of RTE can be greater or less than 0.85. An examinee can be engaged in SB and still have a total RS of 0. An examinee can also randomly guess and receive an RS > 0. The RTE distributions are somewhat similar for MCt RS = 0 and RS > 0, except for the tails where there are more examines with RS > 0 when RTE is $\geq$ 0.85.

Dr. Yin continued by showing a chart of the MCt score distributions after the removal of the floor effect, which includes 96% of the original data. The MCt total raw score distribution looks relatively normal. The floor effect for RS = 0 is significantly reduced. However, RS = 1 seems to stand out, suggesting a secondary floor. The percent of RS = 1 is the second highest in the original data, with 2.97% of examinees answering only one question correctly. The modified RTE method reduced the number of examines with RS = 0. The aggregated RT index has some limitations.

Dr. Yin then showed a table displaying the correlations between MCt raw scores and ASVAB scores based on historical data and before and after removing the examinees identified as “not trying.” Correlation coefficients from this analysis are comparable to historical values. MCt RS correlates moderately with AFQT (around 0.5), and slightly lower with ASVAB subtests. The highest correlation is with AR, and the lowest with Auto and Shop Information (AS). The correlations are reduced slightly after removal based on RTE. Dr. Yin then showed a series of charts displaying results of the analyses.

Dr. Yin then turned to item-level RT distributions, noting that they should be mostly random given the way the items are designed. There is a clear trend for RS = 0, 1, 2, and probably 3. As the item number increases, the mean, median, and interquartile range (IQR) decreases. The IQR is the difference between the upper (Q3) and lower (Q1) quartiles and describes the middle 50% of values when ordered from highest to lowest. For most RT conditional on RS, the RT statistics (mean, median, IQR) are between 2 and 10. For RS = 31 and 32, the RT statistics (mean, median, IQR) are between 4 and 12. Examinees with higher MCt scores tend to spend slightly more time (2 seconds) responding.

The item p-value for examinees with total RS = 1 is very low (< .01) compared to p-values for all examinees. There is no obvious pattern for extremely easy items for RS = 1. For examinees with RS = 1, items 16, 21, and 22 are the easiest, with p-values around .01. Based on display times and adjustments, these are moderately easy items. Dr. Yin then showed a table of the top five response patterns to these items. Based on the key and response pattern, it is unlikely examinees can answer these items correctly by chance or guessing alone. However, it is possible that examinees guess on one or two of the three numbers required, which could increase the probability of answering correctly. As a next step, response patterns will be examined for signs of guessing (e.g., the examinee uses the same response patterns for all items.)

To summarize, Dr. Yin indicated that it is not feasible to identify examinees who are “not trying” on the MCt using RT distributions. This is due to the fact that the RT distribution for MCt is highly skewed with only one mode and is very short (usually less than 10 seconds). It is partially feasible to identify examinees

> who are “not trying” using the aggregated index (RTE), however it needs to be modified. Without adding the requirement of incorrect responses for RTE, more than 50% of examinees will be identified as practicing random guessing. After removing those identified as random guessing, the floor effect is reduced, however there is a secondary floor effect for RS = 1. It is feasible to identify examinees who are not trying by examining the item-level sequential effect, but it is difficult to implement at the individual level. An index to quantify the trend would be helpful. There is a clear trend for RS = 0, 1, 2, and probably 3 where an item number increases the mean/median/IQR decreases. This is not caused by a lack of time. Regarding item difficulty for RS = 1, it is very difficult to answer one MCt item correctly by guessing alone, but it is also possible that examinees guess on one or two of the numbers, which could increase the likelihood.
>
> Dr. Yin concluded by summarizing avenues of future research. These include evaluating response patterns, continuing to evaluate the sequential effect displayed in item-level RT, conducting additional research based on aggregated RTE, and evaluating whether instruction plays a role in the floor effect.

As Dr. Yin explained the definition of “not trying” on a test (slide 10), a committee member said two factors appeared to influence item difficulty: speed and the number of adjustments. Dr. Yin said variations in both areas are distributed randomly throughout the test such that difficulty does not increase systematically over the duration of the test. Dr. Manley commented that the order of items presented is constant across administrations.

As Dr. Yin presented slide 13, Dr. Velgach asked if the measurement of RT begins when the item first appears or after it has been presented. Dr. Pommerich replied that RT starts after an item has been displayed. A committee member asked if the graph on the slide showed RT, and Dr. Yin said it did, but she clarified that the example was only hypothetical.

After Dr. Yin presented the results of the first analysis (slide 22), a committee member asked if it was worth asking any questions about the first analysis, given that it did not work very well. A brief discussion ensued about the lack of a clear bi-modal distribution, after which, the committee member reiterated that the analysis was not helpful.

On slide 24, which referred to the second analysis, a committee member asked if the RTE index measured the percentage of items on which an examinee engaged in solution behaviors. Dr. Yin said it did and noted that higher numbers were better.

After Dr. Yin presented slide 26, a committee member asked if Dr. Yin was suggesting that a method that produced a normal distribution would be the ideal solution. Dr. Yin said that was an interesting question, however, the goal was to explain why so many scored at 0. Another committee member asked if those with a raw score of 1, like those with a raw score of 0, might have failed to understand the instructions. Dr. Yin replied affirmatively. She also explained that the definition of “not trying” based on the modified RTE was based on RT as well as accuracy, but that they only applied accuracy when the raw score was 0. She said the goal was not to create a normal distribution, but to figure out how to reduce the floor effect. The committee member said, yes, to identify people who were not trying. S/he then asked Dr. Yin to clarify the criteria for selecting the second analysis method. Dr. Pommerich responded that this work was just part of the process to answer that question and stated that it cannot really be simulated. Dr. Yin agreed.

Responding to a committee member's question related to RT results (slides 30-31), Dr. Yin explained that there was no limit on the time allowed to answer any single item, but that there was a 30-minute time limit for the test. She said, however, that people were not running out of time. The committee member noted that, after an item is presented, there is nothing to review. When Dr. Yin reported that the total RT for persons who scored 0 was three minutes (slide 32), a committee member suggested that they probably just wanted to get through the test, perhaps because they did not understand the instruction.

On future research plans (slide 37), a committee member asked if there was evidence that persons who scored 0 had not tried on the other tests as well. Dr. Pommerich said no, but that it might not be relevant if examinees knew their scores on MCt would not impact their eligibility. Dr. Pommerich also said she would be less concerned about the floor effect if she could convince herself that examinees were not trying. The committee member responded by clarifying that s/he thought the examinees were, indeed, not trying, probably because they failed to understand the instructions, which caused them to eventually give up. Another committee member asked if examinees get their scores on practice items and noted that the committee had talked previously about adding feedback of that sort. Dr. Pommerich replied that DPAC has not yet made the software changes that would be required to provide feedback. The first committee member then said s/he did not think anyone should proceed with the test until they get at least one or two practice items correct. S/he reiterated that the instructions were not simple. Dr. Yin replied that DPAC would like to ensure those who take the test get at least one practice item correct, but that the requisite changes to the software are a lower priority than other IT efforts currently underway. Dr. Pommerich confirmed that the requirement is ranked "pretty low."

A committee member commented on the p-value figure shown on slide 33, saying that it was clear that the green dots (indicating a raw score of 1) showed little performance difference sequentially, but that some of those items had different demand characteristics, for example, some were more rapid. S/he suggested that, rather than showing these data in sequential order, they could be shown in order of item difficulty, which might facilitate interpretation. The committee member added, however, that s/he doubted that would show anything different. Another committee member commented that total RT was only three minutes, or six seconds per item. Dr. Yin explained that examinees must type in three numbers to respond to an item. The first committee member asked if they had to answer an item to move on, and Dr. Pommerich said they did, which made the patterns shown on slide 29 so interesting: they become very proficient in answering quickly. The committee member said, if a person always entered the same responses, then that would be another indicator that they were not trying. Dr. Manley commented that a response strategy might explain the scores of 1, and the committee member said it would be like answering "A" on every question on a multiple-choice test.

7. <u>Mental Counters Think-aloud Plan</u> (Tab J)

Dr. Ping Yin, HumRRO, presented the briefing.

> Dr. Yin began by summarizing the factors that could be causing the floor effect in the MCt. These include examinees not understanding the instructions, a lack of motivation, the test being too difficult, fatigue or frustration, and combinations of some or all of these. Results presented in the previous presentation suggest that some examinees were not trying when taking the test. Others may be trying, but still did poorly,

scoring at the floor or near the floor. To determine if misunderstanding the instructions is playing a role, the idea of conducting a think-aloud study was presented to the DACMPT and MAPWG meetings in 2018. The DACMPT agreed that this was an idea worth pursuing.

Think-aloud is a research method that systematically collects validity evidence of response processes. In a typical study, participants speak aloud any thoughts in their mind as they complete a task. The method is widely used in usability testing, education, and related fields to determine how people approach tasks and to identify common misperceptions. In this instance, the think-aloud methodology will be used to determine if the MCt instructions are clear, easy to understand, and user friendly. The results can be used to determine if it is possible to simplify or streamline the instructions and identify areas in those instructions that could be contributing to the floor effect (e.g., misunderstanding, lack of motivation, overly difficult). Dr. Yin indicated she would outline the ideal study, a study designed within the constraints that are likely to exist, and a hybrid approach that falls somewhere between the best- and worst-case scenarios.

The ideal think-aloud study would involve randomly equivalent groups of participants, with each group receiving one of two types of MCt instructions (i.e., current or updated). The subjects would be representative of the applicant population regarding major demographics (e.g., age, gender, race/ethnicity). The rule of thumb for boundaries between small and large sample sizes is 25 to 30. A minimum sample size of 25 would be required for each group. Each participant would be scheduled for an individual session in a quiet setting that would allow for talking and audio taping. Dr. Yin provided the committee with a draft script and questionnaires that could be used in the study. She also showed a table that listed the steps to be followed in conducting the research and the estimated time for each, which totaled between 1 and 1 ½ hours. Another step would be to have an additional two random groups of participants take the actual MCt, but with different instructions. This would require four randomly equivalent groups to avoid possible contamination of influence of the think-aloud participants. The examinees in these additional groups would be observed during the session for evidence of rushing, low motivation, or fatigue. This would make it possible to ascertain the impact of factors such as motivation and determine if the updated instructions are effective.

Turning to the study that may be more readily accomplished given the constraints which exist, Dr. Yin indicated that it would be a single-group design with each participant taking the MCt using the current and updated instructions. This is problematic because of an order effect that may occur depending on which instructions are given first and a potential sequence effect where either set of instructions will be affected by the previous set. Counter-balancing the order of instructions is one way to deal with these possibilities, but the fact that the both apply to the same test could still introduce a confound. This would negatively impact the interpretation of the results and is not recommended. Additional constraints may require that the subjects of the study be a convenience sample rather than representative of the test-taking population. Given differences in demographics and motivation, these results could not be generalized to military applicants as a whole. Further, if the sample size is restricted, the statistical power in conducting analyses will be limited. If it is necessary to go with a small sample, focusing on the updated instructions alone might be advisable, and this would constitute a qualitative study. Dr. Yin then mentioned several other constraints, including (a) the current MCt instructions are only provided on a DoD desktop at the MEPS or CAT Lab, which are not necessarily quiet environments; (b) the practice items for the test are only on a DoD desktop and not integrated with the updated instructions; (c) the updated sequence of looping back to the demonstration after failing both easy practice items has not been implemented; and (d) the update of incorporating easier practice items with the instructions has not been implemented. Dr. Yin then presented a flowchart that showed the various options and their outcomes.

Insight into the floor effect has been gained through the analysis of the item-level response times, and efforts to identify examinees who are not trying will continue, but a think-aloud study should also be carried out. One next step will be to develop an item-level index to quantify the observed pattern in item response time for those who were not trying and evaluate the floor effect after implementing the index. It seems more feasible to conduct a small-scale, in-house think-aloud pilot study using a convenience sample, with a focus on collecting qualitative data. Despite its limitations, this could still provide useful information on examinees' understanding of the MCt instructions. After evaluating the results of this effort, additional consideration can be given to conducting a more formal study using applicants and recruits.

As Dr. Yin briefed what she referred to as “the ideal study” (slide 7), Dr. Pommerich responded to a committee member’s question about who would participate in the study, saying that DPAC could consider using Navy applicants, if they were available. A committee member then questioned whether the reference cited on slide 7, in respect to sample sizes for think-aloud studies, was informative. S/he asked if DPAC thought the new instructions were better and said, if so, why waste time with the old instructions. Dr. Yin replied that the ideal study would look at both conditions (i.e., old and new instructions). The committee member responded by comparing that approach to creating a set of worse instructions from a set of good instructions and testing those as well. Another committee member said s/he thought everyone agreed that the current (i.e., old) instructions were not adequate. The first committee member then said the goal should be to have a representative sample, but that the study was not “experimental.” The second committee member emphasized that think-aloud studies, as compared to experimental designs, do not necessarily require random equivalent groups. Continued discussion of the study, however, clarified that the ideal think-aloud study would also include groups who would take the test without performing the think-aloud task. On this point, a committee member requested and received assurance that the think-aloud portion would not constitute a treatment, which would likely improve scores. Dr. Yin stressed that the think-aloud task would only be used to obtain a better understanding and would do so apart from the experimental component of the study. The committee member then commented that the think-aloud component should be separated from the experimental component due to their different purposes.

A committee member asked if DPAC was planning to conduct the study in an operational setting. Dr. Yin said they were, if it is feasible. Otherwise, she said, a convenience sample would be used. The committee member then commented that the study really had two purposes: to gain clarification via think-aloud techniques and to determine the impact of improved instructions via experimentation. S/he said DPAC likely had a better sense of whether they needed to test to see how much the instructions had been improved.

Another committee member suggested DPAC improve the instructions before allowing anyone to take the test. S/he said this would likely yield a normal distribution. Dr. Pommerich replied that they want to do the study before the timeframe in which the software changes could be implemented. The committee member suggested providing the instructions separately, perhaps via P&P.

Continuing discussion of the ideal study, Dr. Pommerich said DPAC would use a convenience sample instead of recruits at MEPS, and so they would have to operate in a constrained environment. A committee member replied that if 9% are currently receiving scores of 0, it holds that the convenience sample would also yield a similar percentage if improved instructions have no impact. Another committee member commented that, if the think-aloud task is performed only by those who score 0, then it would take a large sample size to get enough people who will be useful (i.e., who will struggle) in the think-aloud activity.

Dr. Manley commented that the animated portions of the new instructions would be difficult to implement in a P&P format. He said, though, that they had thought of using a separate computer adjacent to the testing terminal to provide the new instructions. A committee member summarized the situation as this: DPAC needs to have some empirical evidence that the improved instructions make a difference before recommending an investment in their implementation. Dr. Pommerich clarified that the highest IT priorities are moving from WinCAT to iCAT and then to the Cloud and said she cannot ask programmers to work on the MCt at this time. Dr. Segall responded that they do not want to make software changes now and asked if the committee had any other recommendations. In response, a committee member asked if a convenience sample could be used to evaluate the situation by having participants go through the current instructions, but with pauses to ask them what they think they are supposed to do. S/he suggested a script or protocol could include questions such as, "what if I told you this or that, what would you think you were supposed to do then?" S/he also said a think-aloud study could be performed with the current instructions. Dr. Segall asked if it might be possible to talk with MEPS applicants who were struggling. Dr. Pommerich, responding to the committee member, said the logistics of pausing the program would present an issue, but that the product Dr. Manley prepared would allow pausing. She said the committee member's idea was good in theory but presented difficulties in application.

The committee member reasserted that the committee believed that people do not understand what they are doing, but s/he said that does not necessarily explain why they give up. Dr. Pommerich said DPAC was concerned about the instructions, but they also think motivation is a larger issue that they previously believed. A committee member replied that motivation would be minimal when examinees do not understand the instructions. Another committee member replied, however, that if examinees think they understand, but do not, it would not explain the lack of motivation. Another committee member asked what applicants at MEPS are told about the test. Dr. Pommerich said they are *not* told that it does not count. Dr. Watson said that was his understanding as well. Dr. Pommerich asked Dr. Watson if the recruiters tell applicants that the test does not count, and Dr. Watson said they are not supposed to do that. Dr. Heffner said not to count on that. The committee member replied that s/he thought that the applicants were probably motivated. Dr. Watson said he does not know what happens with the recruiters, but Dr. Heffner said they may say something like, "when you hit this part of the test, just don't worry about it." CPT Alex Ryan (USMC) confirmed Dr. Heffner's assumption.

Dr. Watson then asked why DPAC had to measure understanding of the instructions at MEPS and if it could be done at one of his labs. He also suggested correlating MCt scores with some measure of *g*, which he said was a technique he found useful in evaluating the difficulty of other tests. Dr. Pommerich said DPAC could support conducting the study at one of Dr. Watson's labs. Dr. Watson also said he could find a lower cost method of getting the instructions on a computer. Dr. Pommerich then asked Dr. Yin if, given that there was a place to conduct the full study, there would be value in briefing the constraints (i.e., the remainder of the briefing slides). Dr. Yin then proceeded to slide 13, which mapped out the think-aloud study design possibilities. Discussants concluded, however, that it would be advantageous to proceed with the study with Dr. Watson's assistance, and Dr. Watson agreed to support the effort.

The discussion concluded with a committee member asserting that the administration mode was less important than the content of the instructions for the task. Dr. Pommerich replied that the committee member had said earlier that s/he thought the applicants understood the task. Another committee member proposed, however, that if they received feedback on their responses to practice items, they would know whether they understood. A third committee member recommended not letting them proceed with the test until they understand. Dr. Velgach asked if examinees are informed when they get practice items wrong, and Dr. Manley replied that they are only shown the correct response. Dr. Pommerich said examinees can get through 9 practice items, answering each incorrectly, and still progress with the test. She said, however, if they miss the first four, then the TA is alerted to help the examinee, and then they get five more practice items. Dr. Velgach said having a TA come over after answering four questions wrong should be a good indication that they did not understand. Dr. Pommerich explained that they do not track that activity, so records of it are not available. A committee member proposed that it would be helpful if the practice results were recorded. Dr. Pommerich said that would require a software change, however Dr. Segall said it could be done. Another committee member reiterated the need to find participants that would be productive in the study (i.e., people who would not understand the task).

8. <u>CAT-Cyber Test</u> (Tab K)

Dr. Furong Gao, HumRRO, presented the briefing.

Dr. Gao began by explaining that the purpose of this work was to evaluate the feasibility of administering the Cyber Test in a CAT framework. Since 2011, two operational 29-item static forms have been administered via computer. In 2016, CAT pools were constructed using automated test assembly, with 166 items selected from 58 existing items from operational forms and 190 newly-tried-out calibrated items. A two, roughly parallel, form/pool solution was evaluated by simulations examining score information functions (SIF), item usage, and test-retest reliability. This resulted in the decision to develop more items that target the low and middle range of the ability distribution. In 2018, new two-pool and three-pool solutions were evaluated, with items selected from 166 existing items and 242 newly calibrated items. The two pools each contain 130 unique items, while the three pools contain 87 unique items each. Item enemies exist across, but not within pools.

Dr. Gao then presented a slide summarizing the recommendations from the 2016 evaluation. These included (a) maximizing test security by using CAT administration with a maximum exposure rate of 0.40, (b) maximizing reliability by using longer test lengths than the 10 to 15 item administered in the CAT-ASVAB subtests, (c) using a 20-item test length, (d) staying with the two-form solution to ensure forms are parallel, and (e) developing more discriminating items at the low and moderate difficulty level.

The dimensionality assessment was conducted using IRT model-based factor analysis. The assumption is that the test is designed to be uni-dimensional by measuring a single construct but with broad content coverage that may introduce additional, unintended dimensions to the test data. Items are rendered so that the "missingness" in the response data is missing completely at random (MCAR) or missing at random (MAR). Both the CAT-ASVAB and the currently seeded item design produce MCAR data. Confirmatory factor analysis (CFA) was conducted with the assumption that the data will be fit to both a one-factor and a bi-factor model. The latter will include one general factor or dimension with all items loading on *g,* with a group of secondary factors, one for each of the sub-domains, and the factors will be independent of one another. The secondary factors will align with the four broad content areas covered in the test—Computer Operation (CO), Networks and Telecommunications (NT), Security and Compliance (SC), and Software Programming and Web Development (SPWD). The *g*-factor loadings of the two models are compared, with small and neglible differences expected. There is a small role of specific/group factors, but they do not distort the meaning of the general factor that is measured by all items on the test. An indicator of essential uni-dimensionality is explained common variance (ECV). The value of ECV is between 0 and 1, and the larger the value, the stronger the unidimensionality.

To examine the dimensionality of the Cyber Test, item response data were analyzed using operational items, previously seeded items, and newly seeded items. There were a total of 417 items (CO = 142, NT =

127, SC = 100, SPWD = 48). The data were from 108,292 individuals who took Form 1 of the test and 112,221 who took Form 2. The case counts on previously seeded items ranged from 6,492 to 7,678. For newly seeded items the case counts ranged from 3,174 to 3,895. Dr. Gao then showed two charts displaying the item factor analysis results using 417 items. The ECV was 0.925 and increased to 0.938 when adjusted for the standard error of the estimates.

Dr. Gao then turned to analyses of Cyber Test CAT Pools. This involved two 29-item operational forms scaled in 2011 which served as the baseline. In addition, Seed 190, developed, scaled, and equated in 2015 and Seed 242 developed, scaled, and equated in 2018 were used. The notation used is consistent with that used in th 2016 evaluation: two fixed operational forms 02A and 03A; two form pools 01Z and 02Z; and three form pools 01Y, 02Y, and 03Y. Dr. Gao then showed a table displaying the content distributions across forms/pools. This was followed by a chart displaying item parameter distributions which indicated that the *a* parameters in the CAT pool forms are generally lower than those in the two 29-item operational forms. An additional chart suggested that the *b* parameters in the CAT pool forms are generally higher than those in the two 29-item operational forms. Except for one of the forms in the three-form CAT pool solution, all the difficutly distributions of the CAT pools show more spread and have larger interquartile ranges (IQR) than the operational forms. Results displayed in another chart indicated that the mean values of the *c* parameters in the CAT pools are similar to those in the operational forms. However, the median values are much closer to the means in the CAT pools than in the operational forms, indicating less skewness of the distributions.

The evaluation approach involved simulations using the item pools under CAT-ASVAB administration conditions. Test lenghts of 10, 15, 25, and 30 items were used, and no content constraints were applied. This was supported by the dimensionality assesment findings. A target maximum exposure rate of 0.67 was used, which matches the current rate for all subtests on CAT-ASVAB Forms 5-9. Score precision and item usage were evaluated. CAT pool precision was evaluated using score information functions (SIF), which were calculated using simulated data with 500 examinees at each of the 31 equally spaced theta values in [-3,3]. At each theta, the mean and variance of the 500 scores were calculated, and the SIF was approximated using these results. Dr. Gao then showed two charts displaying results suggesting there was higher score precision with the two-form solution and higher score precision than the pools used in the 2016 evaluation. Another chart showed the averaged SIF comparisons, where SIF was averaged across the simulated tests within the two-form or three-form pool solutions. Test scores from the three-pool solution would have lower precision. Another chart indicated that item usage was better across the three-form solution than the two-form solution. In the two-form solution, even with a 30-item test, about 30% of the items were not used.

Results of simulated test-retest reliability were displayed in a table. These were slightly higher than what was reported in the 2016 evaluation, with the average reliability from the two 29-item operational forms being .78. The average simulated test-retest reliabilities of the CAT-ASVAB tests across Forms 5-9 are generally higher, and in the high .80s.

Dr. Gao concluded by stating that, with the additional 242 items, the constructed CAT pools showed higher test score precision and test-retest reliability than previously evaluated pools. Many low-discriminating items in the pool were not used in the simulated tests, and these items would likely be dropped from the pools. This led to the recommendation that the 15-item test length be used for CAT administration, given that score precision is higher than the two operational forms in most ability ranges. The three-form solution should be used because it yields higher score precision than the two operational forms in most of the ability ranges. Two forms should be used for operational CAT administration to replace the two static 29-item forms. The third form should be reserved as a reference form for future item scaling and equating. In addition, two additional CAT pools should be developed targeting more discriminating items in the moderate-to-difficult range. Dr. Gao noted that at the February 20 MAPWG meeting, the Service representatives voted unanimously in favor of a CAT test transition in the future with the recommended 15-item test length and three-form solution.

As Dr. Gao discussed missing items in the response data (slide 5), a committee member asked how, in a CAT environment, there would be missing data, unless the missing data were for the pool as opposed to the administration. Dr. Pommerich replied that the missing data were the items in the pool that were not taken; she said, if a person took 15 items from a pool of 100 items, 85 items would be labeled as missing. Another committee member asked if DPAC imputes missing data, and Dr. Pommerich said they do not. A third committee member asked if the missing data were a function of the design. Dr. Trippe explained that these were seed items, and the committee member said that answered the question.

As Dr. Gao introduced the unidimensionality assessment (slide 7), a committee member asked if the criteria for determining unidimensionality were different for the adjusted index. Dr. Gao responded that the same rules were applied. She also said that, given that the large sample sizes, the adjusted and un-adjusted index values were very close.

In response to a committee member's question about the IFACT results (slide 9), Dr. Gao confirmed that the correlations were on *g*-factor loadings. Another committee member then asked which items were included in the Cyber Test CAT Pools shown on slide 10. Dr. Gao replied that there were 58 operational items, 190 previously seeded items, and 242 new items. She said 73 items were excluded. The committee member then noted that most of the items were seed items.

As Dr. Gao described the evaluation approach, a committee member asked if the intent was to have no content constraints on administration. Dr. Segall said that was the intent, but that the purpose of the dimensionality assessment was to make that determination empirically. He said they would have considered balancing if there had been secondary factor evidence, but that was not the case. The committee member replied that content is one way to think about representation, but another committee member said, statistically, it is not required. Dr. Segall responded that content representation is more important in certification. Dr. Velgach then asked if DPAC was still going to balance items across content areas, and Dr. Segall said they were.

A committee member responded by clarifying that factor loadings were often correlated with item discrimination, which might lead to some content becoming irrelevant. Another committee member agreed that might happen. The first committee member then suggested that DPAC might attempt to cover content, but the CAT might not like it because it would include non-discriminating items. The second committee member recommended that DPAC look at the content to see that one area is not shifting out of line with the others. Dr. Segall said he had done that in the past when looking at different ability levels. The committee member replied that the problem bothered him/her less for difficulty, but when selecting items for administration, it is discrimination that counts. The second committee member added that, in simulations, it is possible to see a propensity for the administered form to be differentially weighted in terms of content; s/he suggested that DPAC could do that for different theta bounds: if theta is low, run the simulation to see what content is really administered, and then do that for low, medium, and high ability to see the tendency. The first committee member said his/her guess was that different content distributions in the items administered would be seen at different levels of ability. S/he reported being more concerned if that was due to some content areas being less discriminating. Dr. Segall said he would expect lower proficiency examinees to get more of the items in the easier content areas. A committee member asked Dr. Gao if she had looked at discrimination by content area, and Dr. Gao said she had not. The committee member said it would just be an interesting piece of information.

When Dr. Gao briefed the two graphs on slide 17, a committee member commented that the two graphs were not comparable because they used different numbers of items. Dr. Gao replied that was because the pool size was different, but she said both score information functions were on the same scale. Dr. Pommerich replied that they were based on the same number of items taken. Dr. Segall then said the score information was defined independent of the number of items taken; he said test length has an impact on the simulation, but then admitted that he might not understand it. Another committee member asked if one could say that the 30-item test from the two-pool solution, which had information functions around 15, and the 30-item test from the three-pool solution, which had information functions around 12, would be comparable. S/he asked if one provided more information than the other. The first committee member said, yes, because the number of items was controlled. S/he went on to say the difference must be a function of the richness of the discrimination of the items. A third committee member replied that, importantly, it measures very high levels of theta. The second committee member said, though, that it had shifted between pool solutions. Dr. Pommerich said they had tried to develop highly discriminating items of moderate difficulty, but that did not work out so well. A committee member replied that, if the Services are looking to discriminate at high levels, this test would do it. Another committee member responded that if the discriminating items are truly this difficult, this was a really hard test. Dr. Pommerich replied that the technical tests tend to follow this trend; that is, not a lot of people know the content.

A committee member then asked how cut scores are set. Dr. Velgach said cut scores for the test were different across the Services. Dr. Heffner said the Army uses a score of 60, and Dr. Manley added that HumRRO had performed analyses to optimize cut scores. Dr. Trippe, of HumRRO, reported that they had used a regression approach to identify those who would have above average performance in training. Dr. Heffner clarified that they want to set cut scores to identify people who will pass the training. A committee member said, once again, this suggested the necessity of using larger numbers of easy and moderately difficult items. Dr. Pommerich replied that they "were not going to try that again."

As Dr. Gao briefed slide 18, a committee member asked if, when scaling is completed, theta would be centered at the average person rather than at the average item. Dr. Segall said it would. The committee member then noted that the pass rate would be 30% or less. Another committee member said it would be less than that, maybe only 15%, one standard deviation below the mean, or 16%. S/he asked if that would support use of the test. Dr. Pommerich replied that, similarly, DLAB was used to place persons who could pass specialized training. Dr. Velgach said the purpose of the Cyber Test was exactly that: to identify people who could succeed in specialized training.

When Dr. Gao briefed item usage for the two-pool versus the three-pool solution (slide 20), a committee member commented that the difference was just a function of the number of items available. Another committee member said the proof was in the information functions.

On test-retest reliability results (slide 21), a committee member remarked that the result was the same as that for the information functions. Another committee member said, if the two-pool solution had the same number of items as the ASVAB, the reliabilities would be in the high 0.70's, though s/he said s/he was not recommending that. S/he then asked what test length DPAC was recommending. The first committee member interjected, suggesting that the ASVAB had better defined constructs. Dr. Pommerich replied that the ASVAB had smaller pools than the Cyber Test, because the items were originally distributed across five pools. She said 87 (the three-pool solution set) would have been the largest set. Dr. Segall said the difficulty of GS, AR, and PC items was more closely targeted and there were more highly discriminating items. He said, now, the ASVAB starts with around 200 items per pool, which is larger than the number of items in the two-pool and three-pool Cyber Test solutions, and that results in greater precision.

As Dr. Gao briefed the recommendation to use the three-pool solution (slide 22), a committee member referred to the graphs on slide 17 and said s/he thought the two-pool solution provided better information. Another committee member, referring to the recommendation to administer a 15-item test, said the decision to use the three-pool solution was surprising. There was then some discussion of the average SIF comparison on slide 18, but then Dr. Gao said the choice to use the three-pool solution was driven by the need to have two operational pools and one reserve pool. A committee member then asked whether more than one form could be drawn from each pool. Dr. Segall responded that people who take adaptive tests will have different items, but he reiterated that they wanted a non-overlapping pool to hold out. He also said there is a tradeoff between precision and the number of pools; he said, if they did not need to have a reserve pool, they would have gone with the two-pool solution to attain higher precision. Dr. Segall said, if they at some point need to equate a new pool to the operational pools, and the operational pools were compromised, then they would have a problem. A committee member replied that if DPAC has landed on the 20-item administration, the three-pool solution would clearly be fine. Dr. Segall replied that they wanted to stick with the 15-item administration, because all the other tests are 15 items. The committee agreed that the 15-item solution would be sufficient for the classification purpose, especially because future pools will specifically concentrate on optimizing information at the cut score level working to increase test-retest reliability.

The briefing concluded the first day of the meeting, and Dr. Velgach asked if there were any comments from the public. There were no comments from the public.

9. Adverse Impact (Tab L)

Dr. Greg Manley, DPAC presented the briefing.

> Dr. Manley began by explaining that adverse impact is the unintended discrimination of a protected class that is the result of a selection procedure. AI is not a property of a test, per se, but may occur when a test's scores are used as the basis for selection. A test may contribute to the occurrence of AI when it shows sizable mean score differences between a majority group and a protected class. Effect sizes of the standardized mean difference provide a method of evaluating potential for adverse impact across individual ASVAB and special tests, where no direct selection occurs. Dr. Manley then displayed the formula for computing effect sizes and for computing confidence intervals about effect sizes. He indicated that small effect sizes start at 0.20, moderate at 0.50, and large at 0.80. The ASVAB testing program evaluates comparisons for four reference group/focal group pairs: (a) males/females, (b) non-Hispanic Whites/Hispanic Whites, (c) non-Hispanic Whites/non-Hispanic Blacks, and (d) non-Hispanic Whites/non-Hispanic Asians. The special tests given on

> the ASVAB platform are (a) MCt, which is a test of working memory only administered by the Navy; (b) the Cyber Test, which assesses basic computer information systems knowledge and is used by all Services; and (c) Coding Speed (CS), a speeded test of assigning code numbers to words, also only administered by the Navy.
>
> Dr. Manley continued by showing a series of charts displaying the effect sizes and 95% confidence intervals for comparisons of the reference and focal groups for the special tests and ASVAB subtests. He concluded by stating that the special tests generally exhibited small to moderate effects that were usually as low or lower than most ASVAB tests. White-Black comparisons were generally larger for MCt than for the other group comparisons. CS usually had very small effects (near 0), but this test suffers from other issues. It is affected by lag time in Internet delivery given that it is a speeded test, it is known to be affected by test delivery device, and it suffers from coachability and susceptibility to invalid strategies that result in high scores. He stressed that the potential for AI is not the only consideration in making changes to the ASVAB.

As Dr. Manley described who is affected by adverse impact, a committee member asked about the sample sizes for the referent groups. Dr. Manley replied that the actual sample sizes were included in the backup slides and that they are all in the multiple thousands.

As Dr. Manley talked through the male-female effect sizes for the MCt, Cyber Test, and CS (slides 7-10), a committee member commented that MCt did not appear to be an outlier in comparison to the ASVAB subtests. Additionally, Dr. Velgach observed that the Cyber Test was on the low side compared to the ASVAB technical tests. Dr. Manley said the technical tests tend to produce that type of male-female effect size. Dr. Velgach then asked if the sample used to determine the CS effect size included only Navy personnel, and Dr. Manley said it did.

Regarding effect sizes for non-Hispanic Whites versus Hispanics for the same three tests (slides 15-18), a committee member asked if the Hispanics category included Black Hispanics. Dr. Manley replied that, if a person identifies as Hispanic, he/she falls in that group.

On effect sizes for non-Hispanic Whites versus non-Hispanic Blacks (slides 19-22), a committee member agreed with Dr. Manley's assertion that the MCt effect size was large in relation to those of the Cyber Test and CS. Dr. Manley said it was similar to the effect sizes found with many of the other ASVAB tests. The committee member next commented that the non-Hispanic White-Black effect was large for many of the ASVAB subtests. Dr. Manley replied that the results were also similar to those found on other standardized cognitive tests. Another committee member reflected that the committee had talked about that in the past, and s/he noted that a half standard deviation difference is persistent in early childhood. Dr. Manley suggested that might be partially due to cultural context, and when a committee member noted that the MCt does not contain cultural content, he explained that there is a significant reading requirement that brings reading comprehension into play.

A committee member asked for a recap of the CS test. Dr. Segall replied that it provides a lookup table that contains a dozen words and associated numbers. He said, for each question, the examinee gets a word and must place numbers in a table. He said this type of test originated in the 1940s when records were kept in file cabinets. He said it was a common task to code for record keeping, but that the task still has validity. A committee member asked if the instructions were heavily verbal, and Dr. Pommerich replied that the verbal requirement was not as extensive as that of the MCt. Dr. Segall said it was more of a step-by-step process. Dr. Pommerich said she did not recall if examinees had to get it right, but Mr. Tiegs said they did. A committee member then said this was different than the MCt test. Dr. Segall replied that there are two modes of test scores, one around chance and one around 90%. He said there is some reason to believe that people do not understand the task. Another committee member agreed that it was like the MCt in that sense.

As Dr. Manley presented the effect sizes for non-Hispanic Whites versus non-Hispanic Asians (slides 23-26), a short discussion clarified that the sample for the Cyber Test was all-Services minus Marines. A committee member then noted that the ASVAB effect sizes were a little larger for the Navy sample (i.e., CS and MCt analyses), and Dr. Manley said he did not know specifically why that was the case, except that there are differences in Service populations. The committee member then asked about the percentages of Asians from different regions (e.g., Japan, China, South-East Asia) and said persons from different regions have very different educational experiences. Dr. Pommerich said DPAC distinguishes among different categories of Asians. The committee member then reiterated that the difference was between the Navy sample and the sample for the Cyber Test, which was across-Service. S/he also said there was a movement among states to better desegregate data in recognition of more specific categories. S/he said if differences between categories is a concern, it would probably be among the Asian categories. Dr. Manley agreed that the current data was based on a very broad sample of Asian categories.

Regarding conclusions related to CS administration (slide 27), Dr. Segall provided more detail about how changes in response modes had led to its susceptibility to invalid response strategies. A committee member then asserted that the adverse impact analysis was incomplete in the sense that selection ratio had not been taken into account. S/he mentioned the high cutoff on the Cyber Test and suggested a more in-depth analysis was warranted. Dr. Manley agreed but clarified that the ASVAB subtest (and some special test) scores are used in the context of composite scores, which would make it difficult to examine the subtests in that manner. Additionally, he said the Services' composites have different cut scores for different occupational specialties, so each composite for each job's cut score would have to be examined individually by specialty. The committee member replied that it could be done with the Cyber Test, and Dr. Manley agreed.

Returning to sample sizes, which were shown on slide 29, Dr. Manley showed that the Asian sample was relatively small for MCt (i.e., N = 1547). A committee member asked about the timeframe for the sample, and Dr. Manley said they try to perform the analysis every two years. He said this report was for FY 2017 data. Mr. Aswell then commented on CS, explaining that it is no longer used by the Army, but that the Navy still uses it due to its relevance to predicting the ability to encrypt and send a message quickly. He said latency differences stemming from delivery would negatively impact the utility of test scores.

10. <u>Device Evaluation</u> (Tab M)

Dr. Tia Fechter, DPAC, presented the briefing.

> Dr. Fechter began by explaining that the goal of this study is to facilitate delivery device expansion of the ASVAB *i*CAT and P*i*CAT by evaluating examinee performance differences among electronic devices (e.g., tablets, smart phones). This will offer more flexibility for ASVAB administration to reduce time spent in the MEPS, increase the number of enlistees, and increase school participation in the CEP. The study will

allow DPAC to make a recommendation regarding which types of electronic devices should be approved or prohibited for ASVAB administration. It will also inform a Next Generation ASVAB user interface that incorporates a responsive design approach, which automatically formats the test display to alternative devices.

Dr. Fechter then cited a personal communication she had with Laurie Davis, Senior Director of Psychometrics at Curriculum Associates, who has done work on this topic. Dr. Davis has found performance across math, reading, and science high school exams to be similar between tablet and computer conditions. For reading, a small device effect favoring tablets was found for the middle to lower parts of the score distribution, with males tending to perform better using tablets. Response time was longer on tablets. It is important to allow for modifications of the test layout to best fit the device in question (i.e., responsive app design). Dr. Davis said that smart phones have not been tried but remain a possibility and should be included as a condition in the study. She is optimistic that results will show comparable performance across devices. Dr. Fechter also cited communications she had with Marine Corps personnel who have delivered the AFQT Prediction Test (APT) and P*i*CAT on 8" and 10" Samsung Tab Active tablets at the recruiting commands. They are currently not experiencing image display issues and are unaware of other issues. However, they don't know the impact on scores. Other findings from the literature include:

- No score differences were found between mobile and non-mobile devices for personality job selection assessments, although it did take examinees longer to complete the tests on mobile devices.
- Score differences have been found between mobile and non-mobile devices for cognitive job selection assessments, although measurement invariance held up for all tests administered.
- Minority groups tend to have access to the internet primarily through smartphones.
- A summary of research shows that it takes longer to take tests on mobile devices.
- Job applicants report more positive reactions to taking tests on mobile devices when the delivery application is specifically designed to support mobile administration.
- Distractions and disruptions are more likely when tests are taken on mobile devices.

Dr. Fechter continued by stating that the DACMPT can be helpful in identifying any barriers to implementing the current evaluation design, offering feedback on the preliminary pilot, and providing recommendations to strengthen or support the analysis plans. The primary questions to be answered through the evaluation research are whether (a) devices differentially impact examinee performance (score, response time) on ASVAB subtests, (b) device familiarity differentially impacts performance, (c) devices differentially impact item difficulty, and (d) item features such as inclusion of graphics interact with the device to increase the probability that item difficulty is differentially impacted.

Dr. Fechter then showed a table summarizing the evaluation sampling plan, which includes recruits at Army, Air Force, Marine Corps, and Navy bases and applicants who are processing into the military at 15 medium-volume MEPS. The goal is to have 9,340 subjects across testing venues. An additional table summarized the evaluation design, which calls for one control condition with subjects testing on notebook computers, and six experimental conditions involving various other devices, web browsers, operating systems, and screen sizes. Dr. Fechter stated that a pilot had been conducted at the San Jose, CA MEPS on March 18-19, involving 24 participants. There was an issue with the "next" button on the Chromebook, and initial issues with obtaining WiFi were resolved by moving the router closer to a door. The experience suggested that TAs may not be familiar with the variety of devices to the degree that they can adequately support examinees, so on-site training of TAs will be necessary. The MEPS Test Control Officers (TCOs) were eager and helpful.

Turning to the analysis plan, Dr. Fechter indicated that the question of differential impact of devices on performance will be examined by conducting MANOVAs after equating the two parallel forms across all device conditions. The dependent variables will be equated subtest scores and response times. The independent variable will be the device used. A total of seven MANOVAs will be conducted, one for each subtest. If the F-test is not significant, no further analysis is needed. If it is significant, post-hoc analyses will be run to determine where the differences are. To address the question of whether device familiarity

> has a differential impact on examinee performance, t-tests will be run between subtest scores pooling across device conditions comparing those familiar with the device used and those unfamiliar with the device. This will be repeated to examine response times, yielding 14 t-tests, one for each subtest and dependent variable. If the t-test is not significant, no further analyses will be required and the familiarity groups can be pooled. If the t-test is significant, consideration will be given to modifying the design to test device effect by adding a categorical covariate (e.g., MANCOVA). To answer the question regarding whether device type has an impact on item difficulty, multi-group IRT calibration will be conducted and item difficulty values compared, as is done with Differential Item Functioning (DIF) analyses. The groups will be the seven device conditions, and items showing DIF will be flagged. The flagged items will then be reviewed to identify any patterns in item features (e.g., includes a graphic) that may explain the differences detected.

When Dr. Fechter described the existing research update (slides 4-5), she mentioned to a committee member that she had contacted Laurie Davis and obtained her most recent research, which she said was informative regarding the use of tablets. Dr. Fechter also mentioned the large sample sizes collected in prior related research (i.e., around 3,500,000), but said the research on mobile devices was based on a smaller, though still adequate sample of around 69,000. A committee member commented on the large size of the samples.

Dr. Fechter relayed research findings that people who took a test on a mobile device had better impressions of the experience than those who had originally taken the test on computer, if, that is, the test was administered via an application designed for a mobile device. She then shared that her personal experience taking the ASVAB on a smartphone had been positive. A committee member asked if the ASVAB she took had been designed for the mobile device, and Dr. Fechter replied that it had.

As Dr. Fechter presented the evaluation questions (slide 7), a committee member asked if there were time limits on the subtests. Dr. Fechter said DPAC had increased the time limits to be more than sufficient. She said they were also asking participants not to focus on the clock, which she said is now set to 59 minutes, regardless of the condition. She said they wanted examinees to have enough time to respond and not to think about pacing; she said this would help them learn more about response-time differences among devices. The committee member then asked whether the response times in the study would translate to response times in the operational environment. Dr. Segall responded that, if they identified that a given device required more time, they would set the operational time constraints accordingly, which he said was the reason for removing time as a factor in the study.

Another committee member asked if DPAC would be able to look at the interaction between the first two evaluation questions (type of device and device familiarity). Dr. Fechter said they would, but that they wanted to start by looking at the main effects. She added that identifying the reasons for performance differences would be a very intricate task; that is, they might want to look at the pattern of differences, perhaps screen size, browser, or the combination of the two. She said their guess is that there will be no differences. The committee member then asserted that research typically looks at interaction effects first, to which Dr. Fechter replied that they wanted to test the device familiarity effect first with t-tests and, if there are differences, use familiarity as a covariate in examining the device effect. The committee member reiterated that if there is a significant interaction, interpreting main effects first would not be appropriate. Dr.

Segall added that the data required to look at interactions would be available, because subjects are assigned to groups randomly.

On the sampling plan (slide 9), CPT Ryan clarified that MCCSSS was the Marine Corps Combat Service Support Schools at Camp LeJune, NC. A committee member then asked if study participants would be Service members, and Dr. Fechter said, yes, and that they would have less than one year of service under their belts. The committee member replied that they would have already qualified, which would restrict the lower end of the population distribution. Dr. Segall then clarified that only half of the 14,000 participants would have already qualified, so the full distribution would be unrestricted accordingly.

A committee member asked how DPAC is convincing applicants to participate in the study. Dr. Fechter replied that MEPCOM had suggested applicants be briefed on the background of the study and asked to volunteer. If they volunteer, they are given stickers so researchers can identify them when they are in the waiting area. She reported that on the first day they tried this approach, and they initially had only four volunteers, but that they then asked the TCO to assist in "recruiting," which resulted in a total of eighteen volunteers for the day. Dr. Segall described situation as more like being "voluntold." Dr. Fechter said 18 volunteers exceeded their goal of ten per day. She said on the second day, they asked the TCO to be less active and relied on researcher solicitation only, which produced six volunteers. She said their plan for future administrations at MEPS is to introduce the study to applicants as they get off the bus and then have the TCO help, but only as needed.

A committee member asked when the study was administered in relation to ASVAB administration and if there might be an order effect. Dr. Fechter said they were not controlling for that but said they could get them either before or after the operational test. Dr. Heffner asked if DPAC would be tracking order, and Dr. Fechter said timestamps would provide that information.

A committee member pointed to the fact that participants in groups 9 and 10 would be taking a longer test and asked if those groups would be tested at the MEPS. Dr. Fechter said, no, that participants in those groups would be active Service members. She then clarified that participants in groups 1-8 would be tested at the MEPS and that, when accounting for all the subjects taking a subtest across groups, a balanced number of subjects would be taking each subtest.

While explaining the methods shown on slides 10-12, Dr. Fechter commented that there is a lot of variance in device characteristics, and this would make it difficult to be explicit in explaining the causes of performance differences.

A committee member asked if recruits ever spend more than one day at the MEPS. Mr. Aswell replied that they usually spend a little less than two days, on average, but explained that it takes two full days including travel time. He clarified that, typically, most of the testing is done in one day.

As Dr. Fechter described what they had to do to obtain sufficient Internet connectivity (slide 13), a committee member pointed out that P*i*CAT administration already requires connectivity, and that there should be existing connectivity specifications for using various devices. Dr. Segall said the military test – fortunately – will not require as much bandwidth as more regular activities, such as watching video. He also said people will automatically recognize whether they have the connectivity required to take the test, due to their familiarity with operating on-line. Dr. Pommerich then explained that there is no WiFi at the MEPS, which requires them to use the 4G network at those locations. She said they use “CraddlePoints,” which require a business contract with the carriers (e.g., AT&T, Verizon).

When Dr. Fechter described some of the issues they experienced at Fort Drum, NY, such as tests freezing mid-way through an administration, a committee member asked how many times that happened. Dr. Fechter said it happened to only a handful of people, and that it happened most often with the smartphones. Dr. Watson then asked if DPAC could resolve the software issues before the Pensacola, FL data collection. Dr. Fechter said they could, and that they have weekly meetings with IT staff to identify and solve problems. Dr. Watson said the availability of the “back” and “next” buttons is “super important,” and asked if fixing issues related to their on-screen visibility is being addressed. Dr. Fechter said this would be difficult to solve during the timeframe of the study, but that it would be solved before implementation. She clarified that only one person has needed assistance due to this glitch, so far. Dr. Watson said it would still be great if the problem could be fixed in the near-term, and Dr. Fechter agreed.

As Dr. Fechter briefed the analysis plan (slides 14-16), a committee member asked why DPAC was planning to pool across devices to identify familiarity effects. Dr. Fechter replied that they were not sure how large the sample sizes would be for people with valid familiarity scores and in which conditions they would occur. The committee member suggested this might hide the effect of the device, but added that DPAC could look more closely, if needed.

A committee member asked if screen size might override the effect of the device; s/he said, for example, all smartphones have relatively small screens, but tablets and notebooks vary more in this respect. Another committee member noted the complexity of the analysis, citing the fact that Apple notebook and tablet screen sizes are larger than their Dell and Samsung counterparts. The first committee member reiterated that screen size might be a critical variable. Dr. Fechter said they were considering examining three categories of screen size: 13”, which is the current requirement, 9.5”, and smaller sizes. A committee member asked if the text is compressed on the smaller screens, and Dr. Fechter said smaller screens require more scrolling than larger screens. She added, however, that most of the items fit on a single screen, even the smaller screens. She also said the user can zoom/pinch in and out, as required to make the text larger or smaller.

Referring back to the sampling design (slide 9), a committee member asked what determined the planned sample sizes. Dr. Fechter said they needed sample sizes large enough to look at item difficulty (slide 16). She said the recommendation was for 500 examinees per device, which added up to 14,000 for all devices. The committee member replied that they did not need anything close to that number to obtain significance, and that they should give more thought to how much of a difference is of practical importance. Dr. Fechter responded that the large number of comparisons that might be required was key in setting the sample size, and that the determination was made using a sample size analysis.

A committee member recommended including device familiarity as a main effect as well as in the interaction effects. S/he said the device may demonstrate an effect, but that being able to use a device effectively would depend on familiarity. Dr. Segall asked if the committee member was suggesting using familiarity as a covariate. The committee member replied that would be using it as a main effect, but that they needed to test the interactions first. The committee member clarified that s/he was talking about one interaction, saying that the others would be confounded in ways that cannot be untangled. Dr. Fechter asked if the committee member believed the analysis had been over-simplified, and the committee member replied that it is just a complex environment and a product of what is available in the device universe. At that point, Dr. Segall said they had tried to identify combinations of device types, device models, and operating systems that were popular rather than combinations that people do not use. The committee member said s/he liked that approach. Another committee member remarked on how a simple research question can induce such a complex experiment.

Next, a committee member inquired as to whether DPAC had already coded all the various types of item features. Dr. Fechter said they had, but that she did not have any examples to show, as those were presented in the previous DAC meeting’s presentation on the topic. She said an example of an item feature would be items that required scrolling, or an item with a graphic. She also mentioned that the presentation of a graphic might have been modified per the device, specifically for the Assembling Objects subtest. To clarify, she said response options are sometimes presented in a 2x2 matrix instead of in a horizontal line. She commented that previous research has shown that distance from the stimulus has been found to affect response choice, so they are likely to find differences. The committee member commented on the difficulty of testing that hypothesis. Dr. Fechter explained that it is difficult to separate the item design effect from the device effect. However, if this is the only subtest where performance differences are seen, and it seems to relate to the item feature, it would be reasonable to assume that the difference is due to the item layout. Dr. Fechter further explained that she plans to conduct item-level calibration analyses to compare device conditions. The committee member then argued that they could not separate the item design effect from the device effect because the design was implemented for the device. Dr. Fechter replied that they would conduct different analyses using multi-group calibration at the item level to compare item difficulty values. She said they would note the items that are flagged for difficulty and look for patterns consistent with item features. After a pause, the committee member suggested that DPAC could use an explanatory item-response theory (IRT) model, and then Dr. Segall proposed a Linear Logistic Test Model (LLTM), but the committee member said s/he did not think that would be required. Another committee member then said s/he believed they would see the patterns. The first committee member, referring to the sampling plan, commented that there were a lot of items, and another committee member asked if all the items were unique. Dr. Fechter replied that they were. The committee member concluded the discussion by asking when the analysis would be complete and said the study looked great and would be helpful of presented in a field publication. Dr. Segall said they would begin the analysis in the October 2019 timeframe.

11. Adaptive Vocational Interest Diagnostic (AVID) Initial Evaluation (Tab N)

Dr. Cristina Kirkendall, ARI, presented the briefing.

Dr. Kirkendall began the presentation by explaining that the U.S. Army has approximately 140 entry-level military occupational specialties (MOS) and that for an interest assessment to be useful for classification it would need to be applicable to all of them. In a survey of over 24,000 Soldiers, "perceived fit" with MOS was the top reason for selecting that MOS. The potential benefits of a vocational interest inventory include providing recruits information about the MOS in which they will be most successful, predicting valued work outcomes across the Army, and minimizing the effects of poor Soldier-MOS fit. The goal of this research is to develop a new generation vocational interest assessment that incorporates recent research and advanced statistical techniques.

Dr. Kirkendall continued by explaining that most major interest measures assess only six primary dimensions, and that most Army jobs cluster into one or two of these. Therefore, broad interest dimensions may not be flexible enough to select and classify Soldiers across a wide range of jobs. To increase assignment potential, the goal was to develop an assessment of basic interests that may be more useful in differentiating across jobs, with a focus on identifying a comprehensive list of basic interest dimensions that would be useful in the Army. The AVID is an IRT-based, computer-adaptive assessment with a forced-choice format. It will more accurately measure the entire range of interests and reduce testing time. It is easily customized to predict performance across a broad range of jobs. The basic interest scales used in AVID are based on a review of previous ARI work on interests and a review of the literature. Dr. Kirkendall then showed a breakdown of the 20 basic interest dimensions that were identified for pretesting. Approximately 1,000 test statements (50 per dimension) were written, and pretest data were collected from 3,300 enlisted Soldiers in reception battalions and basic training. Pretesting established item parameters and social desirability ratings. The correlations between AVID scales and the O*NET Interest Profiler were examined to establish construct validity, which was generally confirmed. This allowed for the development of both static and adaptive forms of AVID.

The static form of AVID includes 123 item pairs to assess 16 of the 20 AVID dimensions. Science, Personal Service, Finance, and Sales were excluded to reduce total testing time. Data were collected from two samples for the validation study. The first included Soldiers in four high-density MOS (Military Police, Combat Medics, Motor Transport Operators, and Wheeled Vehicle Mechanics). To increase the sample size, a fifth group of heath care MOS was included. The second sample included 1,999 Soldiers who were taking part in another project, the majority of whom were E-3s (29%) or E-4s (47%), with the largest MOS being Infantry (n = 343). AVID and the Army Life Questionnaire were administered to both samples, along with Soldiers' ratings of their MOS. The data were cleaned using items to detect unmotivated responding, and 124 Soldiers were excluded from sample 1, and 218 were excluded from sample 2. Correlations and regression analyses were run. The validity of vocational interests is highest when considering the match between individuals and their jobs, so analyses focused on identifying the validity of AVID using Soldiers' fit with their MOS.

Dr. Kirkendall then presented a series of tables and charts showing results for the two samples. The validity of interest fit for predicting overall performance in each MOS was examined. Interest fit was operationalized using regression models with both AVID dimensions and MOS interest scores in the model, which resulted in higher validities. The validities of AVID were often larger than for the TAPAS when interest fit was calculated. The differences across MOS suggest that the AVID may be used for MOS classification. Correlations between MOS-specific composites of AVID scales were strong, but indicated differences across MOS. This provides a useful initial look at AVID, but more research is needed to examine validity in a broader range of MOS and to evaluate the adaptive version. The next steps include conducting a concurrent validation of AVID by collecting additional validity evidence using the static forms and including the four dimensions that were omitted in the initial validation. Five different MOS will be targeted, and simulations will be conducted to evaluate the best method of calculating the match between interest dimensions and job characteristics. The implications of "fit-bandwidth" will also be examined, given that some individuals may be interested in many jobs and others only one. In addition,

> MOS interest profile data have only been collected on a few jobs, so additional data will be collected to explore clusters of MOS with similar interest profiles. Longitudinal data will also be collected from early career Soldiers at reception battalions and then linked to end-of-training data. The outcomes of interest will include 6-month attrition, the Army Life Questionnaire, and performance ratings. Dr. Kirkendall then displayed a graphic showing the timeline for the project. She concluded that AVID has the potential to be a valuable addition to ARI's non-cognitive measures and to contribute to whole-person assessment that more accurately predicts performance behaviors and attitudes. Improved personnel assessment enables greater flexibility to accommodate changes in force size, structure, mission demands, budgets, and availability of qualified applicants. It can also lead to improved person-job match, performance, and retention, saving money by reducing attrition.

As Dr. Kirkendall described the Holland Occupational Themes (also known as RAISEC) model on which the AVID was based, she commented that there are not a lot of artistic jobs in the Army. A committee member asked if photographer was a military occupational specialty (MOS). Dr. Kirkendall said it was, but that it was very small and not a place for the AVID study to invest resources.

When Dr. Kirkendall briefed the initial validation results for Sample 1, a committee member asked if the regression weights were standardized, and Dr. Kirkendall said they were. The committee member then asked if Dr. Kirkendall had a table of correlations. Dr. Kirkendall said she did not have one with her. The committee member, referring to the interest-fit concept shown on slide 7, asked how job fit had been measured. Dr. Kirkendall replied that each test had an item that asked participants about the relevance of dimensions to their job. She explained, however, that the results shown on slide 9 were not job-dependent. Another committee member remarked that writing was the third largest predictor for overall performance. Noting that it was a negative predictor, the first committee member suggested that it may be acting as a suppressor. The other committee member agreed and commented that the charts on slide 10 used the overall dimension weights from slide 9. S/he then asked if there was a relationship between the AVID quintiles and outcomes, as well as, what was an AVID composite? Dr. Kirkendall replied that the overall composite scores were developed with the dimension weights shown in the last column of Slide 9 and said everyone has a composite score. The committee member summarized, then, that it was their weighted score based on the overall composite.

As Dr. Kirkendall presented the results on slide 11, a committee member asked: (a) if the AVID scores were calibrated independently, (b) how were they scored, and (c) were item parameters established through pretesting? Dr. Kirkendall replied that the survey consisted of IRT-based forced-choice measures. The committee member then commented that they probably started with thousands of statements but were now down to 123 item pairs. Dr. Kirkendall said that each dimension has between 30-60 statements and there was not a large loss of statements during the pretesting phase. The committee member then asked if the algorithms had been programmed based on all the items or just the operational items, and Dr. Kirkendall said she was not sure. She also clarified that items presented pairs of statements, and so 123 items represented 246 statements across all dimensions. The committee member said there were, then, about six to eight pairs per dimension, for 16 dimensions. Dr. Kirkendall explained that the statements were paired similarly to how the TAPAS statements had been paired. Dr. Heffner added, however, that the adaptive AVID form would have more items than the current static form. Dr. Kirkendall clarified that she was presenting the first part of the study and that the adaptive part would be conducted later. The committee member commented that management was relevant for all five jobs shown. Dr.

Kirkendall explained that it was easiest to obtain participants from the Advanced Leader Course (ALC), and NCOs in those ranks are typically in managerial type positions, which might explain why management showed up so prominently. Another committee member asked about the variability in the sample, and Dr. Kirkendall said there would be an example illustrating that later in the presentation.

As Dr. Kirkendall explained the differences in the sample 2 table on slide 12 and the sample 1 table on slide 9, Dr Heffner explained that the management loadings for sample 2 were lower, possibly because the participants were of lower rank than in sample 1. Dr. Kirkendall pointed to the weight of 0.42 under "motivation to lead" and said management was one of three motivation-to-lead components.

As Dr. Kirkendall summarized the initial validation (slide 16), she described that the correlations between the MOS-specific composites of AVID scales were strong but still indicated differences across MOS. Here, a committee member asked if there was a way to estimate reliability and if the correlations had been corrected for nonreliability. Another committee member said s/he felt like s/he was missing something and asked if the composite was different here. Dr. Kirkendall replied that they had used the same weights to create each of the MOS-specific composites, but that, previously, they had provided results for specific MOS, whereas now (on slide 16), each Soldier had a score for each MOS-specific composite and the results were based on the entire sample. The committee member asked if the regression tables presented earlier (on slides 9 and 12) showed other outcomes, and Dr. Kirkendall said yes, and that overall performance was a composite of the outcomes shown earlier.

A committee member said s/he thought another slide did not include a performance measure, but only an interest measure. Dr. Kirkendall referred to slide 9, explaining that it showed the overall AVID composite scores and said slide 11 presented the weights used to compute the MOS-specific composites. She said the table on slide 11 showed the MOS breakout and noted that the sample sizes were smaller. Dr. Kirkendall further clarified that slide 11 showed the extent to which interests predicted MOS performance. The committee member then asked how to interpret the correlations shown on slide 16. Dr. Kirkendall explained that a person receives AVID scores for multiple jobs and that the correlations, though high, still indicate that the instrument reveals differences between MOS.

Changing topics, a committee member commented that the composites appear to lack reliability to some degree, and that they should be corrected before using them to suggest the instrument is useful for classification. Another committee member said the jobs could be clustered but, looking at the RAISEC dimensions shown on slide 4, said she could not find management. Dr. Kirkendall said management was the term they were now using for leadership, because the items were more indicative of management than leadership. The committee member then asked if it was of any interest to see the correlations without the outcomes as a function of the RAISEC model. Dr. Kirkendall said she could collapse back into the RAISEC dimensions, and that might gain or lose something. Another committee member suggested that it might show which of the RAISEC dimensions were most predictive. When Dr. Kirkendall said they could go back to that model, the committee member suggested that most would come from the Realistic set.

Discussion of the initial validation study concluded with a committee member expressing his/her understanding of the analysis and communicating wariness of the conclusion that MOS-specific composites indicate differences across MOS. S/he suggested that the procedure did not explain a lot of the variance, though it seemed optimal for predicting overall performance. Another committee member commented that overall performance is not "task performance," as it is traditionally defined, and noted the extent to which MOS scores were correlated. Dr. Pommerich asked if ARI had looked at incremental validity. Dr. Kirkendall said the data in sample 2 could be used for that, but the sample is not large enough yet.

When briefing the next steps in the research (slide 17), Dr. Kirkendall asked the committee if they had any suggestions on how to create composites for the simulation analysis. Dr. Donna Duellberg (US Coast Guard) explained that Sailors often pursue training using Tuition Assistant dollars and that it may be interesting to investigate how interest plays into the type of training Soldiers pursue using Tuition Assistance dollars. That is, do they stick to training relevant to their current MOS or pursue outside interests because that is what they are more interested in?

As Dr. Kirkendall briefed the potential tasks (slide 18), a committee member noted that, with only seven or eight items, the correlations might be as high as they can be. Another committee member encouraged Dr. Kirkendall to present the correlations and regression weights for all the variables to better illustrate the connections; s/he said doing so might reveal something about the negative weight for writing, for example. The first committee member said it looked like Dr. Kirkendall had done a factor analysis, and Dr. Kirkendall replied that a factor analysis had been used to create the dimensions. The committee member then suggested investigating the use of Confirmatory Factor Analysis (CFA) to create the composite score, which would account for variance in measure scores instead of the current approach, which would account for variance in overall performance scores. The committee member then asked if Dr. Kirkendall wanted to account for variance in the measure (predictor) versus the variance in outcome (performance). The committee member provided a recommendation to use factor loadings to inform composite score weights instead of validity coefficients. Dr. Kirkendall replied that they did want to account for variance in the predictor and would look into the use of CFA for creating composite scores.

12. <u>ASVAB Time Limits</u> (Tab O)

Dr. Furong Gao, HumRRO, presented the briefing.

> Dr. Gao began by explaining that the purpose of the analyses she conducted was to evaluate the current ASVAB testing time limits to ensure examinees have sufficient time to complete the tests. The current time limits were set based on analyses of data collected in the Forms 5-9 equating study. They have been in place since the initial Forms 5-8 were implemented in 2009 for tests without seeding and 2014 for tests with seeding. Currently, when items are seeded, 15 are added per subtest. Dr. Gao showed a table that listed the number of operational items per subtest and the time limits with and without seeded items. In the seeding configuration, examinees are randomly assigned to one of five groups and seed items are randomly dispersed among the operational items in the subtest. To determine if sufficient time is allocated for each subtest, empirical response time distributions were examined, a theoretical statistical distribution/model was fit, and the various quantiles of the statistical distribution were evaluated with regard to the current time limit.
>
> The data came from 2015-2018 WinCAT and *i*CAT administrations. The analyses were conducted for each subtest by analyzing examinees' response time distribution on the test and by (a) seeding status (without or without seeded items); (b) year of testing; (c) test administration platform (WinCAT, *i*CAT); and (d)

combining platforms across 2015-2018. The analyses examined trends and variations across years and platforms. Dr. Gao then showed a graph depicting response time distributions for GS without seeded items by year and platform, which indicated they were similar across years, with *i*CAT examinees tending to take slightly longer. Another chart showed response time distributions for GS with seeded items by year and platform, which were similar. Distributions for EI, with and without seeded items, were also displayed. The questions that arise are why *i*CAT examinees take longer, and should time limits be adjusted as a result? Dr. Gao stated that additional analyses suggested that examinees who took the *i*CAT seemed to have slightly lower abilities, which likely explains the response time differences. Examination of the empirical response time distributions indicate that some tests may be slightly speeded (e.g., GS), and some tests may require less time than currently allocated (e.g., AI).

Dr. Gao then showed a table listing the proportion of examinees who did not complete each subtest by platform and combined. The data suggest possible speededness in GS, AR, and MK, with the effect being more severe for MK. She noted that if an examinee does not complete a subtest, he/she receives a "penalized" theta score, with non-completed items scored as though they were answered randomly. Another table showed the proportion of non-completes by seeding status. Regarding the impact on AFQT scores, Dr. Gao indicated that among the 932,746 examinees (across platforms and years) who took the ASVAB, about 4% received a lower AFQT score due to incompletion of at least one of four subtests (AR, MK, WK, PC). Of these, 1,041 fell below the cut score of 31 and 879 fell below the cut score of 50.

To estimate the appropriate time limit, Dr. Gao fit a theoretical statistical distribution to the empirical test response time distribution. Response time (item or test) generally follows a log-normal distribution. If the test is speeded, then the empirical distribution will be right-censored at the time limit. She examined the 95th, 98th, and 99th quantiles of the fitted distribution to evaluate the current and potentially adjusted time limits. Based on these results, the recommendation is that the time limits be set so that at least 90% of examinees can complete the test in the time given. Examinees do not need to take all the time allowed. When a given subtest is completed, they can move on to the next subtest. Dr. Gao concluded by showing a chart displaying the new time limit recommendations.

As Dr. Gao briefed slide 8, Dr. Segall responded to a committee member's question about differences in the administration between the iCAT and WinCAT. He said there were differences in the technologies and in the testing population. He explained that the WinCAT was administered at MEPS and the iCAT was administered at remote processing centers. He also said applicants tend to score higher on the WinCAT, possibly because recruiters take the more promising applicants to the MEPS in the hope that they will process faster. He summarized by saying that the difference in iCAT and WinCAT scores was probably due more to aptitude than the test.

When Dr. Gao described the proportion of examinees who did not finish the tests (slide 13), a committee member asked if examinees had to end the test and move on to the next test when time expired. Dr. Gao replied that they did. Another committee asked if the scores of examinees who did not finish were flagged. Dr. Segall explained that all scores were used, regardless of whether examinees had finished the test. He noted, however, that non-finishers were tracked and could be identified if needed.

In response to Dr. Gao's presentation of new time limit recommendations (slide 24), a committee member remarked that the mean testing times had not changed that much, and that this was consistent with the marginal levels of skewness of test time distributions. S/he added, however, that a small percentage of examinees had not completed the test and that changing the time limit to accommodate that segment of the population could result in rather large changes in allotted administration time. Dr. Segall agreed but said that the length of time required to complete individual subtests did not correlate highly with overall testing time. When a committee member commented on the 16-minute increase for AR, Dr. Segall replied that DPAC would continue to monitor testing time, but that they did not expect to see a big difference.

The committee member then asked if examinees see a timer when they are testing, and Dr. Segall said that they do. The committee member suggested that the presence of the timer might affect examinee behavior. Dr. Segall replied that it could, but that they cannot go back to previous items, which would tend to move them along more expeditiously. He also said that the total time would only increase for the 1% of examinees who required it. In response, Mr. Aswell asked if all examinees would have the longer amount of testing time blocked off. Dr. Segall explained that, at MEPCOM, examinees are only told they have a half day or a full day for testing and that the fact that some of them might be there a little longer than others would not affect the guidance they receive. He added that extended testing by a few individuals would not cause operational issues. Mr. Aswell asked if examinees can get up when they are finished, and Dr. Segall said they can. Mr. Aswell also asked if the extended time limits would eliminate recent issues associated with examinees not having sufficient time to complete the overall test, and Dr. Segall said that it should. He said the extended time limits account for 99% of the test taker population.

Continuing the discussion, a committee member pointed out that the current rate of item seeding would not continue indefinitely. Dr. Segall agreed and said that after the current seeding effort is complete, DPAC wound continue to add only about half the number of seed items that are currently being administered. He said this would result in only about 10 additional minutes of testing time.

Two committee members then noted the very small proportion of examinees who would still not be able to complete a subtest even with the extended time limits. Dr. Segall responded by explaining that examinees who do not finish an adaptive test would receive a penalty on their scores. He added, however, that an option would be to take an estimated score based on the number of items completed (e.g., 14). He said this would be akin to scoring without penalty. A committee member then asked if this approach would result in a slight truncation of range, and Dr. Segall said that it would. Another committee member asked if a penalty would be applied if an examinee responds randomly on items he/she did not have time to answer thoughtfully. In response, Dr. Segall said the examinee would get credit for those items answered correctly, which would not represent a penalty (a penalty is currently applied for items not answered).

To conclude the discussion, Dr. Velgach asked the committee if it concurred with the recommendations presented on slide 23. She said that if the committee concurred, DPAC could implement the recommendations in short order. Dr. Segall commented that the Services are looking for ways to qualify more people and that the addition of testing time may facilitate this end. A committee member observed that more examinees who know time is running out may respond more quickly, and so providing more time might result in a more accurate measurement. S/he then said that if the extended time limits were affordable, it should be alright. S/he also added, however, that it would only affect 1-2% of examinees. Dr. Pommerich remarked that it was an issue of fairness. The committee member said s/he could appreciate the impact on classification and that the committee would note that in its letter. Dr. Velgach replied that DPAC will need to submit a letter to AP for approval to make updates, which she said would be good news to the Services. Mr. Aswell said the change would affect a lot of people, especially for the Army.

13. TAPAS Review Update (Tab P)

Dr. Tim McGonigle, HumRRO, presented the briefing.

Dr. McGonigle began by explaining that some stakeholders have raised technical concerns about the TAPAS, especially related to low test-retest reliability. RAND recently completed an independent evaluation of the reliability and validity of TAPAS. They found small, significant incremental validity over education credential in predicting attrition and evidence of low test-retest correlation in some conditions. DPAC requested a TAPAS Evaluation Project (TEP) to independently review the body of TAPAS research and make recommendations regarding the readiness of TAPAS for operational use. The evaluators were to review related research conducted by the Services, both on TAPAS and other instruments such as interest inventories. They will then comment on the readiness of TAPAS for operational use and make recommendations for future research and development. Dr. McGonigle listed the members of the TEP.

The first panel meeting was held in October and included attendees from DPAC, ARI, the US Air Force, the US Navy, RAND, and Drasgow Consulting Group (DCG). Evaluators received a copy of the RAND report before the meeting. The agenda focused on TAPAS research and development, with presentations made by DCG, ARI, and Air Force representatives. Detailed minutes were provided to the TEP. Dr. McGonigle then summarized the highlights from the presentations, including:

- Dr. Steve Stark from DCG reviewed the research leading to the development of TAPAS, described its underlying personality and measurement theory, summarized the history of its development and use, and described research to improve reliability, use of marginal reliability index, recalibration of item pools, smart-CAT algorithm, and use of item triplets.
- Dr. Chris Nye of DCG and Dr. Len White of ARI described the validity of TAPAS composites for predicting Will-Do, Can-Do, and Adaptation criteria; assessed the incremental validity for Will-Do (over AFQT) and Adaptation (over AFQT and Educational Tiers); described different predictors of 36-month and misconduct attrition; summarized evaluations of validity for MOS-specific TAPAS composites; compared Soldiers' predicted performance in their current MOS to their predicted performance in other MOS; and examined use of TAPAS for in-service testing, such as to identify high-potential individuals for special duty assignments that are only available to experienced Soldiers (Special Forces, Recruiters, Drill Sergeants, Instructors).
- Mr. John Trent of the Air Force presented research comparing TAPAS scores across three conditions ([1] operational, pre-accession, [2] retest administration post-accession under honest conditions, [3] retest administration post-accession under directed faking conditions); presented comparisons of TAPAS-Five Factor Model (FFM) reliability to other FFM measures from two meta-analyses; summarized results suggesting that Physical Conditioning, Non-delinquency, Dominance, and Adjustment most strongly correlated with training and job outcomes across select Air Force careers; and indicated that SMEs rated Adjustment, Achievement, Self-Control, and Even-Tempered as most important across three Air Force careers.

Evaluators also reviewed and approved the TEP charter, elected a chair, and identified topics for the second meeting.

The second TEP meeting was held in January 2019, and was attended by representatives of DPAC, the Air Force, ARI, the Marine Corps, RAND, DCG, and the Office of the Undersecretary for Personnel and Readiness of DoD. Evaluators requested relevant articles, chapters, and technical reports recommended by the Services, which were provided by ARI and the Air Force. The agenda focused on evaluation and operational use of TAPAS. The presenters included representatives from RAND, the Personnel Decisions Research Institute (PDRI), ARI, the Air Force, and the Marine Corps. Individual Q&A sessions were conducted with RAND and ARI personnel, and the TEP was again provided detailed notes on the proceedings. Highlights of the presentation included:

- RAND found test-retest correlations ranged from .19 – .59; reported low but significant validity for some TAPAS scores in predicting attrition, but little practical effect on reducing attrition;

discussed simulation showing observed levels of correlation may be due to large proportion misrepresenting or responding randomly; encouraged both operational and lab-based research strategies to improve TAPAS.

- PDRI described a project aimed at reviewing previous research findings and analyzing data to make recommendations regarding which aspects from the Navy Employs Enlisted Computer Adaptive Personality Scales (NCAPS) and Self-description Inventory (SDI) would be worth incorporating into the TAPAS system; coded 30 articles that had psychometric information on TAPAS; presented a pattern of predictive relationships between TAPAS facets performance and attrition measures; and presented subgroup differences.
- ARI presented simulated data to examine the impact of range restriction on test-retest correlations and retest score gains, and to predict Can-Do, Will-Do, Adaptation, and Good Conduct criterion composites.
- The Air Force reviewed current operational use of the TAPAS in the Air Force (for some classification; not for selection); described testing policies and procedures, including retesting policy; discussed positive (validity, faking resistance, flexibility) and negative (limitations on administration time) experiences with TAPAS; described the potential future of TAPAS in the Air Force, including adding classification models, modifying TAPAS format and content, and potentially using TAPAS in selection.
- The Marine Corps described current testing policies and use (administered to all recruits but scores automatically waived); and discussed the operational challenges the Marine Corps faces, such as small staff.

Dr. McGonigle concluded by discussing upcoming TEP activities, including a third meeting scheduled for May and a final meeting in the summer of 2019. He then presented a list of the articles, chapters, and technical reports provided by ARI and USAF.

As Dr. McGonigle commented on the test-retest correlations (slide 7), a committee member asked about the intervals between tests. When Dr. McGonigle said the test-retest intervals varied, Dr. Heffner clarified that it was as least 30 days for Army personnel. Dr. Manley said it could range up to 19 months. Dr. Velgach then explained that there were several reasons for retaking the test, and this prompted a committee member to inquire about the equivalency of the first and second administrations. Dr. Heffner replied that the test was adaptive, which meant that retests did not include the same items as the original tests. Another committee member noted that using alternate forms would tend to reduce the correlation between tests. Dr. Heffner added that participants might also receive feedback on their scores, which could also affect second administrations. Dr. McGonigle summarized the situation as being one from which it was difficult to draw inferences.

As Dr. McGonigle described upcoming activities (slide 8), Dr. Velgach noted that the request for additional discussion of reliability had come from RAND and not from the panel members. She said RAND had wanted the opportunity to comment further on the matter of test-retest correlations versus test-retest reliability.

Regarding operational policy, a committee member asked if the panel was planning to address using the TAPAS for more than just predicting attrition. Dr. McGonigle replied that the panel's charge was to make recommendations on the test's operational use. The committee member then asked if the original issue was RAND's concern with the prediction of attrition. Dr. McGonigle replied that the RAND report had said many things, but that its stated concerns with low reliability and low incremental validity in predicting attrition were the driving factors behind the panel's discussions. He added, however, that even low incremental validity was likely to have a large positive impact on overall costs (accessions and training). Dr. Velgach responded that the panel's mission was about more than just the RAND report; she said the mission concentrates on evaluating the proper uses of the test.

A committee member asked what the panel hopes to gain from looking at the dimensionality analyses. S/he also asked how many items were in each facet. Dr. Manley replied that each administration includes 8 items per facet, and Dr. McGonigle explained that the pool of items per facet is much larger.

As Dr. McGonigle presented the list of TAPAS-related literature (slide 9), he explained that Dr. Tracy Kantrowitz from PDRI had presented a summary of the results of each study, but the panel found it difficult to interpret. The panel requested a quantitative summary of the TAPAS-related literature to show reliability and validity results under each condition. He said his team was working with Dr. Kantrowitz and the Services to gather the literature and extract the relevant information. A committee member commented that most of the references were produced by Dr. Fritz Drasgow and his team and suggested that there might be a bias in the corpus of literature. Dr. McGonigle replied that the list of sources he had presented was not intended to represent the universe of literature on the subject, and he added that they were providing a broader summary of the literature as well. Another committee member commented that many of the Drasgow sources had been published in peer reviewed journals. Dr. McGonigle agreed that the large number of Drasgow publications was an issue, but he said the summary document would include approximately 30 other sources as well.

A committee member asked if the panel was going to produce a report after its last meeting. Dr. McGonigle said the report would be completed in October 2019. He said the report would include the panel's recommendations on use of the TAPAS as well as ideas for further research. A committee member said that sounded great.

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

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)
- Adverse Impact
- P*i*CAT/VTEST (Verification Test) Updates
- 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, Working Memory, Abstract Reasoning including Adverse Impact)
- Adding New Non-Cognitive Measures (personality and interest measures)
- Automatic Item Generation
- Web and Cloud efforts
- Device Evaluation Study

As Dr. Segall presented a proposed list of topics for the next DACMPT meeting, he said the meeting would likely be held prior to October 2019, which would affect the list of projects that could be briefed. In response, a committee member asked for an update on the device evaluation study, and Dr. Fechter said she should have some preliminary analyses by then. The committee member also asked for an update on automated item generation, and another committee member mentioned Abstract Reasoning. Dr. Pommerich said she was trying to track down sufficient data on the Abstract Reasoning Test in order to conduct analyses. Another committee member requested an update on the TAPAS expert panel recommendations.

A committee member then asked if there was an update on the APT. Dr. Pommerich replied that there were no changes and, therefore, she would not have an update. The committee member then asked about the P*i*CAT work. Dr. Segall replied that it was stable and said DPAC was seeing an increase in usage rates, which he said should continue.

Another committee member asked about the progress being made on the across-Service criterion development project. Dr. Kirkendall replied that her team had developed a performance taxonomy and recommended measures, so she should have something to present.

A committee member asked if DPAC could present on the transition to the Cloud. Dr. Segall replied that he and Dr. Pommerich spend over half their time on IT matters related to the transition. He also said he did not know how interesting the process would be to the committee, but that they could present a summary if the committee wants to hear about it. Another committee member then commented about his/her concerns regarding the security of the tests once they are on the Cloud. Dr. Segall replied that they could address that concern. Another committee member asked if the move to the Cloud was DoD-wide, and Dr. Segall replied that it was. The first committee member then referred to the prevalence of hacking, but another committee member suggested that the most vulnerable content is housed locally. S/he then mentioned that DPAC might address security as well as the technology structure and how it differs from the current process.

A committee member asked if DPAC could address the ASVAB validity framework, and Dr. Pommerich said they could have something on that. A committee member summarized by saying that sounded like a good agenda.

On the broader topic of future DACMPT meetings, a committee member asked if AP could project out at least one year in scheduling future meetings. Dr. Velgach pointed out that the DACMPT schedule depends, in part, on the ability of MAPWG and Service members to support the schedule. In addition to timing, she mentioned funding considerations for members of those organizations. Dr. Pommerich added that DPAC generally projects in 6-month cycles. Another committee member replied that a year out would be good and that three months was too close to allow proper planning. Dr. Segall responded that DPAC could try to support the planning of DACMPT meetings one year out; he said they could accommodate MAPWG meetings either in person or by phone, as required. A committee member asserted that March and September were not good months for DACMPT meetings.

Finally, a committee member suggested that 12 programmatic briefings were too many to cover in one meeting. Dr. Segall replied that he would like to lighten the agenda, but that the number of topics is driven by AP requirements for DACMPT input. Dr. Velgach said it is important for AP to have DACMPT feedback on certain efforts before they can move ahead.

Following the discussion of the agenda for the next meeting, Dr. Velgach asked if the Services had any further input or comments before the meeting came to a close. Mr. Aswell responded, saying that senior Army leaders had recently communicated interest in three areas related to accessions testing. First, he noted their interest in having a Spanish version of the ASVAB, which he said they believed would increase the number of accessible recruits. He then reported that their interest had been tempered somewhat upon hearing an explanation of what it would take to develop such a test. Additionally, he reported explaining to leaders that the Army already accepts applicants with ASVAB scores as low as 21 if language is believed to be the cause of the low score. He also said he told the leaders that between 500 and 1,000 such applicants a year are sent to the Defense Language Institute for an opportunity to improve their English. Second, he said the leaders were thinking about using the ASVAB to screen in ROTC cadets, or even to screen in officers. He said he was not sure how this would play out, but that it should be kept on AP's radar. Third, he mentioned the leaders' sustained interest in the use of calculators on the ASVAB. He also said that recruiters continue to ask that calculators be allowed, because they believe it will result in qualifying more recruits. He added that he did not think the desire to allow calculators would go away, even given an understanding of the collateral impact it would have on the testing program as a whole.

In response to Mr. Aswell's report, a committee member said s/he had seen increased interest in the public education sector for Spanish versions of non-English-loaded tests, like math. However, s/he expressed uncertainty that such an approach would improve measurement. Another committee member added that it might be feasible to provide instructions in Spanish, if nothing else. Dr. Velgach replied that she thought the instructions had been administered in Spanish in Puerto Rico. Dr. Velgach further reiterated that providing the test in one alternate language would likely beg the need to provide it in other languages, which would lead to a difficult series of decisions to determine which languages to target. Dr. Velgach said she was unaware of any official interest in using the ASVAB for officer selection, but that she was familiar with the push by the Secretary of the Army to use calculators on the ASVAB.

As the meeting closed, the committee chair thanked all participants for their contributions and commented on the large amount of work that is being accomplished. Another committee member said it was nice to see the collaboration with HumRRO. Dr. Segall replied that HumRRO has been a great resource and that HumRRO personnel do not get the credit they deserve because so much of their work is done behind the scenes. Dr. Pommerich noted that some HumRRO personnel work in DPAC's facility. Dr. Velgach then thanked everyone and asked if there were any comments from the public, of which there were none.

Tab A

LIST OF ATTENDEES

Defense Advisory Committee on Military Personnel Testing (DACMPT)
March 28-29, 2019, The Pine Inn, Carmel-By-The-Sea, California

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 Lincoln	Professor Emeritus	DACMPT, University of Nebraska-
Dr. Kevin Sweeney*	Vice President, Research and Development	The College Board
Dr. Sofiya Velgach	Designated Federal Officer (attendance req'd by FACA)	Accession Policy Directorate
Mr. Christopher Arendt**	Deputy Director	Accession Policy Directorate
Mr. Christopher Graves	Senior Scientist	Human Resources Research Organization
Dr. Daniel Segall	Division Chief	Defense Personnel Assessment Center
Dr. Mary Pommerich	Deputy Director	Defense Personnel Assessment Center
Dr. Shannon Salyer	Manager, Career Exploration Center Program	Defense Personnel Assessment Center
Dr. Greg Manley	Personnel Research Psychologist	Defense Personnel Assessment Center
Dr. Tia Fechter	Personnel Research Psychologist	Defense Personnel Assessment Center
Dr. Richard Riemer	Personnel Research Scientist	Defense Personnel Assessment Center
Mr. Doug Keindl	IT Specialist/Applications/Systems	Defense Personnel Assessment Center
Ms. Olga Fridman	Analyst	Defense Personnel Assessment Center
CPT Alex Ryan	Operations Research Analyst	US Marine Corps, Manpower and Reserve Affairs
Dr. Donna Duellberg	Voluntary Education Program Manager	US Coast Guard
Dr. Tonia Heffner	Supervisory Research Psychologist	US Army Research Institute

Dr. Cristina Kirkendall	Research Psychologist	US Army Research Institute
Mr. Paul Aswell	Deputy Chief of Staff for Personnel	US Army, G-1
Mr. Ken Schwartz	Air Force Enlistment Policy	Headquarters, Air Force Personnel Policy
Mr. Brad Tiegs	Testing Director	Headquarters, U.S. Military Entrance Processing Command
Dr. Steve Watson	Director	Navy Selection and Classification Policy
Dr. Tim McGonigle	Program Manager	Human Resources Research Organization
Dr. Matthew Trippe	Senior Staff Scientist	Human Resources Research Organization
Dr. Furong Gao	Senior Staff Scientist	Human Resources Research Organization
Dr. Ping Yin	Senior Staff Scientist	Human Resources Research Organization
Mr. Tom Blanco	Vice President	S&T Consulting
Mr. Hector Jimenez		Public

*Not in attendance
**Participated telephonically

Tab B

DEFENSE ADVISORY COMMITTEE ON MILITARY PERSONNEL TESTING AGENDA

March 28-29, 2019
The Pine Inn
Carmel-By-The-Sea, California

March 28, 2019

0800-0830	Complimentary Buffet Breakfast in Dining Room	
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 to include JAMRS* Brief	Dr. Sofiya Velgach, OASD (M&RA)/AP
0945-1030	CEP* Update	Dr. Shannon Salyer, DPAC/OPA*
1030-1045	*Break*	
1045-1115	Milestones and Project Schedules	Dr. Mary Pommerich, DPAC/OPA
1115-1200	ASVAB* Evaluation Plan	Dr. Mary Pommerich, DPAC/OPA
1200-1230	CAT*-ASVAB New Forms Update	Dr. Matt Trippe, HumRRO*
1230-1330	Lunch – “Dining Room”	
1330-1415	Mental Counters – Rapid Guessing Behavior	Dr. Ping Yin, HumRRO
1415-1500	Mental Counters Think Aloud Plan	Dr. Ping Yin, HumRRO
1500-1515	*Break*	
1515-1600	CAT-Cyber Test	Dr. Furong Gao, HumRRO
1600-1615	*Public Comments*	
1615-1730	Executive Session	Dr. Michael Rodriguez, Chair

March 29, 2019

0800-0830	Complimentary Buffet Breakfast in Dining Room	
0830-0900	Executive Session	Dr. Michael Rodriguez, Chair
0900-0930	Adverse Impact for Special Tests	Dr. Greg Manley, DPAC/OPA
0930-1015	Device Evaluation	Dr. Tia Fechter, DPAC/OPA
1015-1045	AVID* Initial Evaluation	Dr. Cristina Kirkendall, ARI*
1045-1100	*Break*	
1100-1145	ASVAB Time Limits	Dr. Furong Gao, HumRRO
1145-1215	TAPAS* Review Update	Dr. Tim McGonigle, HumRRO
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
ARI = US Army Research Institute for the Behavioral and Social Sciences
ASVAB = Armed Services Vocational Aptitude Battery
AVID = Adaptive Vocational Interest Diagnostic
CAT = Computer Adaptive Testing
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
HumRRO = Human Resources Research Organization
JAMRS = Joint Advertising Market Research & Studies
OASD (M&RA)/AP = Office of the Assistant Secretary of Defense (Manpower & Reserve Affairs)/Accession Policy
P*i*CAT = Unproctored Pre-Screening Internet CAT-ASVAB
TAPAS = Tailored Adaptive Personality Assessment System

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

April 24, 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 March 28-29, 2019, in Carmel by the Sea, California. Below, we provide summaries and recommendations from the DACMPT. The DACMPT members appreciate the commitment of all presenters, their thorough presentations, and thoughtful responses to questions and discussion.

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

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

Accession Policy Update

As Chris Arendt was unable to attend the meeting, Dr. Sofia Velgach provided a briefing on the ability of the various services to meet their recruitment goals. The Army has fallen short of active duty recruitment goals and has had to select a small percentage of AFQT CAT IV applicants. The Navy Reserve has also fallen short of recruitment goals. Dr. Velgach presented results of a survey conducted by JAMRS regarding perceptions of the military among the youth market. The survey indicated that today's youth are disconnected from the military and are not familiar with military life. They recommend a sustained outreach effort to youth to change perceptions, knowledge of the military and the benefits that accrue from military service. This is particularly important as the negative outcomes of service in the military often get the most attention from the media. The DACMPT agrees that the services must continue to direct resources and effort to provide accurate and balanced views of the nature and outcomes associated with military service.

Career Exploration Program Update

Dr. Salyer provided a briefing on the CEP. The CEP metrics continue to be strong and promising. In addition the CEP iCAT numbers are increasing. As the CEP continues to reduce the number of paper-pencil testing sessions and is encountering bandwidth limitations in some areas. It was also noted that there continues to be only one form, as another backup form is currently being created. Dr. Salyer reviewed numerous steps and advances as the program staff continue to respond to the expert review panel and needs assessment. Upon review of the PTI (post-test interpretation) proficiency training materials, the DACMPT noted that the materials are comprehensive and highly professional.

Recommendations: The move to iCAT administration is promising and the Department plans to move test administration to the Cloud will be an important step. Along with continuous monitoring of school bandwidth to support iCAT administration, it is important to track and increase the availability of personnel for administration, particularly in those states that are now adopting CEP statewide. One potential source of administrators is the school district assessment coordinators and their designees who are trained to support standardized state test administration. The DACMPT supports the collaboration with CAVEON to monitor online presence of potentially compromised materials. The DACMPT also supports efforts by the CEP to clarify the purposes of the CEP, particularly with respect to claims that ASVAB CEP may be used for purposes other than career exploration as suggested by some in response to ESSA requirements.

Milestones and Project Schedules

Dr. Pommerich provided a briefing on the status of ASVAB R & D efforts, including milestones and project schedules for each effort. Topics covered in the briefing were ASVAB Development, Career Exploration Program (CEP), ASVAB and ETP Revisions, the AFQT Predictor Test, Air Force Compatibility Assessment (AFCA), and the Defense Language Aptitude Battery. Many of the topics were ones that received additional focused briefings during the DACMPT meeting (specifically the development of new CAT-ASVAB Item Pools, Career Exploration Program, Mental Counters, Cyber Test, Non-cognitive Measures (TAPAS and AVID). Summaries and recommendations from those briefing are included in this letter.

Most of the projects and efforts are on schedule. Notable new efforts include the use of automated item generation (AIG) for Arithmetic Reasoning (AR) and Mathematical Knowledge (MK) (in progress) and the exploration of AIG for General Science (GS). Preparation of Technical Bulletins is ongoing, with completions anticipated in 2019 for CAT-ASVAB pool 10 and CEP *i*-CAT, and for pools 11-15 in Summer 2020. New cognitive tests (including nonverbal reasoning) are under consideration, including an abstract reasoning test. These efforts are often centered in service-specific efforts using small sample sizes that limit generalization to other services and uses. The Air Force Compatibility Assessment (AFCA) has been programmed to be available on the WinCAT platform in Spring 2019 but deployment is delayed until final stages of QC has been completed. It is currently intended for exploratory use only.

Efforts to extend test availability include a plan to migrate special tests to web/cloud delivery over the next 2 years (e.g., Cyber Test, Coding Speed, TAPAS). The plan is to decommission WinCAT in 2020 and deploy *i*CAT to the cloud at that time.

Comment: The DACMPT appreciates these regular updates on ASVAB R & D efforts.

ASVAB Evaluation Plan

Dr. Pommerich provided a status report on the ASVAB evaluation plan, primarily focused on creating a framework for next generation ASVAB and continuous maintenance. This includes addressing the question as to what tests should be administered as part of the ASVAB or on the ASVAB platform, essentially about the distinctions between the battery and special tests. The DACMPT inquired as to whether the special service-specific interests tests would be included in this larger plan. In addition, the DACMPT was interested in hearing more about how results from a pseudo-standard setting for the technical tests might be used – with interest. The plan is comprehensive and thoughtful and reflects previous conversations with the DACMPT.

Recommendations: The DACMPT noted that many of the sources of evidence described in the evaluation plan are important sources of validity evidence as classified by the Standards for Educational and Psychological Testing. It would be useful to adopt the language and categories of evidence that are consistent with the Testing Standards for consistency and adherence to the standards. The DACMPT also requested additional information about all tests that currently comprise the battery (and noted that they are described in the briefing) versus all of the special tests that are being delivered currently on the platform.

CAT-ASVAB New Forms Update

Dr. Trippe gave a presentation on the development of new ASVAB pools 11 – 15. The presentation summarized efforts to identify enemy items and, once identified, to distribute them across the ASVAB test pools. The presentation also described the CAT simulations that support test form assembly. Dr. Trippe's presentation concluded with identification of tasks to be completed in preparation for final pool construction to support form administration.

Because of the large numbers of items in the development process, due in part to the aggressive item seeding/tryout schedule, a methodology is needed to identify items that cue others and therefore should not be delivered to a test taker in a single administration. Two strategies are being deployed to address this concern: a) a more automated method for identifying potentially enemy items and b) distribution of these enemy items across the item pools so they are not available for item selection during the use of a single pool. Two strategies were tried out for automatic item enemy flagging (content based judgments and a "hotspot" tool).

Simulations were used to explore the information functions derived from applying the item selection algorithms across the newly developed pools, using differing number of items per CAT administration and across 2 or 3 pools.. The results suggest that in the simulated forms, some subtests (notably Mathematical Knowledge) are comprised of more difficult items than were previously administered using the paper-and-pencil tests. Mitigation of this result is expected from infusing middle range difficulty items and additional newly developed items. Equating efforts are underway for pool 10, and are in preparation for new pools 11-15.

Comment: The DACMPT had no specific recommendations regarding this presentation and look forward to seeing updates on this process for enhancing the ASVAB items pools continues.

Mental Counters – Rapid Guessing Behavior & Think Aloud Plan

Dr. Yin provided two briefings on research related to the Mental Counters test. A long term problem with the Mental Counters test is that between 4 and 9 percent of the examinees fail to get any of the 32 items on the test correct. The distribution of scores for the remaining examinees is relatively normal. Dr. Yin described a careful and clever examination of the degree to which it is possible to identify those examinees who are not trying, hence get scores of 0. In examining response times to individual items and items in sequence, the research team was able to identify those who did not appear to be trying to answer the items correctly and remove them from the distribution of scores. Their removal resulted in a near normal distribution with the odd result that an unexpectedly high number of examinees now received a score of 1.

In a second report, Dr. Yin described a planned research effort in which examinees are asked to think aloud while taking the exam in an effort to identify why some examinees may fail to get any items correct. The DACMPT agreed that a first effort should be to modify the instructions so that it can be determined that all examinees actually understand the instructions. Shortened instructions along with a requirement that examinees get at least one practice item correct before proceeding to take the test may be effective, a practice that is used successfully in other testing programs.

Recommendation. Before investing in a think aloud experiment, the DACMPT recommends that an experiment be conducted in which efforts to modify instructions are evaluated. This modification would involve shortening the current instructions insofar as is possible and to provide instructions and practice that would ensure each examinee gets one practice item correct before taking the test to ensure that the instructions are understood.

CAT Cyber Test

Dr. Gao briefed the DACMPT on efforts to create a computerized adaptive version of the Cyber Test. In order to achieve this goal, analyses were needed to (1) identify whether the assessment is sufficiently unidimensional to support the item response theory model underlying the CAT and (2) how well the CAT performs under simulated conditions.

The unidimensionality analysis indicated that the assessment is sufficiently unidimensional to support its use for a CAT. An analysis of the item pool indicated there was an appropriate distribution of items in the pool in terms of content distribution and item parameter similarity to previously administered 29-item fixed forms. Simulations were run to help decide whether to implement a 2-pool or 3-pool CAT solution of varying lengths.

Recommendation: The DACMPT concurs with the recommendation of the MAPWG to transition at a future date to a 15-item CAT using the 3-pool solution. Even though the 3-pool solution with 15-item CAT showed lower precision at the cut than for the current 29-item fixed form, the decrease in precision was slight and by having 3 pools to draw from it is possible to have a back-up pool available as a reserve form. This slight decrease in precision at the cutscore will likely be mitigated when future items are brought into the pools that are either targeted at the cutscore or are more discriminating at low-to-moderate difficulty.

Adverse Impact for Special Tests

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

Recommendation: The DACMPT recommends continued monitoring of subgroup differences and adverse impact particularly for tests that are used with relatively high cutoff scores.

Device Evaluation

Dr. Fechter provided an update on the Device Evaluation Study that will investigate the feasibility of expanding the devices for ASVAB iCAT and PiCAT administrations. Following a review of the literature on the impact of test scores when administration is completed on mobile devices, a decision was made to limit the study to the consideration of specific Notebook, Tablet and Smart phone devices. Because these devices are products of different vendors and use multiple operating systems, there are limitations to the generalizability of the results beyond these specific devices.

Administration is currently on-going and will employ both recruits and applicants over a 6 month period involving several ASBAB subtests (GS, AR, WK, PC, MK, MC, and AO). Data analysis plans were presented with the intent to use MANOVA with equated subtest score and response time as the two dependent variables, as well as a separate analysis to address the impact of device familiarity on examinee performance. Analyses will also consider if there are item features (such as the inclusion of a graphic) that interact with examinee performance for the different devices.

Recommendation: The DACMPT is eager to see the results of the study. Consideration should be given to doing the full model MANOVA (including interaction terms) instead of looking first at the main effect of the model.

AVID Initial Evaluation

Dr. Kirkendall briefed the DACMPT on the development of the Adaptive Vocational Interest Diagnostic (AVID) through the U.S. Army Research Institute. The AVID is intended to provide recruits relevant information about the 140 entry-level specialties. The validity information was useful, but additional descriptive information would be helpful in interpreting validity coefficients. The outcomes used in the validity studies are appropriate and informative.

Recommendations: The DACMPT recommends that all presentations of validity coefficients include descriptive information, particularly means and variances. When validity coefficients are reported, it may be useful to report both uncorrected and corrected correlations (disattenuated for measurement error and when appropriate for range restriction). The AVID composite is currently optimally weighted to predict the composite outcome. Another approach would be to create a composite that is optimally weighted for the measurement information contained in the test components – from a confirmatory factor analysis model or some other measurement model. This would maximize the measurement information in the AVID, rather than rely on sample-specific regression coefficients when predicting a composite outcome (that may differ in variability from a larger generalizable population).

ASVAB Time Limits

Dr. Gao presented the results of a study to investigate whether there is sufficient time for examinees to complete the tests that are on the ASVAB platform by looking at actual examinees response time distributions. In this study, data from 2015 – 2018 across WinCAT and *i*CAT administrations were used. It should be noted that starting in 2015, an aggressive item seeding/pretest strategy was implemented on the WinCAT platform that resulted in additional items (and additional testing time) for test takers. This study also anticipated the implementation in the future of administrations of the ASVAB across multiple devices that might also have implications for test administration time.

Through use of inspecting the actual test score distributions for candidates who completed 95% and 99% percent of the items, two subtests appeared to demonstrate speededness with the 95% completion criterion (Mathematical Knowledge and Assembling Objects).

Recommendation: The DACMPT concurs with the recommendation to set time limits so that at least 99% of examinees are able to complete the test in the testing time allotment.

TAPAS Review

Dr. McGonigle provided an update on the evaluation of the TAPAS research by a special committee. The committee has met twice to consider a Rand report on TAPAS as well as reports by the Drasgow Consulting Group that developed TAPAS, and other users as well as the body of published papers and technical reports on TAPAS. They will meet two more times this spring and summer and expect to provide a report on their work by October.

Comment: The DACMPT agrees that a competent and well-respected committee has been convened to conduct this important review and looks forward to a summary of their report and findings at a future meeting.

Future Topics

As a final session, the DACMPT considered future topics for committee briefings.

Recommendation: Among the ongoing projects and regular updates, the DACMPT requests additional updates regarding the following topics:

- ASVAB validity framework
- Common performance metrics across services
- CEP expert panel recommendation updates
- Information on the content, purpose and uses of the Abstract Reasoning special test
- Cloud efforts, addressing the structure of the "cloud", security, and flexibility

Finally, the DACMPT requested that future meetings be planned a full year ahead, rather than the next half-year meeting. This would allow easier planning since many other technical committees plan a full year in advance.

The DACMPT agrees that the meeting was informative and useful. We continue to appreciate the high quality efforts of Accession Policy and DPAC staff, and the research staff of the services, as well as their frank interactions with the committee. We look forward to our next meeting.

Sincerely,

Michael C Rodriguez

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

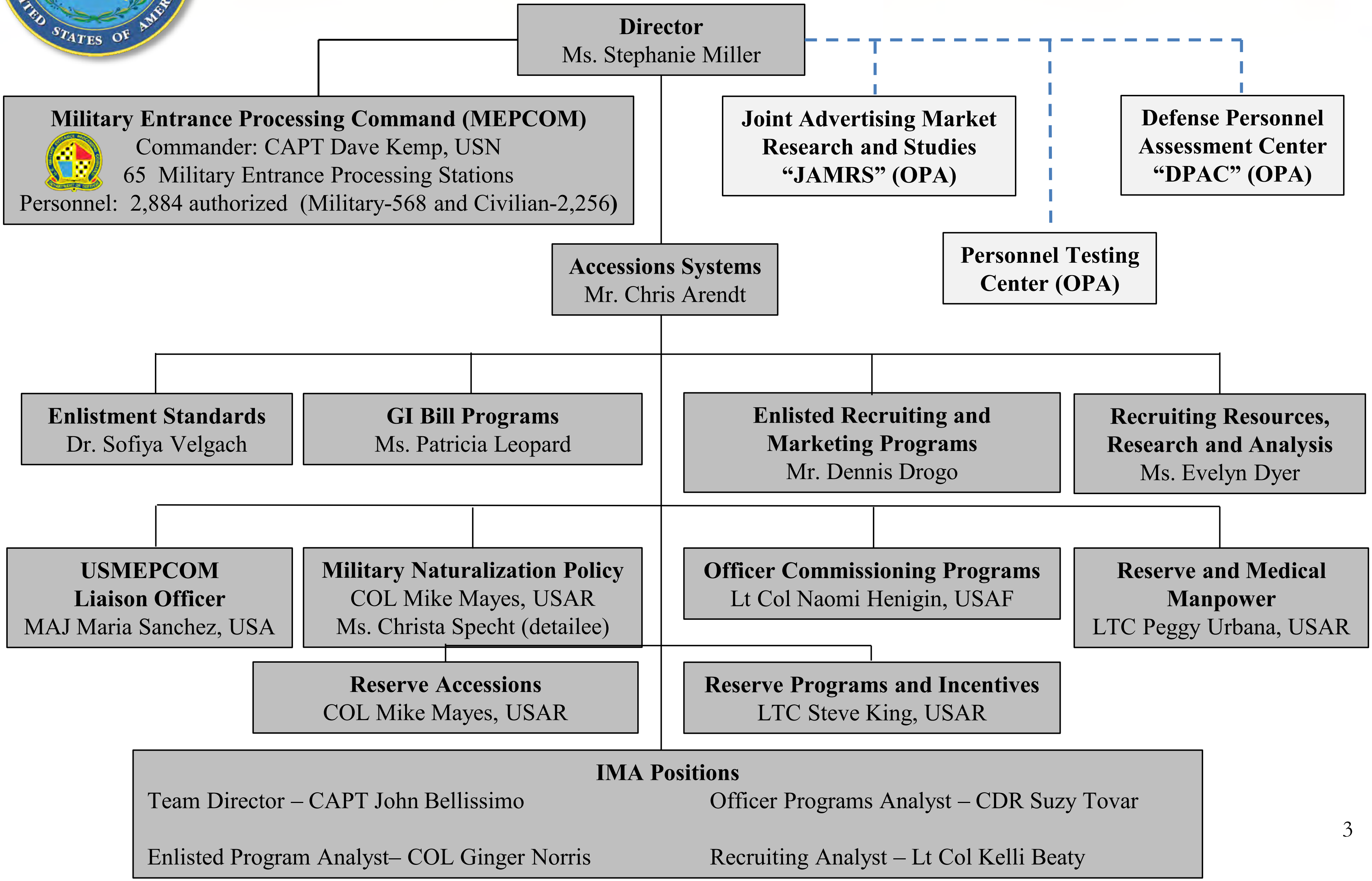
Military Personnel Policy
(Accession Policy)

Our Mission

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

"Stewards of the All-volunteer Force"

Accession Policy



Director
Ms. Stephanie Miller

Military Entrance Processing Command (MEPCOM)
Commander: CAPT Dave Kemp, USN
65 Military Entrance Processing Stations
Personnel: 2,884 authorized (Military-568 and Civilian-2,256)

Joint Advertising Market Research and Studies “JAMRS” (OPA)

Defense Personnel Assessment Center “DPAC” (OPA)

Personnel Testing Center (OPA)

Accessions Systems
Mr. Chris Arendt

Enlistment 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

Military Naturalization Policy
COL Mike Mayes, USAR
Ms. Christa Specht (detailee)

Officer Commissioning Programs
Lt Col Naomi Henigin, USAF

Reserve and Medical Manpower
LTC Peggy Urbana, USAR

Reserve Accessions
COL Mike Mayes, USAR

Reserve Programs and Incentives
LTC Steve King, USAR

IMA Positions
Team Director – CAPT John Bellissimo
Officer Programs Analyst – CDR Suzy Tovar
Enlisted Program Analyst– COL Ginger Norris
Recruiting Analyst – Lt Col Kelli Beaty

Fiscal Year 2019 Mission

Service	Goal
Army – Active, Guard, and Reserve	122,600
Navy – Active and Reserve	48,162
Marine Corps – Active and Reserve	41,370
Air Force - Active, Guard, and Reserve	47,132
DoD Total	259,264

Source: Services

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

Mission Attainment-February 2019

- Fiscal Year 2019 -	Active Recruiting/Accession Data				
	Annual Goal	FYTD Goal	FYTD Accessions	FYTD Percent of Goal	
Army	68,000	20,455	19,932	**97.44**	Y
Navy	40,000	15,614	15,682	**100.44**	G
Marine Corps	32,967	11,817	11,840	**100.19**	G
Air Force	32,300	13,239	13,356	**100.88**	G
Total	173,267	61,125	60,810	**99.48**	■

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	16,546	16,904	**102.16**	G
Army Reserve	15,600	6,105	6,264	**102.60**	G
Navy Reserve	8,162	3,227	2,979	**92.31**	Y
Marine Corps Reserve	8,403	3,209	3,578	**111.50**	G
Air National Guard	9,422	3,621	3,771	**104.14**	G
Air Force Reserve	5,410	2,937	3,937	**100.00**	G
Department of Defense (Total)	85,997	35,645	36,433	**102.21**	■

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	**92.3**	G	**59.7**	Y	**3.38**	G
Navy	**97.4**	G	**71.2**	G	**0**	G
Marine Corps	**99.2**	G	**69.2**	G	**0**	G
Air Force	**98.4**	G	**82.3**	G	**0**	G
Reserve Components						
Army National Guard	**97.5**	G	**64.1**	G	**3.30**	G
Army Reserve	**97.1**	G	**65.1**	G	**0.67**	G
Navy Reserve	**96.5**	G	**73.5**	G	**0.0**	G
Marine Corps Reserve	**99.2**	G	**72.9**	G	**0.0**	G
Air National Guard	**99.7**	G	**76.7**	G	**0.0**	G
Air Force Reserve	**99.6**	G	**75.4**	G	**0.0**	G

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

OPA

OFFICE OF PEOPLE ANALYTICS

State of the Recruiting Market

March 2019

JAMRS

FOR OFFICIAL USE ONLY

BLUF: State of the Youth Market

- **The youth market is disconnected from today's Military, resulting in few youth being interested in or considering military service.**
  - The declining veteran population and shrinking military footprint has contributed to a market that is unfamiliar with military service.

- **A lack of familiarity with the Military leads to youth relying on stereotypes of what they think life is like in the Military.**
  - For youth, the risks of service are top of mind, and they don't see the inherent value of the benefits and opportunities afforded by serving.

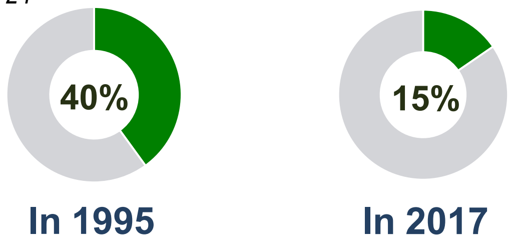
- **Outreach efforts must be deliberate, sustained, and relevant to targeted markets in order to move youth, and their influencers, beyond their preconceived notions of military service.**
  - Messaging that propensed youth are receptive towards will not have the same impact on non-propensed youth.
  - Influencers are largely unfamiliar with the benefits of service. Those that consider themselves knowledgeable about the U.S. Military are more likely to recommend and support youth for military service. Targeted outreach to key influencers should build awareness and advocacy for service.

Youth Market: Awareness and Knowledge

Proportion of Youth with a Parent Who Served

YATS (1995) and Youth Poll (Fall 2017)

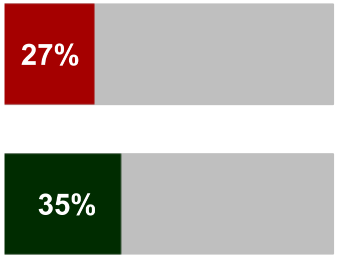
Youth ages 16–24



Awareness

Military Ad Tracking Reserve Study (April–June 2018)

Young adults ages 17–35



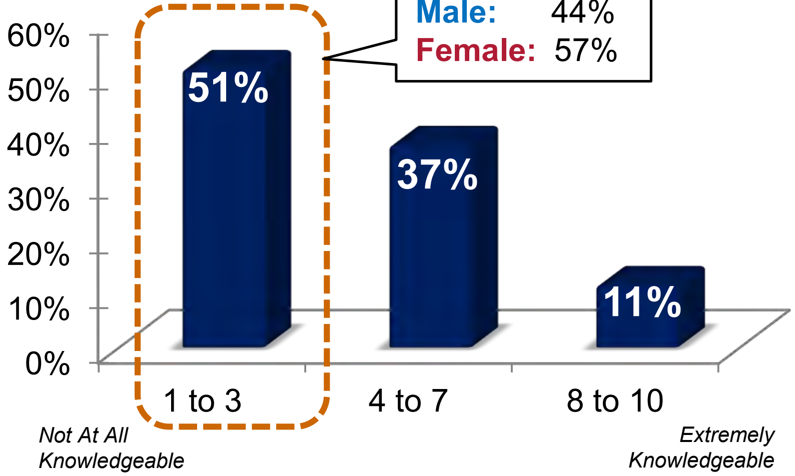
…can name **all five** active duty Services.

…**do not know** there is a difference between an Officer and an enlisted person.

Self-Reported Knowledge of Active Duty Service

Military Ad Tracking Reserve Study (April–June 2018)

Young adults ages 17–35



Male: 44%
Female: 57%

"[Service members] don't get to plan out anything because their whole plan could change right away."

"Pets aren't allowed on base…right? Are you allowed to have a dog in the Military?"

"Is there a penalty for getting pregnant (in the Military)?"

"There's also people that come out and they don't know what to do in the real world anymore."

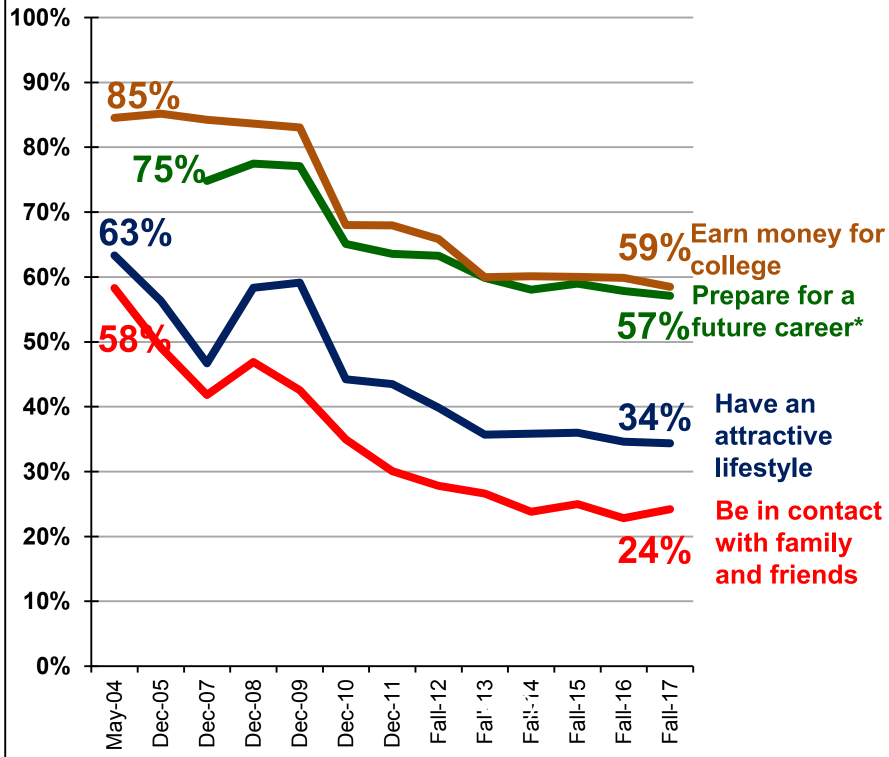
Many young adults lack close family ties to the Military and basic knowledge about the Military. What they think they know is often wrong.

Youth Market: Perceptions of the Military

How likely is it that joining the U.S. Military would allow you to…

Youth Poll (2004 to 2017)

Youth ages 16–21
% Responding 5, 6, 7 "Extremely likely"

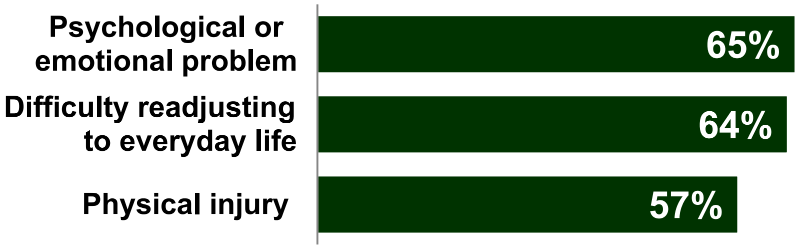


*"Prepare for a future career" began tracking in Dec 07.

How likely do you think it is that someone getting out of the Military will have…

Military Ad Tracking Study (April–June 2018)

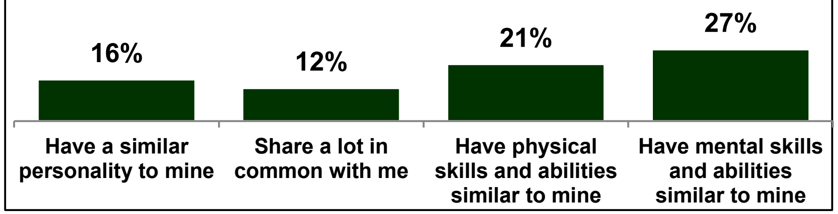
Youth ages 16–24
% Responding Likely/Very Likely



People in the U.S. Military…

Youth Poll (Spring 2017)

Youth ages 16 to 21
% Responding "Agree/Strongly Agree"

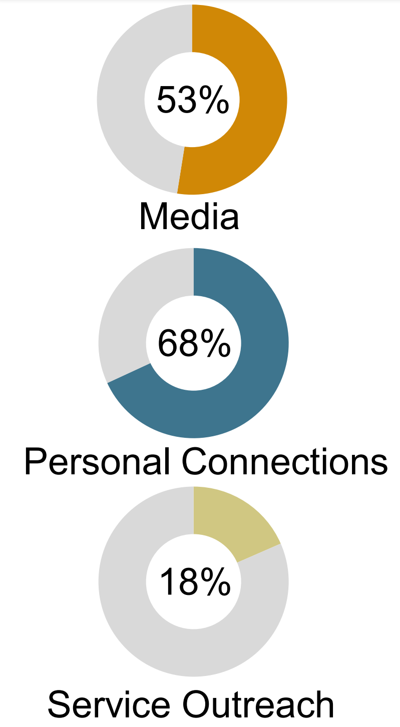


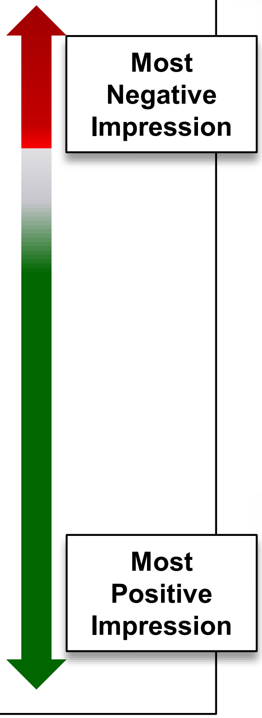
Youth are more familiar with the risks than the benefits of serving and do not relate to people who join the Military.

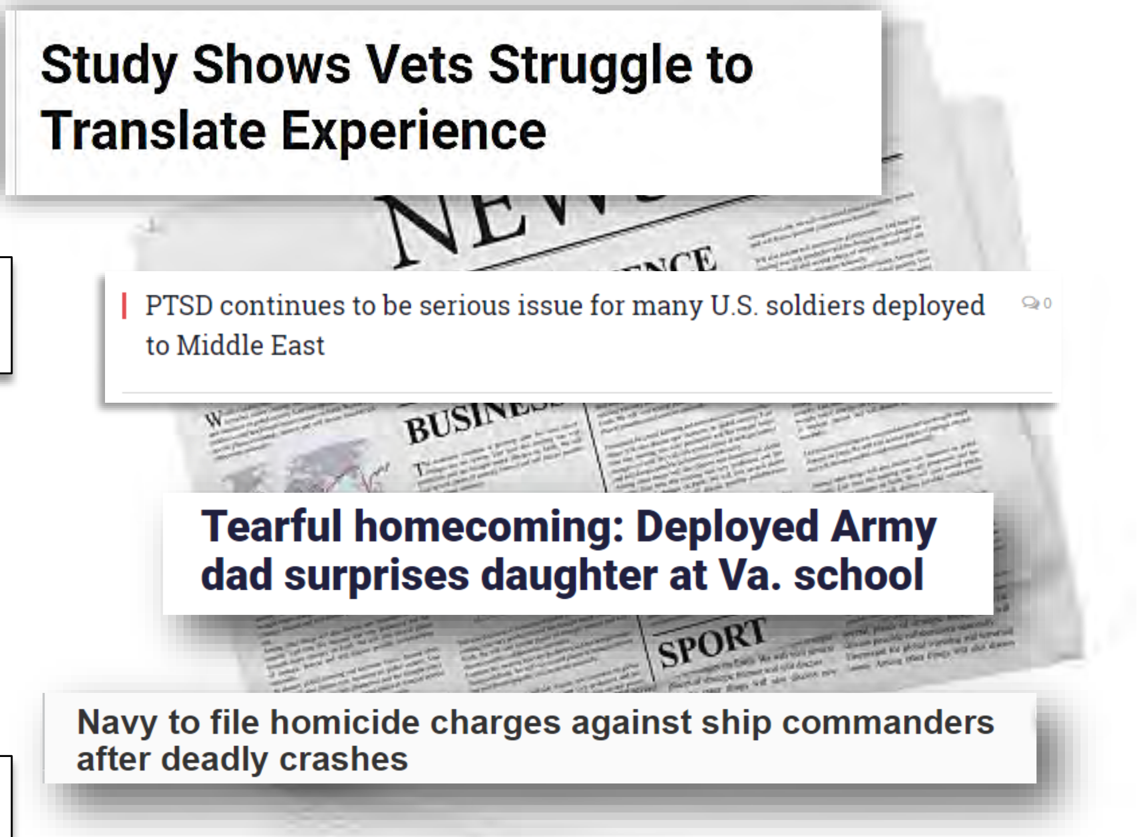
Sources of Impressions of the Military

From what people or sources of information have you gotten the majority of your impressions about the Military?

Military Ad Tracking Reserve Study (April–June 2018)







Study Shows Vets Struggle to Translate Experience

PTSD continues to be serious issue for many U.S. soldiers deployed to Middle East

Tearful homecoming: Deployed Army dad surprises daughter at Va. school

Navy to file homicide charges against ship commanders after deadly crashes

Most of the narrative in youths’ environment is not controlled by the DoD and disproportionately focuses on sacrifice.

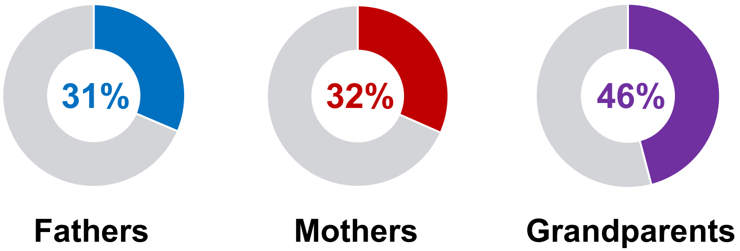
Note: Young adults ages 17–35.

Influencer Market: Social Norms for Service

History of Service

Military Ad Tracking Influencer Study (April–June 2018)

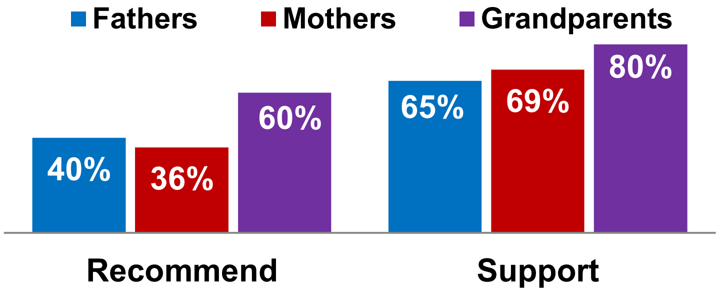
Influencers of youth ages 12–21;
% Responding they or a household member have served in the Military



Likelihood to Recommend and Support Decision to Join

Military Ad Tracking Influencer Study (April–June 2018)

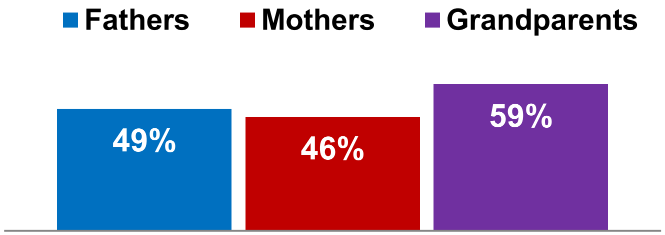
Influencers of youth ages 12–21;
% Responding Likely/Very likely and Agree/Strongly agree



How much would you feel pride if [relation] wanted to join the Military?

Military Ad Tracking Influencer Study (Oct–Dec 2017)

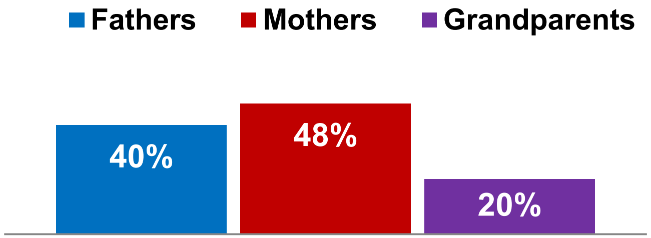
Influencers of youth ages 12–21; % Responding A lot/A great deal



I believe there are positive aspects of the Military, but the negatives outweigh the positives in my opinion for [relation].

Military Ad Tracking Influencer Study (April–June 2018)

Influencers of youth ages 12–21; % Responding Agree/Strongly agree



Parents are less positive about military service than grandparents, but parents are also less connected to the Military.

Questions?

Tab E

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB Career Exploration Program

February 2019

Agenda

- CEP Metrics
- CEP IPR (Accession Policy, MEPCOM , and OPA)
- Needs Assessment and Recommendations
- State Usage of ASVAB CEP
- PTI Proficiency Training
- FYI Revisions

ASVAB CEP 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**	**559,375**

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**	**10,490**	**45.7%**

**School year runs from July 1- June 30. Data as of 1 February, 2019.

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Year-to-Date**

Paper and Pencil Numbers

	Examinees 17-18	Examinees 18-19
TOTAL	662,564	513,236**

CEP iCAT Numbers

	Examinees 17-18	Examinees 18-19
TOTAL	51,213	46,139**

**Total students as of 1 February, 2019.

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**	**6,356**	**3,138**	**2,818**	**3,025**	**244**	**15,581**

**School year runs from July 1- June 30. Final numbers (as of 1 February, 2019).

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

CEP IPR

Preliminary Recommendations for Elimination of P&P

COA 1: DPAC develop an access code process valid for a specific period of time or test window, similarly to current Single Site Testing procedures. A school would be provided the code to access the CEP iCAT and be able to test their students on their schedule based on the availability of their computer lab(s) within the assigned test window. The schools would be able to test without a MEPS TA/ITA present.

DRAWBACK: Increased test administration support required from school personnel

COA 2: USMEPCOM purchase tablet computers for each MEPS to make available to test administrators when needed for use at a school. A suitable number of tablets would be determined and issued to each MEPS. When a school desired to test and did not have the right equipment to test, or no computers at all, the tablet computers would be issued to a TA in order to conduct the test. This would require a version of the CEP iCAT compatible with tablets. DPAC would need to develop alternate interfaces and a new test delivery application to run on tablets of interest. DPAC would need to conduct a study to evaluate whether there are any psychometric issues such as performance differences across the different tablets. This work is underway at DPAC and projected for completion in Sept 2019.

DRAWBACK: Increased costs associated with the purchase and maintenance of tablets and associated security and test delivery software.

COA 3: USMEPCOM purchase and configure a government laptop as a server or hotspot to provide the test via local network or wifi to school tablets and or student phones, by developing a CEP "app" that the student could take the CEP on their phone in the school setting. This would require programming and test development by both USMEPCOM and DPAC. DPAC would have to be consulted regarding feasibility as well as the time and effort required to complete the programming and new test versions. Additionally, DPAC would need to conduct a study to evaluate whether there are any psychometric issues with this option.

DRAWBACK: Increased costs associated with developing an iCAT application that runs on existing student/school devices

COA 4: Administer the P&P CEP test as a "low stakes" test where applicants would take a verification test at the MEPS if the results are used for enlistment. This would allow continued use of the P&P test. DPAC would need to develop a verification test process similar to that used for PiCAT, so DPAC would need to determine feasibility and level of effort. This option would require MAPWG approval prior to conducting any level of effort analysis.

DRAWBACK: Discontinued use of CEP P&P ASVAB scores for operational enlistment purposes (negatively impact accessions)

COA 5: Develop a test version which could be administered via DVD. This would allow a MEPS to send an appropriate number of disks to a school to have them plug the disk into their computers and administer the test. This option would require DPAC programming to administer the test via a disk. DPAC would have to be consulted regarding feasibility as well as the time and effort required to complete the programming. DPAC would need to conduct a study to evaluate whether there are any psychometric issues with this option.

DRAWBACK: Security and cost implications of developing a stand-alone (non-internet) version of CAT-ASVAB for use in CEP.

Conclusions and Next Steps:

- P&P-ASVAB testing should continue in the CEP for schools that lack sufficient infrastructure (i.e., computers and internet connectivity)
- Backup P&P forms should be identified for use in the CEP in the event of a compromise
  - Identify form/s from P&P – ASVAB 20 – 22 series
  - Tiger Team with DPAC, MEPCOM, and AP to develop PO&AM for implementation
- Where possible, steps should be taken to increase utilization of iCAT
  - DPAC has identified DMDC team leads to address monitoring issues
  - DMDC assigned POCs to resolve connectivity issues
  - Requested RAM increase
  - Requested serves increase
  - DPAC continues to work with DMDC to create systematic plan to accommodate approximately 2 million additional users (transition of WinCAT and CEP P&P to iCAT).
  - Ultimate transition from iCAT to Cloud
- Conducting Device Evaluation Study

Conclusions and Next Steps:

- Allow iCAT administration using 12” monitors
- Expand iCAT browser options to Safari
  - DPAC developing implementation timeline
- Continue to work with CAVEON to identify potentially compromised material
  - The largest online arena of potential ASVAB threats was from free and for profit test prep sites
  - Other areas where reported potential threats occurred were Video Archives, Flashcard Sites, Social Media, Mobile Apps, Document Archives and Brain dump Sites.
  - Provide additional test item content to CAVEON for review

Needs Assessment and Action Items

Actions Completed:

- Reviewed past reports on pilot efforts to incorporate iCAT into the CEP and observation forms provided by test administrators
- Conducted observations of current paper-and-pencil and CEP iCAT sessions
- Obtained input from DPAC and MEPCOM personnel, ESS, and others in the field on the status of iCAT CEP
- Made recommendations for improving the current situation and provide alternative business models for iCAT CEP

Recommendations:

- Paper-and-Pencil and iCAT Administrations
  - Reinforce rules related to proper proctor behavior
  - Institute a database management system to issue session numbers
- Paper-and-Pencil ASVAB Administrations
  - Done in accordance with recommended procedures
  - Review instructions with the aim of making them more succinct, eliminate repetition

Recommendations, cont:

- iCAT Administrations
  - Eliminate data fields in test that are not routinely completed by students (e.g., address, avowal of physical fitness)
  - Identify method to circumvent issues with students needing to reenter information exactly as first entered when there is a need to log-in a second time
  - Standardize log-in procedures
  - Streamline log-in procedures for test administrators
  - Allow TAs to reset passwords rather than having to call Help Desk
  - Address bandwidth issues
    - Stress the requirement that school IT personnel be consulted during scheduling to ensure sufficient bandwidth and minimal conflicts
  - Stress administration protocols
  - Ensure that Help Desk support is available during testing and that server maintenance doesn't occur during testing
  - Institute a nationwide scheduling system so Help Desk and IT personnel know when testing is occurring and where
  - Expand browsers through which iCAT can be accessed
  - Address issues regarding screen resolution

Recommendations, cont:

- Score reporting
  - Investigate possibility of allowing ESS to access score reports with user name and password vs CAC
    - Opens up possibility of same-day testing and interpretation
    - Same-day PTIs would reduce ESS scheduling and travel burden
  - Address issues reported by ESS with printing score reports
- Post-Test Interpretations (PTI)
  - Standardize PTIs – will ensure students receive the most relevant and accurate information
  - Focus on score reports, using ASVAB and FYI scores to explore careers, and accessing information in the future
  - Employ registration feature that ties student's email to their access code, so only email and password are needed to access sites
  - Include PTIs in performance metrics of MEPS/ESS, rather than just number of students tested
  - Consider alternate methods of providing PTIs
    - Train school personnel (e.g., counselors, teachers)
    - Develop online, interactive modules to guide students through the process

State Usage of ASVAB CEP

Current list of States and ASVAB CEP:

Four states have legislation requiring them to provide CEP state wide:

- Texas
- Indiana
- Arizona
- Utah

16 states have legislation which impacts certain parts of the State, either by district or school type (e.g. Career Tech Ed)
Arkansas, California, Colorado, Georgia, Kansas, Maryland, Missouri, New Jersey, New York, Nevada, Oregon, San Juan, Tennessee, Vermont, Washington, West Virginia,

Mississippi and Kentucky are considering legislation (2/14/2019)

State Usage of ASVAB (Cont.)

- As of August 2018, 12 states use ASVAB CEP in some capacity as a career exploration tool
  - Colorado
  - Indiana
  - Kentucky
  - Maine
  - Mississippi
  - Missouri
  - Nevada
  - New Jersey
  - Texas
  - Virginia
  - West Virginia
  - Wyoming
- Some states (IN, KY, ME, TX) mention ASVAB CEP specifically
  - Often as part of graduation or college/career readiness track
- Other states fail to mention CEP, limiting discussion to ASVAB
  - Typically in terms of requiring a minimums score on the “ASVAB” (likely the AFQT, given that cutoffs are almost certainly percentile scores)

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

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

PTI Proficiency Training

Program Initiative: PTI Proficiency Training

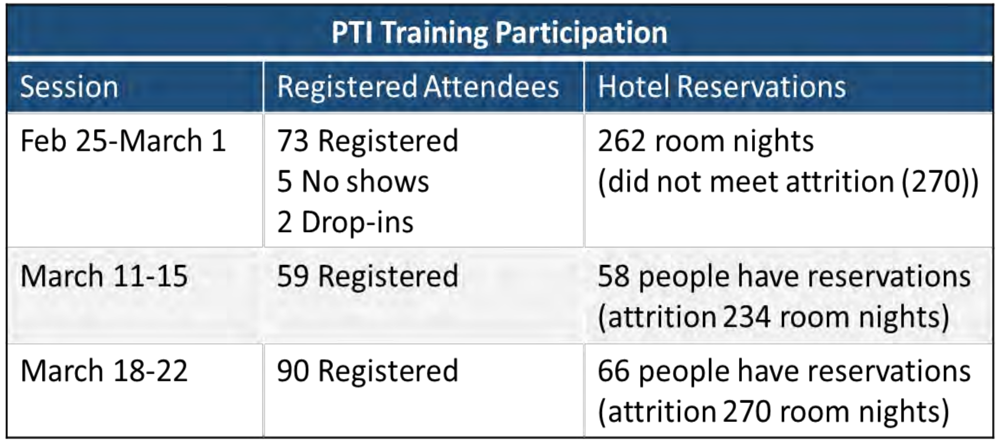
Goal: Standardize the process by which post-test interpretation (PTI) sessions are conducted

Purpose: Address Expert Panel Recommendations to orient all attendees to the ASVAB CEP enhancements, and help attendees learn the strategic purposes of collaborating with others operating within their territory to achieve missions

Stakeholder Involvement: OPA, US MEPCOM, Army, Navy, Army National Guard, Air Force Reserves, Coast Guard

In progress: Established process for becoming proficient. Virtual training “pre-requisite” deployed to attendees 25 JAN via Moodle; topics include:

- ASVAB Measurement, Data and Use
- Interpreting and Discussing ASVAB Scores
- ASVAB CEP Components
- Conducting a Post Test Interpretation
- Becoming PTI Proficient



PTI Training Participation

Session	Registered Attendees	Hotel Reservations
Feb 25-March 1	73 Registered 5 No shows 2 Drop-ins	262 room nights (did not meet attrition (270))
March 11-15	59 Registered	58 people have reservations (attrition 234 room nights)
March 18-22	90 Registered	66 people have reservations (attrition 270 room nights)

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 ESS (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

ASVAB Score Interpreting

The Understanding Your ASVAB Results video should be shown at the beginning

Military Line Scores*

Passed □

Notes:

OCCU-Find Orientation

The Exploring the OCCU-Find video should be shown at the beginning of this section and can be used to guide the instruction.

ASVAB CEP Post-Test Interpretation Session
PTI Proficiency Evaluation Metric

Instructions: For each of the items below, choose from these options:

- **Excellent**
- **Above Average**
- **Average**
- **Below Average**
- **Needs Improvement**

Descriptions are given for performance that is Excellent, Average, and Needs Improvement for each item. Scores of Above Average and Below Average are to be given for execution that shows some, but not all, of the described practices.

In order to be considered proficient, personnel must receive a score of Average (3) or higher on all of the required components, as well as a score of Average (3) or higher on 3 complementary components. Required components have been marked with an asterisk and are shaded in purple. Complementary components have been shaded in teal. Next to each section, put a checkmark if the individual received a score of 3 or higher on that section. At the bottom of each page, add up the checks you have for the required and non-required components.

Introduction

Establishing Expectations* Passed □

Notes:

○ 5 **Excellent:** Explained directions, procedures, and intended activity outcomes. Asked students about their post-high school plans to guide PTI session. Informed students that military service is not required. Checked for understanding.

○ 4 **Above Average**

○ 3 **Average:** Presented agenda and session information. Informed students that military service is not required.

○ 2 **Below Average**

○ 1 **Needs Improvement:** Began first activity without providing an overv

Website Log-In*

○ 5 **Excellent:** Ensured that al students with

Virtual Training Consists of:

- Log in with username and password
- Objectives
- Learning Goals
- Multimedia Content: including videos, print materials, social media, etc
- Concept Checks
- Reflection and Application Activities
- An area to upload supporting documentation
- An area to view all people who are proficient at giving PTIs, regardless of job function or affiliation
- An area to assign access codes that are pre-populated with scores and can be used for a long period of time
- A communication system for all people who are conducting PTIs across the US.
- An ability to collect information about training needs

Comments From Training Evaluations

- "This entire program blows my mind. From the quality of the paper to the useful content it is evident those running this program know what is needed to improve actions in the field."
- "The structure was great. It gave some very valuable information to a very diverse audience. It was nice to place a face with some of the organizations we conduct business with.“
- "The new matrix and checklist are the best. It makes life so much easier for training others to conduct the PTI and keeps consistency across the US.“
- "As a recruiter, I did not understand some of the analytics or processes/roles ESS, TC, etc . It was good to have the mixed groups to help me understand that side of this program."
- "New information showing both MEPCOM and recruiting ESS they need to work together for the program to succeed."
- "Thank you to all who have worked above and beyond to accomplish this mission. Great teamwork."
- "I like the most of all how it was conveyed that we were going to be ALL EQUAL for the week."
- "The virtual training was wonderful because I wasn't aware of all the PTI had to teach.“
- "I liked sharing the content and having discourse with so many different viewpoints based on position and service."
- "This training was wonderful but to be 100% effective, having Service leadership attend would be great. Having recruiting leadership here to see how PTIs are supposed to happen would alleviate some issues (proctor no shows and recruiters scheduling PTIs without MEPS knowledge.)“
- "Class exceeded expectations. Facilitated a learning environment."
- "Training was phenomenal. Best I have received in my last 5 years. I can tell the thought and strategic logistics, planning, and overall group effort.“
- "Great resources from other attendees. Networking provided insight into other geographical areas, best practices. Great opportunity and training.“
- "Outstanding training. Relaxed, professional, and complimentary! Enjoyed the professionalism of the staff and attendees. Kudos to the staff and all that assisted with coordinating this course.“
- “No more homework!”

Expert Panel: FYI Revisions

Expert Panel Members: Jim Rounds, Patrick Rottinghaus,
Contract Project Manager: Rod McCloy,
Government: Dan Segall, Marry Pommerich, Shannon Salyer

Background

- OPA convened an ASVAB CEP Expert Review in 2017 to comment on revisions to CEP and its measures, ASVAB and FYI
- Review recommended thorough revision of the FYI
  - Item revisions and additional items are required to assess a broader array of basic interests (e.g., information technology, leadership, health services)
    - Most commercial interest inventories are evaluated/revised routinely, typically every 5 to 7 years
- Review recommended an 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
- Review recommended integrating basic interest scales into the CEP
  - Research suggests that basic scales provide . . .
    - a richer interpretation of interests than Holland types do (Day & Rounds, 1997; Gasser, Larson, & Borgen, 2007; Liao, Armstrong, & Rounds, 2008; Ralston, Borgen, Rottinghaus, & Donnay, 2004; Su et al., 2018)
    - more transparent linkages between a person's interests and the interests and activities associated with work environments

Background (cont.)

- The review discussed the need to revise the FYI on a routine basis to ensure item content is relevant, and to consider incorporating advances in interest measurement (e.g., computerized adaptive testing)
    - Process would involve several steps, including (a) a content analysis of the items to determine the extent to which the Holland types are covered and (b) exploration of potentially new content domains critical to both the current occupational landscape and overall relevance to the target population of students using the ASVAB CEP.
- The review discussed the need for more precise measures of basic interest domains (e.g., science, mathematics, mechanical activities, healthcare, public speaking), which are analogous to those offered in the Strong Interest Inventory (SII; Donnay, Morris, Shaubhut, & Thompson, 2005) and could supplement the existing Holland measures (Day & Rounds, 1997; Ralston et al., 2004)
- In addition to an expert review of the FYI items, existing data could be used to conduct an item analysis, including exploration of differential item functioning (DIF) and an investigation of the factor structure to inform the revision process.
- Review members offered differing opinions on whether to change the Like-Indifferent-Dislike (LID) format in favor of a 5-point scale in future revisions
    - Different scale options could be evaluated by focus groups, and data from pilot studies could inform decisions on selecting the best response scale for the revised FYI.

Tasks Conducted in Initial Research on FYI Revision

I. Obtain a representative sample of existing FYI data.

II. Conduct a preliminary expert review of existing FYI items. Prior to expert review, conduct descriptive item analyses (M, SD, and item response distributions) by sex and race. These data are to be used in item reviews.

III. Conduct a structural analysis of the RIASEC scales by gender using multidimensional scaling, randomization tests (Rounds, Tracey, & Hubert, 1992), and circular unidimensional scaling (Armstrong, Hubert, & Rounds, 2003).

IV. Conduct a structural analysis of the RIASEC scales by race/ethnicity using multidimensional scaling, randomization tests (Rounds et al., 1992), and circular unidimensional scaling (Armstrong et al., 2003).

V. Conduct a content analysis of the current FYI RIASEC scales to identify facets (basic interests). Each RIASEC scale (15 items) will be analyzed according to the RIASEC basic interest RIASEC classification proposed by Su et al. (2018) to identify item coverage.

VI. Recommend a set of new basic interest scales and other specialized scales (e.g., work styles) informed by the previous item content analyses and structural analyses of the FYI; identify potential new target domains suggested by the 2017 ASVAB CEP Panel Report (e.g., Information Technology, Leadership, Public Speaking, and others included in the U.S. Department of Education's States' Career Cluster Initiative [e.g., STEM, Health Sciences]).

VII. Make final recommendations for further development of a Holland-based RIASEC measure.

Sample

- Data were collected from a national sample of 384,391 students who completed the FYI as part of participating in the ASVAB CEP
  - 177,994 (46.3%) females
    206,397 (53.7%) males
  - 71,919 (18.7%) 10th-grade students
    232,216 (60.4%) 11th-grade students
    79,257 (20.6%) 12th-grade students
    864 (0.2%) post-graduate individuals
  - 193,897 (74.8%) White
    29,183 (11.3%) Hispanic
    18,880 (7.3%) African American,
    10,616 (4.1%) Asian/Pacific Islander,
    6,477 (2.5%) Native American (1.7%), and
    47,608 (12.4%) who identified as multi-ethnic
    77,730 (20.2%) participants did not report race/ethnicity.

Results by Sex

- For both females and males, the RIASEC scales showed strong internal consistency
  - Cronbach’s alpha= .90 to 94 (females), .90 to .95 (males)
- Sex differences across the RIASEC scales
  - Males scored higher than females on the Realistic (*d* = -0.98) and Investigative (*d* = -0.23) dimensions
  - Females scored higher than males on the Social (*d* = 0.77) and Artistic (*d* = 0.38) scales.
  - There were no significant gender differences on the Enterprising and Conventional scales.
- Structural analyses
  - Current FYI items poorly fit the RIASEC model for males
    - A likely reason is that the item selection procedure that mirrored interrelations among the RIASEC types was not applied to a male sample
    - A revision of the item pool should use the same procedures with males as was used with females to select items
    - It is also important to include all racial-ethnic groups in the item selection process, including White students.

Next Steps

- DPAC to review recommendations and discuss a way ahead for the update of the FYI.
- Will keep MAPWG and DACMPT informed of these efforts.

Shannon Salyer, Ph.D.

Shannon.d.salyer.civ@mail.mil

Tab F

Major ASVAB R&D Efforts

Milestones and Project Schedules

Mary Pommerich

Briefing presented to the DAC

Carmel, CA

March 2019

Projects

- **ASVAB Development**
  - New CAT-ASVAB Item Pools*
  - Developing New CAT Item Pool for CEP
  - Automating Generation of WK Items†/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
- **AFQT Predictor Test** †
- **Air Force Compatibility Assessment**
- **Defense Language Aptitude Battery**

*Will be presented/discussed at this meeting.

† Moved to completed projects

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: May 2020

- **Subtasks**
  - Write items ✓
  - Pretest items (Summer 2018) ✓
  - Calibrate and scale items (Summer 2018) ✓
  - Conduct item screenings (Feb 2019)
  - Identify item enemies (Mar 2019)
  - Complete preliminary/final form assembly (Apr 2019–May 2019)

New CAT-ASVAB Item Pools (continued)

- **Subtasks** (continued)
  - Modify, test, and deliver CAT-ASVAB software and item pools to MEPCOM (Jun 2019–Jul 2019)
  - Collect and analyze IOT&E data (Aug 2019–Mar 2020)
  - Implement operationally in WinCAT and iCAT (Apr 2020–May 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

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 2018
- **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 (Spring 2019)
- **Successors**
  - Implementation of new CAT pool for CEP *i*CAT

† Dates impacted by DMDC Cyber Hardening Initiative

†† Dates are dependent upon MEPCOM's QA and deployment schedule

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**
  - Sep 2019
- **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 (Feb 2019–May 2019)
  - Assess MK item quality and parameter accuracy (Jun 2019–Jul 2019)
  - Conduct AR data collection (Jun 2019 – Sep 2019)
  - Assess AR item quality and parameter accuracy (Oct 2019–Jan 2020)
  - Provide final generator, interface, and documentation (Apr 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 (Spring 2019)
  - CAT-ASVAB Pool 10 for CEP *i*CAT (Spring 2019)
  - CAT-ASVAB Pools 11–15 (Summer 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 (in progress)

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)
  - Conduct predictor and criterion data analyses (TBD)
  - Examine projected impact of operational use of MCt scores for selected jobs (Summer 2019)
- **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)
    - 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) ✓
      - 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 (in progress)
      - Document critical aptitudes for cyber jobs
      - Summarize literature & recommendations on how serious gaming could be used to enhance assessment of cyber aptitude
    - 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 (TBD)
    - 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 (in progress)
    - Phase 5: Combine developed and validated measures into one cohesive computer-administered self-scoring test (TBD)
    - Phase 6: Validate to cyber populations (TBD)

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 (TBD)
    - 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)
    - TAPAS pilot testing of Marine Corps officers using paper & pencil (FY17–ongoing)
    - TAPAS pilot testing of Marine Corps applicants on the ASVAB platform (FY15–FY18) ✓
    - TAPAS operational administration for Marine Corps applicants (FY18–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)
  - 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 (Q3, 2019)
    - Proof of Concept for gaming environment vice self report format (Q4, 2019)

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 (Spring 2019)
  - Deploy in production environment (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) ✓
  - Identify administration platform in lieu of WinCAT† (TBD)

† WinCAT is slated to be decommissioned in March 2020.

Expanding Test Availability: Web/Cloud Delivery of 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 (TBD)
  - Develop cloud infrastructure (TBD)
- **Subtasks**
  - Identify requirements and design transition (Jan 2018–Sep 2018) ✓
  - Migrate Test 1 to DMDC web-based platform (Oct 2018–Mar 2019)[†]
  - Modify iCAT software to accommodate special tests (Oct 2018–Mar 2019)
  - Modify iCAT-A&R software to accommodate special tests (Oct 2018–Mar 2019)

[†] Test 1 is tentatively slated to be the Cyber Test.

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

- **Subtasks** (continued)
  - Develop web service for transferring scores to MEPCOM (Oct 2018–Apr 2019)†
  - Migrate TAPAS to the cloud platform (Feb 2019–Mar 2020)††
  - QA Test 1 on DMDC web platform (Apr 2019–Jun 2019)
  - Deploy Test 1 to Production on DMDC web platform (Jul 2019–Jul 2019)
  - Migrate Tests 2 and 3 to DMDC web platform (Apr 2019–Sep 2019)†††
  - QA Tests 2 and 3 on DMDC web platform (Oct 2019–Dec 2019)
  - Deploy Tests 2 and 3 to Production on DMDC web platform (Jan 2020–Jan 2020)
  - Migrate iCAT and iCAT-A&R to the cloud, including Tests 1-3, and QA (Jul 2019–Mar 2020)

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

†† 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 2 and 3 are tentatively slated to be AFCA and Coding Speed.

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

- **Subtasks** (continued)
  - Deploy iCAT & iCAT-A&R to Production in the cloud (Mar 2020–Mar 2020)
  - Deploy TAPAS to Production in the cloud (Mar 2020–Mar 2020)
  - Decommission WinCAT (Mar 2020–Mar 2020)
  - Migrate Tests 4 and 5 to the cloud platform (Apr 2020–Sep 2020)†
  - QA Tests 4 and 5 on the cloud platform (Oct 2020–Dec 2020)
  - Deploy 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)

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

Expanding Test Availability: Moving to the Cloud

- **Objective**
  - Examine the feasibility of moving test delivery to the cloud
- **Projected Completion**
  - Dec 2021
- **Predecessors**
  - Cyber hardening and code modernization (TBD)
  - Web delivery of special tests (TBD)
- **Subtasks**
  - Develop a business case analysis (Oct 2016) ✓
  - Assess cloud hosting options (Mar 2017) ✓
  - Obtain internal approvals (Spring 2017) ✓
  - Develop cloud infrastructure (Summer 2018) ✓
  - Test cloud infrastructure (Ongoing)
  - Submit package for IATT (Interim Authority To Test) (Aug 2018) ✓
  - Obtain IATT (Sep 2018) ✓
  - Conduct initial gap analysis on iCAT-A&R for cloud compatibility (Aug 2018–Oct 2018) ✓

Expanding Test Availability: Moving to the Cloud (continued)

- **Subtasks (continued)**
  - Conduct initial gap analysis on iCAT suite for cloud compatibility (TBD)
  - Obtain ATO (May 2019)[†]
  - Migrate TAPAS to the cloud platform (Feb 2019–Mar 2020)
  - Deploy TAPAS to Production in the cloud (Mar 2020–Mar 2020)
  - Migrate iCAT and iCAT-A&R to the cloud, including Tests 1-3, and QA (Jul 2019–Mar 2020)[††]
  - Deploy iCAT & iCAT-A&R to Production in the cloud (Mar 2020–Mar 2020)
  - Migrate Tests 4 and 5 to the cloud platform (Apr 2020–Sep 2020)[††]
  - QA Tests 4 and 5 on the cloud platform (Oct 2020–Dec 2020)
  - Transition DLAB2 from WDLPT to special test environment (Jan 2021–Dec 2021)

[†] The IATT is good for 6 months. Obtaining an ATO is dependent on the gap analysis and testing outcomes; as such, this date could shift.

[††] Tests 1-5 are tentatively slated to be (1) Cyber Test, (2) AFCA, (3) CS, (4) Mental Counters, and (5) Abstract Reasoning.

Appendix A
Completed Projects

Automating Generation of Word Knowledge Items

- **Objective**
  - Develop procedures for automating Word Knowledge (WK) item generation so that WK item pools can be replaced on a frequent basis
- **Projected Completion**
  - Sep 2018
- **Subtasks**
  - Develop Statement of Work and Independent Government Cost Estimate (Jun 2015) ✓
  - Contract Award (Sep 2015)) ✓
  - Kickoff meeting with HumRRO/ETS (Sep 2015) ✓
  - Build item difficulty model (Feb 2017) ✓
  - Generate tryout items (May 2017) ✓
  - Conduct data collection on tryout items (Aug 2017) ✓
  - Conduct CAT simulation (Oct 2017) ✓

Automating Generation of Word Knowledge Items (continued)

- **Subtasks** (continued)
  - Evaluate WK generated items (Dec 2018) ✓
  - Refine difficulty model (Feb 2018) ✓
  - Expand templates for contextual items (Mar 2018) ✓
  - Refine WK generator (May 2018) ✓
  - Generate and review 3000 WK items (Sep 2018) ✓
  - Provide final generator, interface, and documentation (Sep 2018) ✓

AFQT Predictor Test (APT)

- **Objective**
  - Develop a short screening test that will accurately predict AFQT
- **Projected Completion**
  - Summer 2018
- **Subtasks**
  - Develop test items (Jun 2012–Jul 2013) ✓
  - Develop and evaluate item selection and scoring algorithms (May 2012–Apr 2013) ✓
  - Elaborate requirements/needs of recruiters by conducting structured interviews (Mar–Nov 2013) ✓
  - Develop web-based software (July 2013–Sep 2014) ✓
  - Government review of software (Sep - Oct 2014) ✓
  - Prepare for implementation on production servers (July 2016–Feb 2017) ✓
  - Conduct pilot testing (May 2017–Jun 2017) ✓
  - Implement operationally nationwide (Summer 2017) ✓
  - Conduct initial validation (Feb 2018) ✓
  - Update prediction algorithms (Jul 2018) ✓
- **Successors**
  - Implementation of APT as a tool for use by military recruiters ✓

Appendix B
List of Acronyms

List of Acronyms

AF	Air Force
AFCA	Air Force Compatibility Assessment
AFCT	Armed Forces Classification Test
AFQT	Air Force Compatibility Assessment
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
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

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 G

OPA

OFFICE OF PEOPLE ANALYTICS

Status Report on ASVAB Evaluation Plan

Mary Pommerich

Defense Personnel Assessment Center

DAC Meeting
March 28-29, 2019
Carmel, CA

PURPOSE AND OVERVIEW

- Provide background and update on status and plans for the evaluation of the tests on the ASVAB.
  - Next Generation ASVAB and ETP
  - Next Generation ASVAB progress report
  - ASVAB evaluation—revisiting why
  - ASVAB evaluation plan refresh
  - ASVAB evaluation plan status report
    - Steps 1–11
  - Future steps

Next Generation ASVAB and ETP[†]

Focus: What tests should be administered as part of the ASVAB or on the ASVAB platform[‡] in the future?

[†] ETP = Enlistment Testing Program

[‡] ASVAB platform = The test delivery modality for the ASVAB and various special tests administered in the ETP

NEXT GENERATION ASVAB: PROGRESS REPORT

- Continue efforts to evaluate and resolve (as needed) issues/concerns pertaining to the new tests of interest (TAPAS, Cyber Test, Mental Counters).
  - ☑ *Ongoing: To be briefed later in this meeting.*
- Continue efforts to evaluate tests currently in the ASVAB.
  - ☑ *Ongoing: Details to follow in this briefing.*
- Complete effort to apply argument-based approach to validation of the ASVAB.
  - ☑ *Ongoing: Status to be briefed at a future meeting.*
- Review and update the psychometric checklist, as needed, for the purpose of evaluating tests to be administered as part of the ASVAB.
  - ☑ *Ongoing: Briefed at the Feb. MAPWG meeting; updates TBD.*

Next Generation ASVAB: Progress report

- Services/proponents complete the updated psychometric checklist for new tests of interest, documenting all new information since a checklist was previously completed.
  - 🕔 *Future effort to follow additional study of new tests.*
- Stakeholders develop a shared vision that defines the purpose and general makeup of the next generation ASVAB.
  - 🕔 *Future effort to follow completion of the argument-based approach to validation of the ASVAB.*
- Establish a systematic process to follow for evaluating potential changes and making decisions regarding tests in the ASVAB.
  - 🗀 *DPAC presented a proposed process for potential changes to the ASVAB in 2014. The proposed process will be revisited and refined in the future as other efforts progress.*

Next Generation ASVAB: Progress report

- Revisit logistical questions with stakeholders, including the feasibility of lengthening the ASVAB and the feasibility of dropping existing tests.

  🕓 *Future effort to follow ASVAB evaluation.*

- Stakeholders summarize impact of potential modifications to the battery and identify resources to support a revised battery.

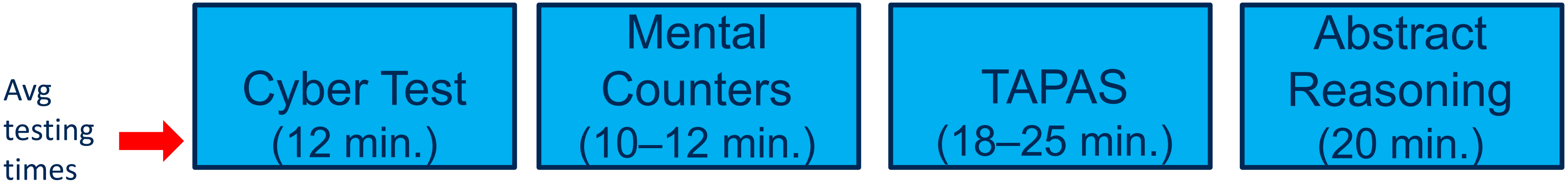
  🕓 *Future effort to follow ASVAB evaluation and evaluations of new tests of interest.*

- Compile all information, then identify and discuss potential changes to the contents of the ASVAB and tests administered in the ETP.

  🕓 *Future effort to follow completion of all above steps.*

ASVAB Evaluation—Revisiting Why

- The Services/DPAC are continuing to study new tests of interest with an eye toward use with *Next Generation ASVAB*:



Avg testing times

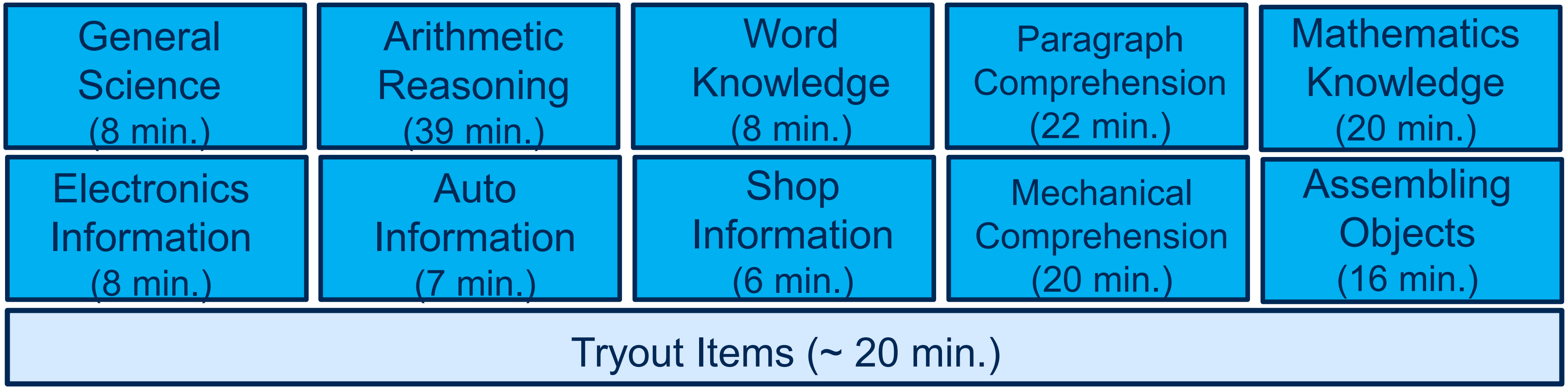
Cyber Test (12 min.)

Mental Counters (10–12 min.)

TAPAS (18–25 min.)

Abstract Reasoning (20 min.)

- Total testing time across the ASVAB and special tests (as well as potentially dated content) continues to be a concern.



General Science (8 min.)

Arithmetic Reasoning (39 min.)

Word Knowledge (8 min.)

Paragraph Comprehension (22 min.)

Mathematics Knowledge (20 min.)

Electronics Information (8 min.)

Auto Information (7 min.)

Shop Information (6 min.)

Mechanical Comprehension (20 min.)

Assembling Objects (16 min.)

Tryout Items (~ 20 min.)

- Hence, there is a strong interest in assessing how the ASVAB might be modified to accommodate new tests.

ASVAB Evaluation—Revisiting Why

- Potential changes to accommodate new tests in *Next Generation ASVAB* could include any combination of the following:

  Dropping existing tests

  Combining existing tests

  Shortening existing tests

  Merging new tests with existing tests

- Research has been ongoing to evaluate the new tests of interest, but the existing ASVAB tests have not systematically undergone similar scrutiny.

A comprehensive assessment of the tests currently in the battery will give insight into their utility, quality, and potential modifiability.

ASVAB Evaluation Plan*

- DPAC has initiated an extensive plan to evaluate the current ASVAB tests in order to determine their desirability/expendability, including:
  - ❑ Reviewing the history of current ASVAB tests and why they were originally included in the battery.
  - ❑ Completing the psychometric checklist and evaluating psychometric value/limitations for each test.
  - ❑ Evaluating the usefulness/appropriateness of existing tests with the current population.
  - ❑ Evaluating item/form development costs.
  - ❑ Evaluating ease/difficulty of developing good, quality items.
  - ❑ Evaluating durability of test content.
  - ❑ Evaluating appropriateness/efficiency of content coverage across tests.
  - ❑ Evaluating vulnerability of content to compromise and other unwanted effects.
  - ❑ Evaluating efficiency of each test.
  - ❑ Evaluating psychometric impact of shortening or combining various tests.
  - ❑ Evaluating psychometric impact of dropping various tests.

*The full plan was briefed to the MAPWG/DAC in 2015.

Step 1: Trace History of Current Tests

- Goal: Document where the ASVAB tests came from and why they were originally included in the battery.
- Team: Tia Fechter, Greg Manley
- Resources:
  - Maier & Sims (1986)
  - Maier (1993)
  - Oppler et al. (1990s)
  - Other possible resources = ?
- Status:

Information regarding the provenance of AO and PC has been found.

Other tests are still in progress.

❑ *PC was included to increase the literacy requirements in the AFQT, in response to findings that recruits had difficulty reading the instructional materials in their training courses.*

❑ *It is a popularly held belief that AO was selected, in part, because it was one of the few ECAT tests that could be administered across both CAT and P&P platforms. In reality, AO was one of the best looking of the 9 ECAT tests, when considered over all analyses.*

Step 2: Complete Psychometric Checklists

- Goal: Complete the psychometric checklist for current ASVAB tests (and possibly Coding Speed) and evaluate psychometric value/limitations of each test.
- Team: Tia Fechter, Greg Manley
- Status: Checklists have been completed for AO and PC. ☑

PROS for PC:

- Incremental Validity
- Contributes less to adverse impact compared to other verbal measures (i.e., WK, GS)
- Resistant to coaching, cheating, and compromise
- Possible candidate for automated item generation through the use of natural language processing

CONS for PC:

- Possible susceptibility to multidimensionality dependent upon screen size
- Ceiling effects evidenced in the past
- Sensitivity concerns related to content
- Durability concerns related to content
- Testing time requirements are lengthy compared to other ASVAB tests
- Limited to one item per passage in CAT-ASVAB modality

Step 2: Complete Psychometric Checklists

PROS for AO:

- Nonverbal
- Unique domain (spatial ability)
- Less potential for adverse impact
- Predictive validity with training criteria
- Has demonstrated incremental validity over other ASVAB tests
- Good potential for classification efficiency
- An excellent candidate for automated item generation
- Less vulnerable to practice effects than other psychomotor or spatial tests
- Less vulnerable to compromise
- Less critical to update pools frequently

CONS for AO:

- Significant ceiling effect due to low item difficulty
- Multidimensionality concerns
- Necessary to break the scale and start over for next generation AO (i.e., introduce separate AC and AP scales)
- Possible platform efforts
- Somewhat labor intensive in terms of item formatting requirements

Step 3: Evaluate Usefulness, Appropriateness

- Goal: Evaluate the usefulness and appropriateness of existing tests with regard to the current population.
- Task 3a: Track trends in test scores over years 1984–2019.
  - Team: Tia Fechter, Robert Hamilton, Lihua Yao, Ping Yin
  - Status:

    Located data:
    - 1997–current for CAT-ASVAB
    - 2002–current for P&P-ASVAB

    Not yet located data:
    - 1990–1996 for CAT-ASVAB
    - 1984–2001 for P&P-ASVAB
- Task 3b: Evaluate what fraction of the population possesses the knowledge/skill assessed by the test.
  - Task 3b(i): Evaluate overlap between latent ability and score information for current testing population.
    - Team: Mary Pommerich, Ping Yin
    - Status: Required programs in place. Need to apply to current data.

Step 3: Evaluate Usefulness, Appropriateness

- Goal: Evaluate the usefulness and appropriateness of existing tests with regard to the current population.
- Task 3b(ii): Conduct pseudo-standard setting and evaluate percent in each category over time (technical tests only).
- Team: Tia Fechter, Dan Segall
- Status:

In Progress:

- Plan standard setting for technical tests—AI, SI, EI, and MC (4 weeks).

Next Steps:

- Implement pseudo-standard setting (8 weeks).
- Analyze data collected (1 week).
- Recommend cut scores (1 day).
- Summarize percent in each category over years (1 week).
    - [Will use archival data from Task 1].

Step 4: Evaluate Item Development Costs

- Goal: Identify estimated yearly costs for item development.
  - Task 4a: Identify cost per item per test.
  - Task 4b: Identify desired form replacement schedule.
  - Task 4c: Identify number of items needed per year per test.
  - Task 4d: Identify total yearly cost per test.
  - Team: Jeff Harber, Mary Pommerich
  - Status = Completed ☑

Subtest	Cost Per Item (Approx.)	# Pools Per Year (Target)	# Items Per Year (Target)	Total Yearly Cost (Estimated)
General Science	$X	4	800	$X
Arithmetic Reasoning	$X	4	800	$X
Word Knowledge	$X	8	1600	$X
Paragraph Comprehension	$X	4	800	$X
Mathematics Knowledge	$X	4	800	$X
Electronics Information	$X	2	400	$X
Automotive Information	$X	2	400	$X
Shop Information	$X	2	400	$X
Mechanical Comprehension	$X	2	400	$X
Assembling Objects*	TBD	TBD	TBD	TBD

Step 5: Evaluate Ease of Developing Good Items

- Goal: Evaluate the overall ease/difficulty of developing good quality items.
  - Task 5a: Identify finiteness of domains [more finite = less ease].
  - Task 5b: Evaluate feasibility of using automatic item generation with test content [less feasible = less ease].
  - Task 5c: Identify item retention rates [less retention = less ease].
  - Team: Jeff Harber, Mary Pommerich, Matt Trippe

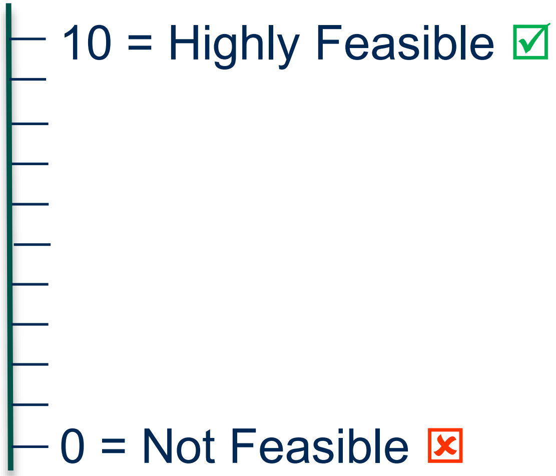
Subtest	Finiteness Rating
General Science	6
Arithmetic Reasoning	10
Word Knowledge	6
Paragraph Comprehension	10
Mathematics Knowledge	8
Electronics Information	4
Automotive Information	4
Shop Information	4
Mechanical Comprehension	5
Assembling Objects	10

10 = Infinite ☑

0 = Finite ☒

Step 5: Evaluate Ease of Developing Good Items

Subtest	Ease of AIG Rating	Quality of AIG Rating[†]
General Science	TBD	TBD
Arithmetic Reasoning	TBD	TBD
Word Knowledge	TBD	TBD
Paragraph Comprehension	?	?
Mathematics Knowledge	TBD	TBD
Electronics Information	?	?
Automotive Information	?	?
Shop Information	?	?
Mechanical Comprehension	?	?
Assembling Objects	TBD	TBD



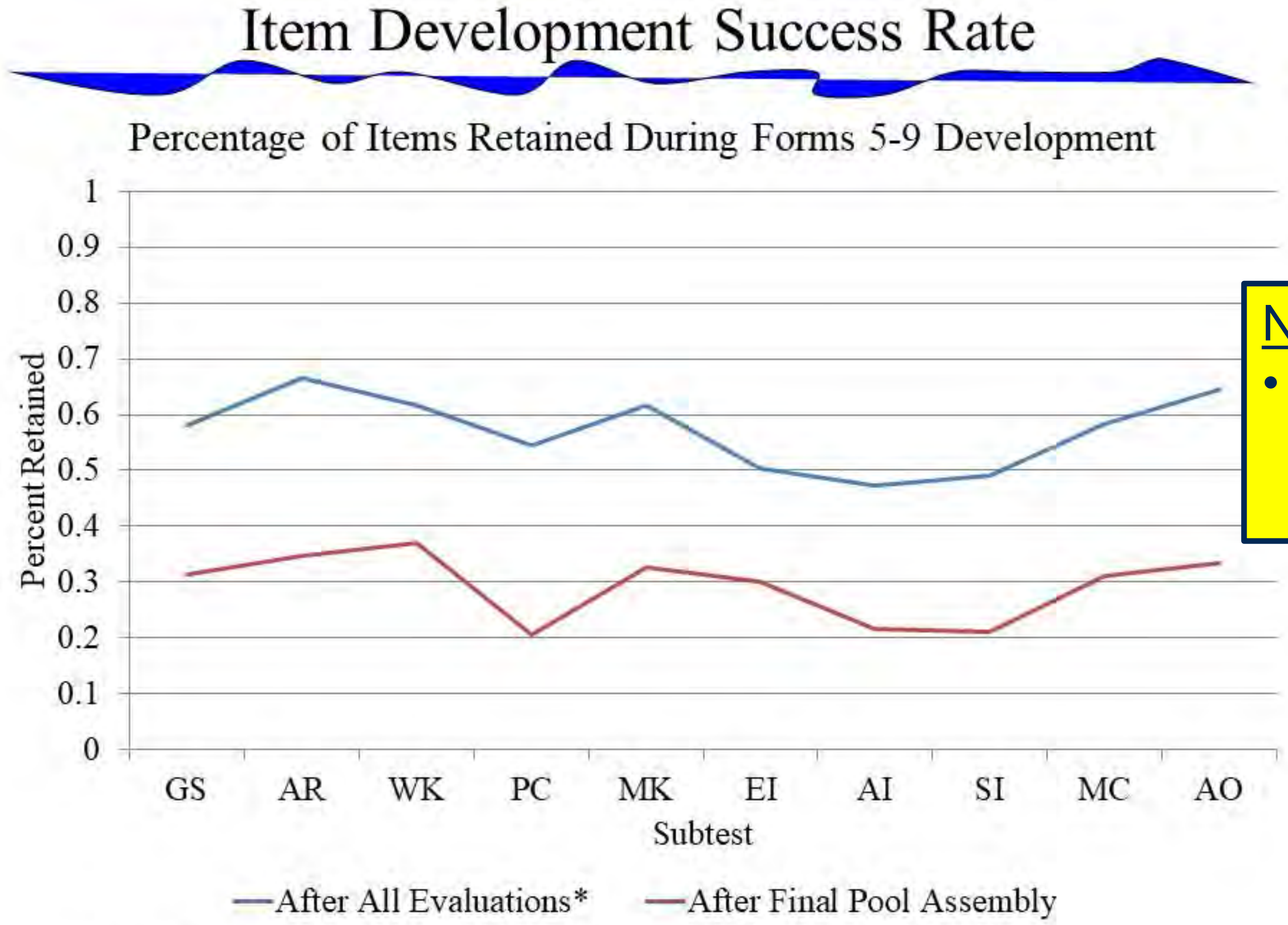
10 = Highly Feasible ☑

0 = Not Feasible ☒

[†] For tests where AIG has been introduced, Quality of AIG rating should take into account the following questions:

- How much formatting/manipulation is required after generation?
- Can AIG be applied to all item types within a subtest?
- Can traditional item calibrations be eliminated or requirements reduced?
- What percentage of items are estimated to be usable?

STEP 5: EVALUATE EASE OF DEVELOPING GOOD ITEMS



* Excludes all items in the bottom quarter of score information

Next Step:

- Update based on Forms 11–15 development.

Step 6: Evaluate Durability of Test Content

- Goal: Evaluate how likely content is to stand the test of time.
- Task 6a: Evaluate extent to which content is (or appears) less relevant to today's applicant population (see also Step 3).
- Task 6b: Evaluate extent to which content is prone to obsolescence.
- Task 6c: Evaluate extent to which content is in need of frequent updating in order to stay current.
- Task 6d: Evaluate extent to which it is difficult to keep up with new technology or changes in technology.
- Team: Tia Fechter, Jeff Harber, Sachi Phillips

Next Step:

- Develop rating scales for each task and rate subtests.

Step 7: Evaluate Efficiency of content coverage

- Goal: Review prior research and summarize findings regarding the efficiency and adequacy of content coverage (i.e., redundancies and gaps).
- Task 7a: Identify redundancies in content coverage across tests.
- Task 7b: Identify gaps in content coverage.
- Task 7c: Identify potentially unnecessary content coverage.
- Team: Tia Fechter, Jeff Harber

Next Step:

- Review relevant literature.

Step 8: Evaluate Vulnerability to Compromise

- Goal: Evaluate the vulnerability of item content and item pools to compromise.
- Task 8a: Identify features of tests that could make them easy to compromise.
- Task 8b: Identify features of item pools that could make them easy to compromise.
- Task 8c: Identify previous incidences of compromise on the ASVAB and tests that were breached.
- Team: Tia Fechter, Jeff Harber, Sachi Phillips, Dan Segall

In Progress:

- Information gathering and review of prior compromise history in ETP, AFCT, and CEP is underway.

Next Step:

- Develop rating scales to summarize vulnerability across the various factors.

Step 9: Evaluate Other Vulnerabilities

- Goal: Evaluate the vulnerability of item content to other unwanted effects.
- Task 9a: Coachability
- Task 9b: Practice Effects
- Task 9c: Hardware Effects
- Task 9d: Mode Effects
- Task 9e: Local Dependence
- Team: Tia Fechter, Jeff Harber, Sachi Phillips, Mary Pommerich, Dan Segall

Next Steps:

- Review prior findings for ASVAB tests.
- Develop rating scales to summarize vulnerability across the various factors.

Step 10: Evaluate Efficiency of Each Test

- Goal: Evaluate the relative efficiency of each test, with regard to testing time allotted and testing time used.
- Task 10a: Summarize total testing time allocated on CAT-ASVAB.
- Task 10b: Summarize observed testing times for applicants, total and per test.
- Task 10c: Summarize time allocated versus time spent, per item and per test.
- Team: Furong Gao, Mary Pommerich, Dan Segall

In Progress:

- Details on time allocated and time used to be briefed later in this meeting.

Next Step:

- Develop rating scales to summarize testing efficiency.

STEP 11: SYNTHESIZE FINDINGS

- Goal: Synthesize findings across all evaluation criteria and tests and summarize the desirability/expendability of each test.
- Team: Tia Fechter, Furong Gao, Jeff Harber, Greg Manley, Sachi Phillips, Mary Pommerich, Dan Segall, Matt Trippe, Lihua Yao, and Ping Yin

Next Steps:

- Identify a way to concisely summarize results over all steps.
- Identify a way to aggregate findings and compute an overall rating.

Any suggestions would be greatly appreciated!

FUTURE STEPS

- Goal: Evaluate the impact and feasibility of dropping, combining, or shortening existing tests, or merging new and existing tests.
- Step 12: Evaluate psychometric impact of shortening various tests (Mary Pommerich, Ping Yin).
- Steps 13–17: Evaluate (1) psychometric impact of shortening AR and/or MK and computing a math composite score, (2) feasibility and psychometric impact of combining AR & MK into a single test, (3) feasibility and psychometric impact of combining EI and Cyber Test into a single test (Furong Gao, Lihua Yao, Ping Yin).
- Steps 18–24: Evaluate the psychometric impact of dropping AI, SI, AO, EI, MC, GS, WK (TBD, Service technical reps).

Tab H

ASVAB Psychometric Support

Development of ASVAB Forms 11–15

Presenter : D. Matthew Trippe, HumRRO

March 28, 2019

Objectives

- Short recap of previous briefings
- Item enemy identification
- ASVAB form assembly
- Form assembly CAT simulation analyses
- Schedule
- Questions/Discussion

Recap of Previous Briefings

- Goal is to develop more ASVAB forms on a more aggressive schedule
- Begin with forms 11–1**5**, which will be assembled from experimental items administered under old and new configurations
- **Replace operational forms & PiCAT**
  - **Will develop one additional form over original goal of 4 forms**
- Forms 11–1**5** will be assembled from **ten** experimental item "series," or sets, of 100 experimental items per test
- Each experimental item is reviewed for psychometric (e.g., model fit, information) and content quality
- Items that survive review process move on to form assembly

Recap: Forms 11–15 Development—Completed Steps

Test	89000	89100	89200	89300	89400	89500	89600	89700	89800	89900
WK	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Item Enemy ID	Item Enemy ID
GS	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Item Enemy ID	Item Enemy ID
AR	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Item Enemy ID	Item Enemy ID
PC	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Item Enemy ID	Item Enemy ID
MK	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Item Enemy ID	Item Enemy ID
EI	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Initial Screening	Initial Screening
AI	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Initial Screening	Initial Screening
SI	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Initial Screening	Initial Screening
MC	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Prelim FrmAsmbl	Initial Screening	Initial Screening
AC	External Reviews	External Reviews	Content Review	Item Analyses	Item Analyses	–	–	–	–	–
AP	External Reviews	External Reviews	Content Review	Item Analyses	Item Analyses	–	–	–	–	–

Enemy Item Identification

Enemy Item Identification: Background

- Local dependence (LD) analysis was conducted prior to assembly of forms 5–9 (Pommerich & Segall, 2008)
  - Results suggest MK and MC tests susceptible to LD
- Mitigating LD requires identification of item enemy groups
  - Items likely to trigger LD if administered to the same person
  - Two or more items that measure similar or highly related content
- Before assembling forms 5–9, DPAC developed a content framework for identifying enemy groups
  - 700+ items per test evaluated for match with enemy group
    - MC: 95 content areas
    - MK: 155 content areas
- HumRRO developing a procedure to optimize human judgment + quantitative roles

Enemy Item Identification: Methods

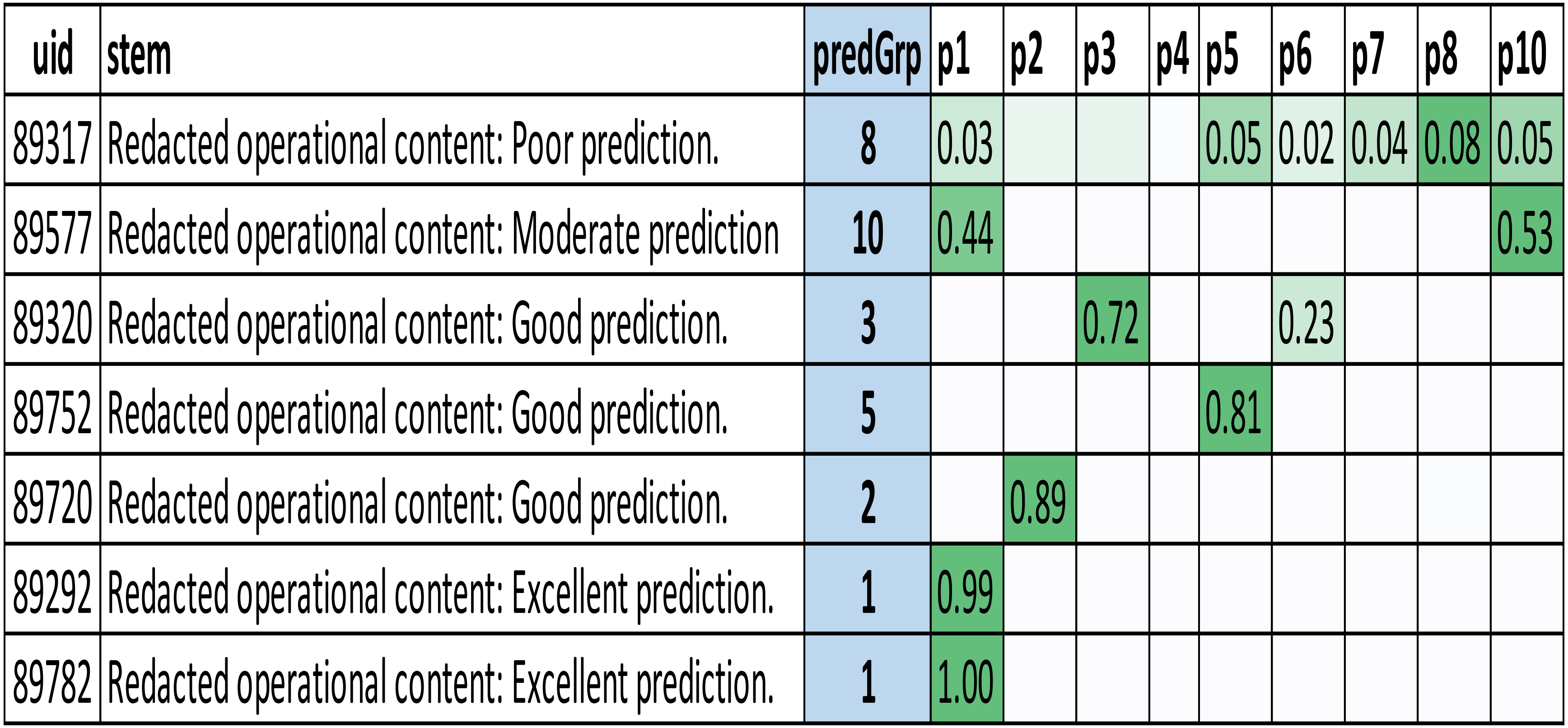
- Mechanical Comprehension & Math Knowledge Tests
  - Method 1
    - Two humans independently link each item to the DPAC defined categories; identify new categories as necessary
    - Resolve disagreements with third rater
  - Method 2
    - Unsupervised classification using text analysis of items
    - Supervised classification analysis based on existing DPAC content framework developed during form 5–9 construction
      - Outcome “Y” we are predicting is enemy group label
  - Method 3
    - Sparse data local dependence analysis
    - Addresses the local dependence concerns directly through Q_3 (Yen, 1984)
    - Will be possible for seed series to be included in future form assembly efforts
- All other tests
  - Unsupervised classification using text analysis of items
  - Human review of “heatmap” hot spots

Enemy Item Identification: Results

- Math Knowledge (MK)
  - Prediction is generally good at higher-order group level (e.g., angles)
  - Prediction is not possible at sub-group level (e.g., angles complementary, angles obtuse)
  - Human judgment cannot be replaced, but a complicated and tedious task can be simplified and accelerated with model-based tools
    - Group assignment probability matrix
    - Heatmap
- Mechanical Comprehension (MC)
  - Prediction is generally not good
  - Much of the information in MC is stored in images/artwork, which is (so far) difficult to quantify or tokenize for this type of analysis
  - Many group membership assignments are based on similarity to existing items in the group rather than an entirely discrete concept
  - There is more conceptual overlap between groups in MC than in MK
  - Will continue to work on ways to improve this process for MC

Enemy Item Identification: Results

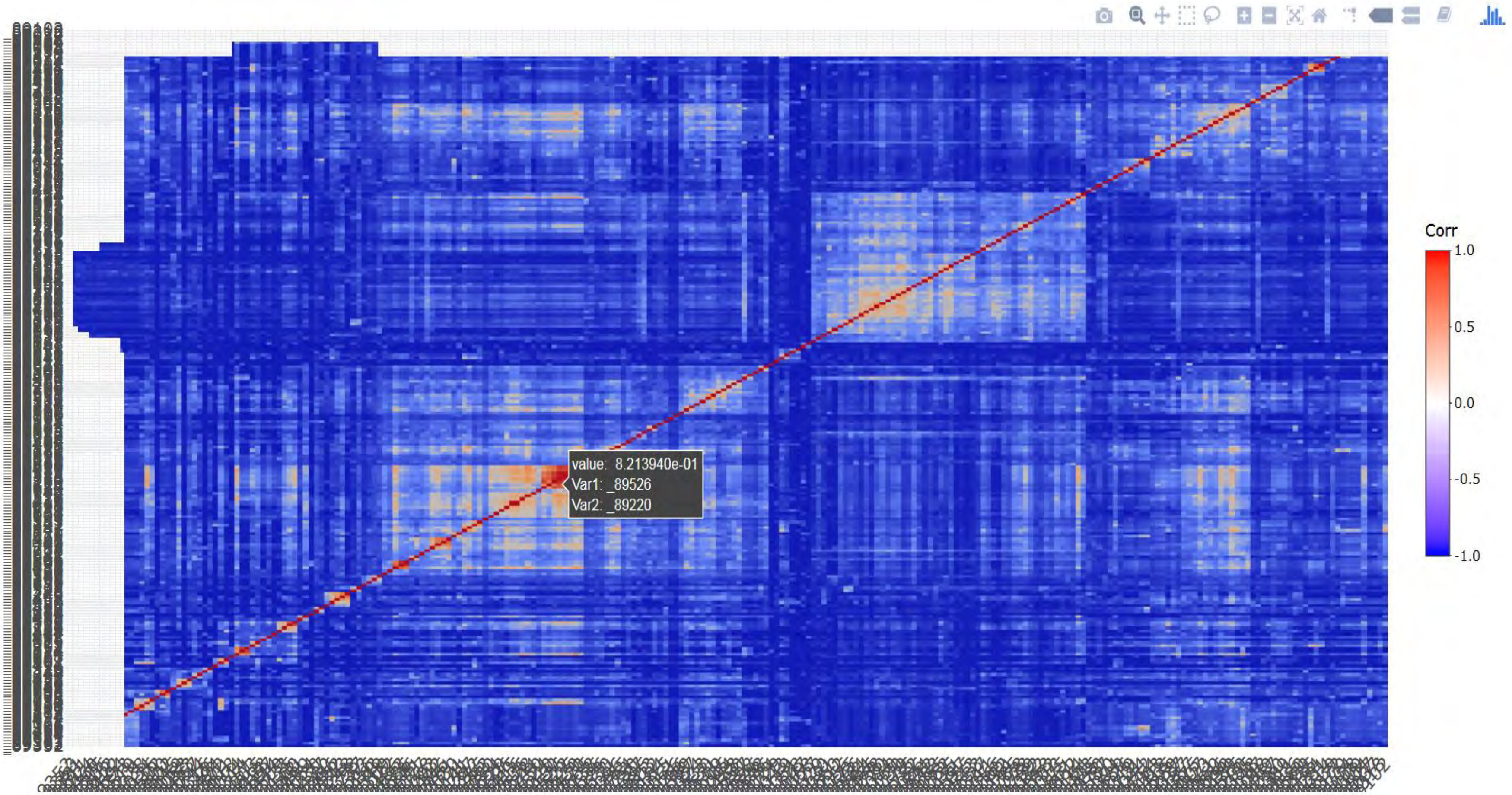
- Example tool to facilitate MK item enemy review



uid	stem	predGrp	p1	p2	p3	p4	p5	p6	p7	p8	p10
89317	Redacted operational content: Poor prediction.	8	0.03				0.05	0.02	0.04	0.08	0.05
89577	Redacted operational content: Moderate prediction	10	0.44								0.53
89320	Redacted operational content: Good prediction.	3			0.72			0.23			
89752	Redacted operational content: Good prediction.	5					0.81				
89720	Redacted operational content: Good prediction.	2		0.89							
89292	Redacted operational content: Excellent prediction.	1	0.99								
89782	Redacted operational content: Excellent prediction.	1	1.00								

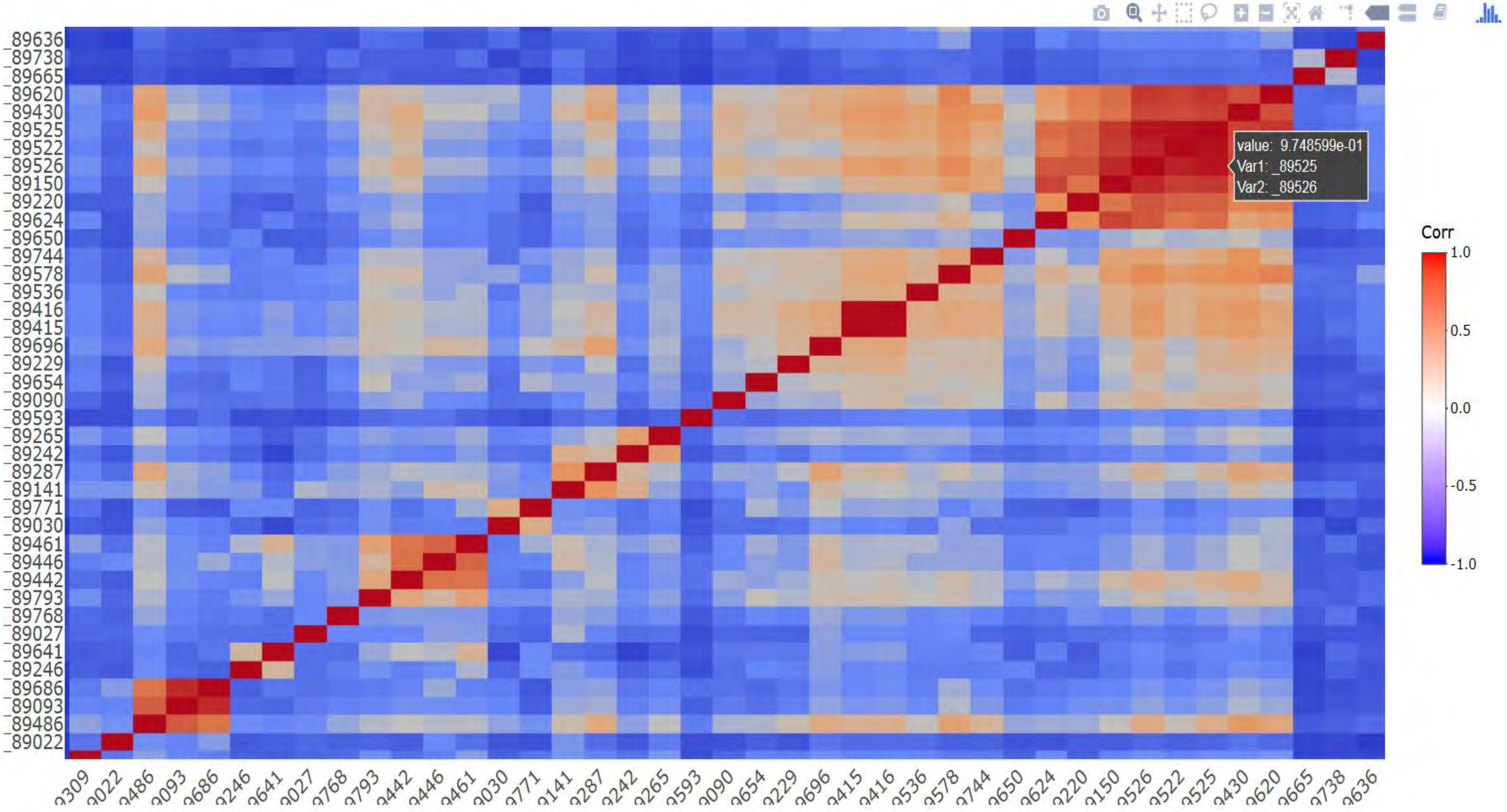
Enemy Item Detection: Results

- Example (GS) tool to facilitate item enemy review in all tests



Enemy Item Detection: Results

- Example (GS) tool to facilitate item enemy review in all tests



Form Assembly

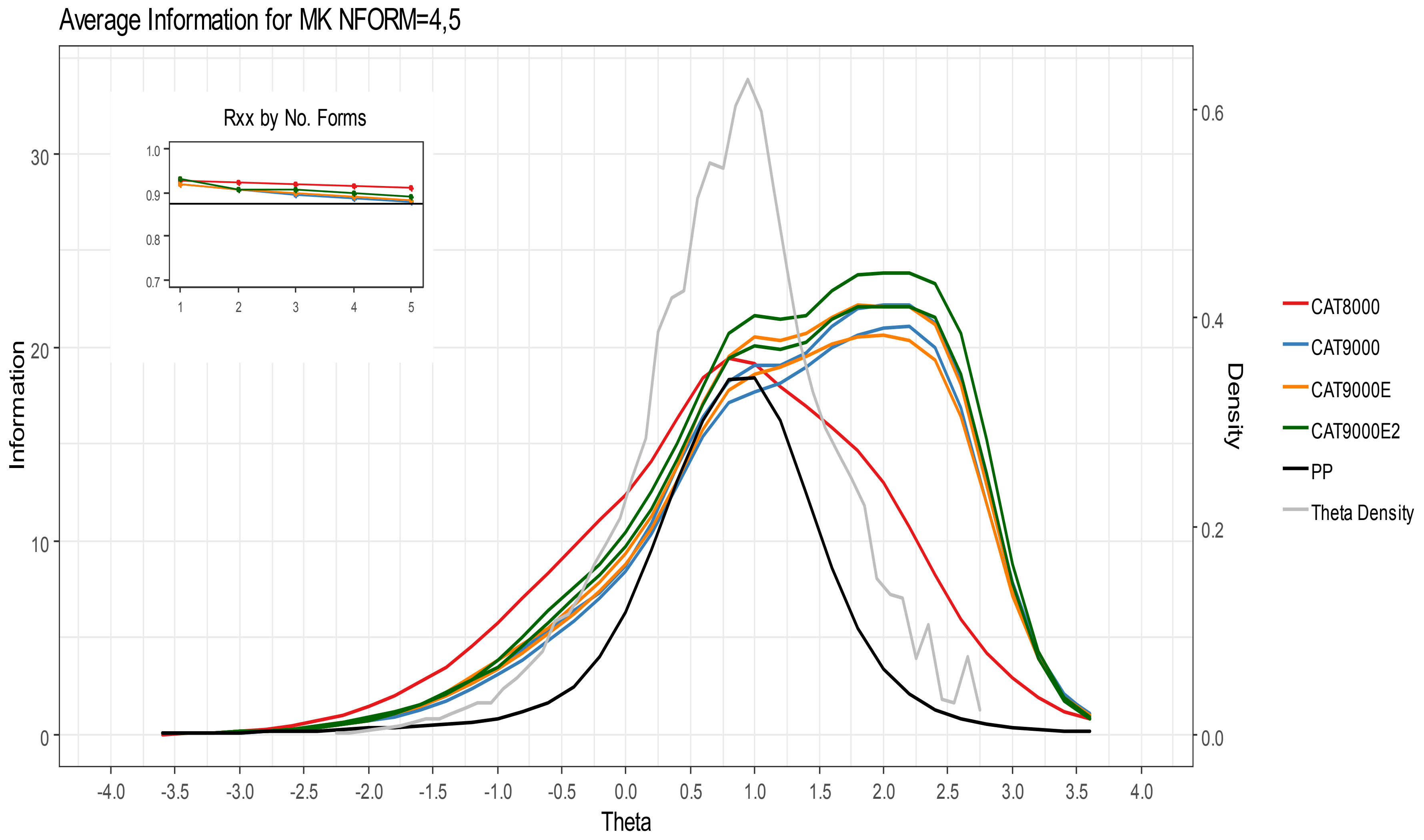
ASVAB Form Assembly

- CAT forms
  - CAT administration is based on forms from which a *potentially* unique set of items is administered to each examinee
  - Forms need to contain items from the full range of content and difficulty
  - Forms need to contain sufficient information/score precision across the full range of ability
- Form assembly goals
  - For each test, assign each item to one of five forms (11, 12, 13, 14, 15)
  - Maximize conditional precision levels of each form
  - Constrain conditional precision levels to be comparable across forms
  - Account for "enemy" items—distribute them evenly across pools
  - Account for content taxonomies where applicable (GS, AO)

ASVAB Form Assembly: Simulation Analyses

- Assemble forms algorithmically to optimize stated goals
  - Assembled main analytic functions in Fortran as dynamic-link library (DLL)
  - Develop R package to wrap FORTRAN functions and to implement CAT analyses
  - Best of both worlds approach, combining speed of FORTRAN and flexibility of R, facilitates changes to problem configurations, analysis of results, and promotes QC.
- Compute test information using CAT simulations
  - Partition entire item pool into four or five candidate forms
  - Calculate item exposure parameters using preliminary CAT simulations
  - Generate large sample of scored responses using another round of CAT simulations
  - Approximate test information based on sample of scored responses
  - Trim items not administered in the second round of CAT simulations
- Compare information to
  - Original P&P ASVAB
  - Current operational forms (5–9)
  - Observed theta density
- Expand items available in form assembly
  - Unused/unassigned items from forms 5–9 assembly
  - Additional item series (89800 & 89900)

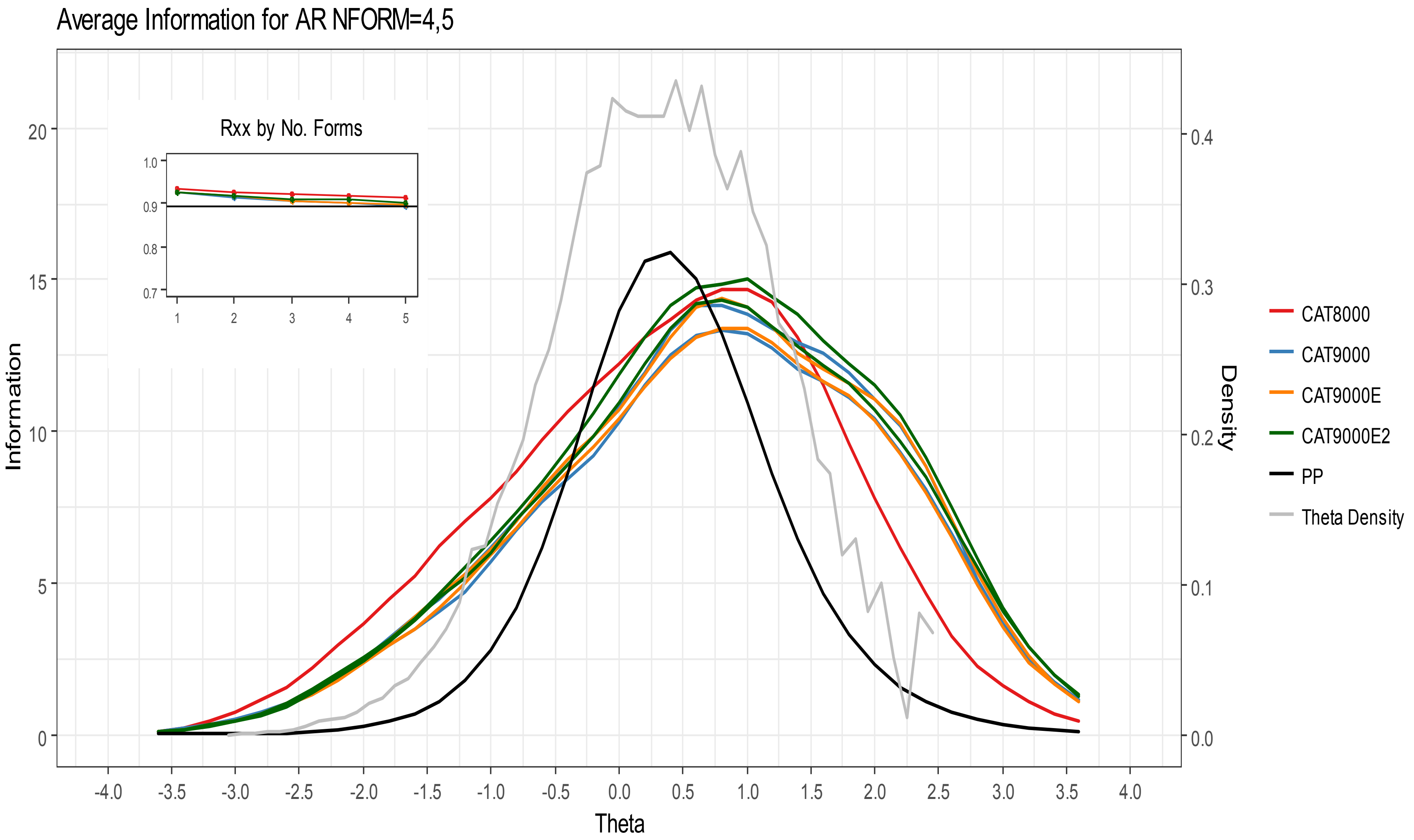
ASVAB Form Assembly: MK Simulation Results



ASVAB Form Assembly: Simulation Findings

- MK information is not well aligned with existing operational forms or observed applicant ability
  - Including previously unassigned items mitigates this issue in the middle range of ability
  - Including additional series (89800 & 89900) provides more information where needed (low to middle theta)
- All other tests
  - Information alignment is comparable to existing operational forms
  - Including additional series will help maintain information with new goal of five forms
- Psychometric team must coordinate with item development team(s) regarding information alignment with observed ability distribution

ASVAB Form Assembly: AR Simulation Results



Schedule

Schedule

- Tasks remaining include:
  - 89800 & 89900 series technical tests seed items
    - Data cleaning
    - Calibration
    - Rescaling
    - Initial screening (begin here for AFQT and GS tests)
    - Enemy identification
  - Preliminary form assembly
    - Identify additional item enemies in tests other than MC and MK
      - Heatmap analysis should reduce effort here
    - Evaluate score information functions
  - Final form assembly
  - Equating and equating analyses/evaluation
    - HumRRO team recently completed equating on form 10 and is on top of the learning curve for equating forms 11–15

Questions?

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

HumRRO Team

- Adam Beatty
- Maura Burke
- Ted Diaz
- Amanda Koch
- Justin Purl
- Peter Ramsberger
- Matthew Reeder
- Matthew Trippe

Tab I

OFFICE OF PEOPLE ANALYTICS

Mental Counters

Identifying Examinees Who Are “Not Trying”

Presented to the DACMPT

Ping Yin, HumRRO
Mary Pommerich, DPAC
March 28, 2019 | Carmel-By-The-Sea, CA

OVERVIEW

- **Brief review**
  - The Mental Counters Test (MCt)
  - Observed floor effect
- **Purpose of the study**
- **Review of previous research**
- **Analysis and results**
- **Summary**
- **Future research**

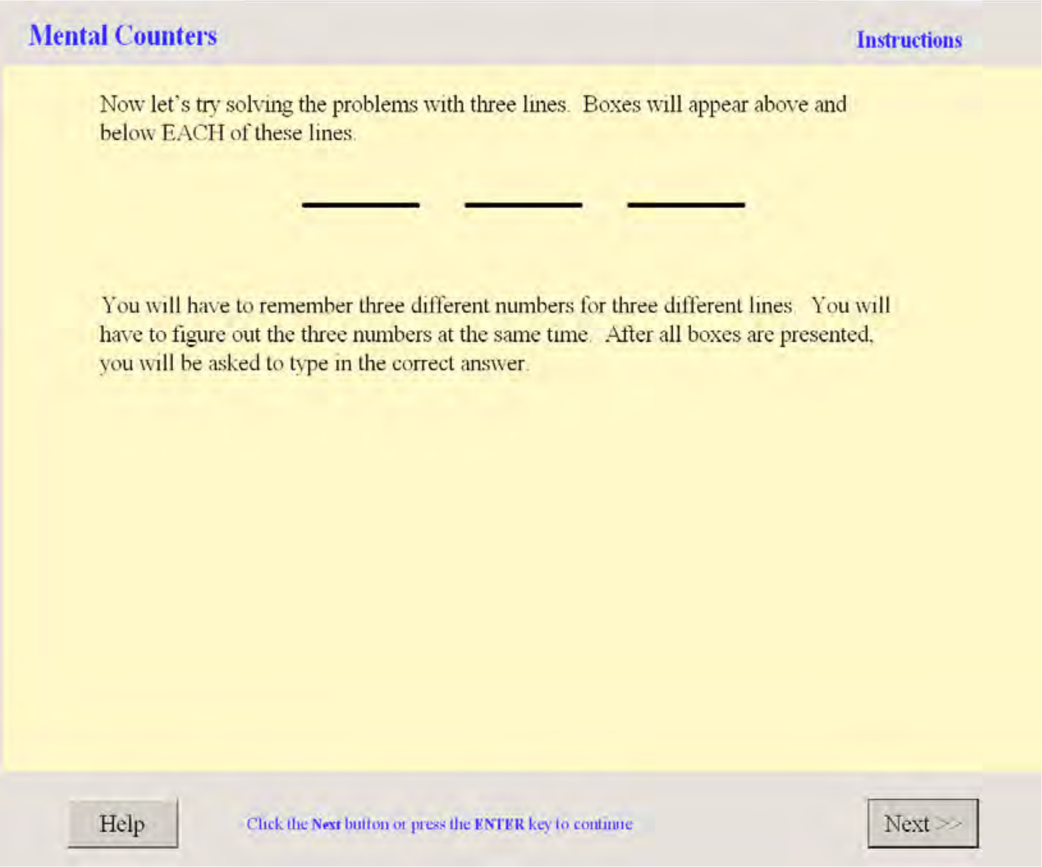
Brief Review

MENTAL COUNTERS: REVIEW

- **Mental Counters (MCt) is a test of working memory (WM), originally developed by the Navy and studied as part of the Enhanced Computer-Administered Test (ECAT) battery evaluation**
  - 32 items
  - Currently administered to Navy applicants on the CAT-ASVAB platform
  - Measures a unique domain not represented on the ASVAB: WM has a short duration of 10–15 seconds and may hold 4 to 5 pieces of new information
    - Evidence of incremental and predictive validity
    - Evidence of classification efficiency
    - Evidence of excellent reliability
    - No adverse impact for gender
    - Very short testing time
    - Excellent candidate for automatic item generation
    - Unique domain not represented in the ASVAB or other special tests

MENTAL COUNTERS: REVIEW

- The MCt test requires the examinee to count the number of boxes that flash above or below one of three stationary lines on the computer screen



Mental Counters

Instructions

Now let's try solving the problems with three lines. Boxes will appear above and below EACH of these lines.

You will have to remember three different numbers for three different lines. You will have to figure out the three numbers at the same time. After all boxes are presented, you will be asked to type in the correct answer.

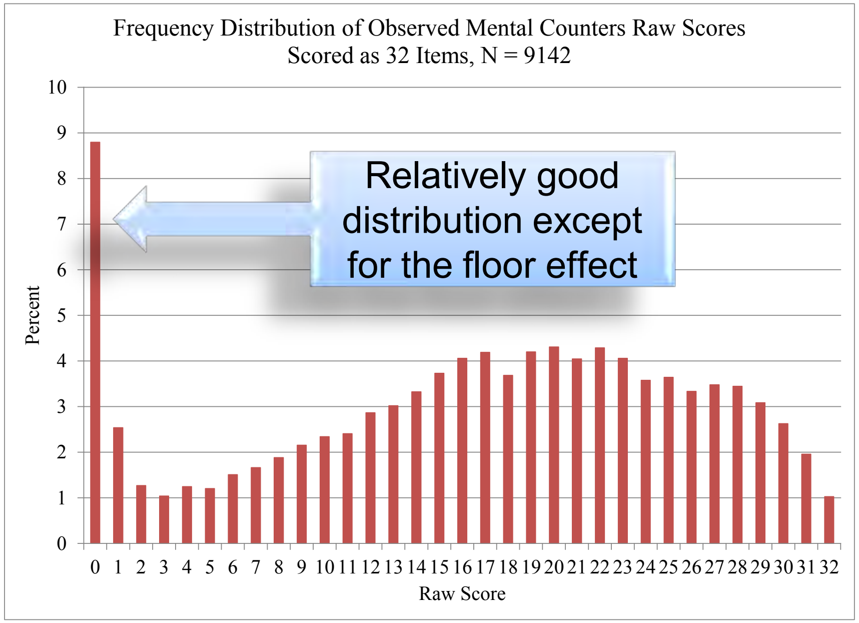
Click the **Next** button or press the **ENTER** key to continue

MENTAL COUNTERS: REVIEW

- **There are three counters for each MCt item**
- **Counters for each line start at 5**
- **A value of 1 is [added to]/[subtracted from] the counter for each line if a box appears [above]/[below] the line**
- **After all boxes are presented, an examinee is asked to enter/type the correct answers (the three numbers) in the correct sequence**
- **A MCt item is answered correctly only if all three numbers are entered correctly**
  - If any of the three numbers is incorrect, or the numbers are not in the correct sequence, the item is scored as incorrect

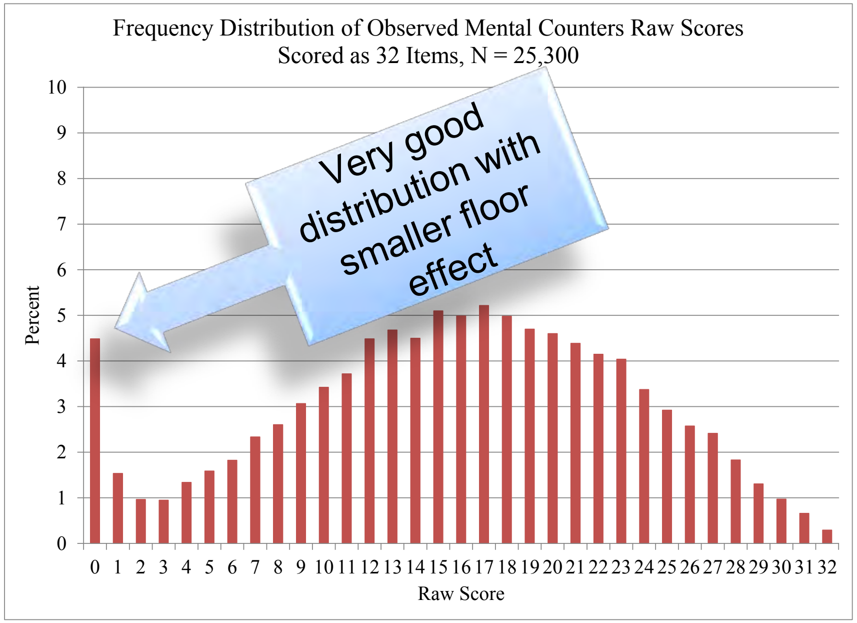
OFFICE OF PEOPLE ANALYTICS

REVIEW OF OBSERVED FLOOR EFFECT


Frequency Distribution of Observed Mental Counters Raw Scores
Scored as 32 Items, N = 9142

Relatively good distribution except for the floor effect

Version 2.0 (2013)

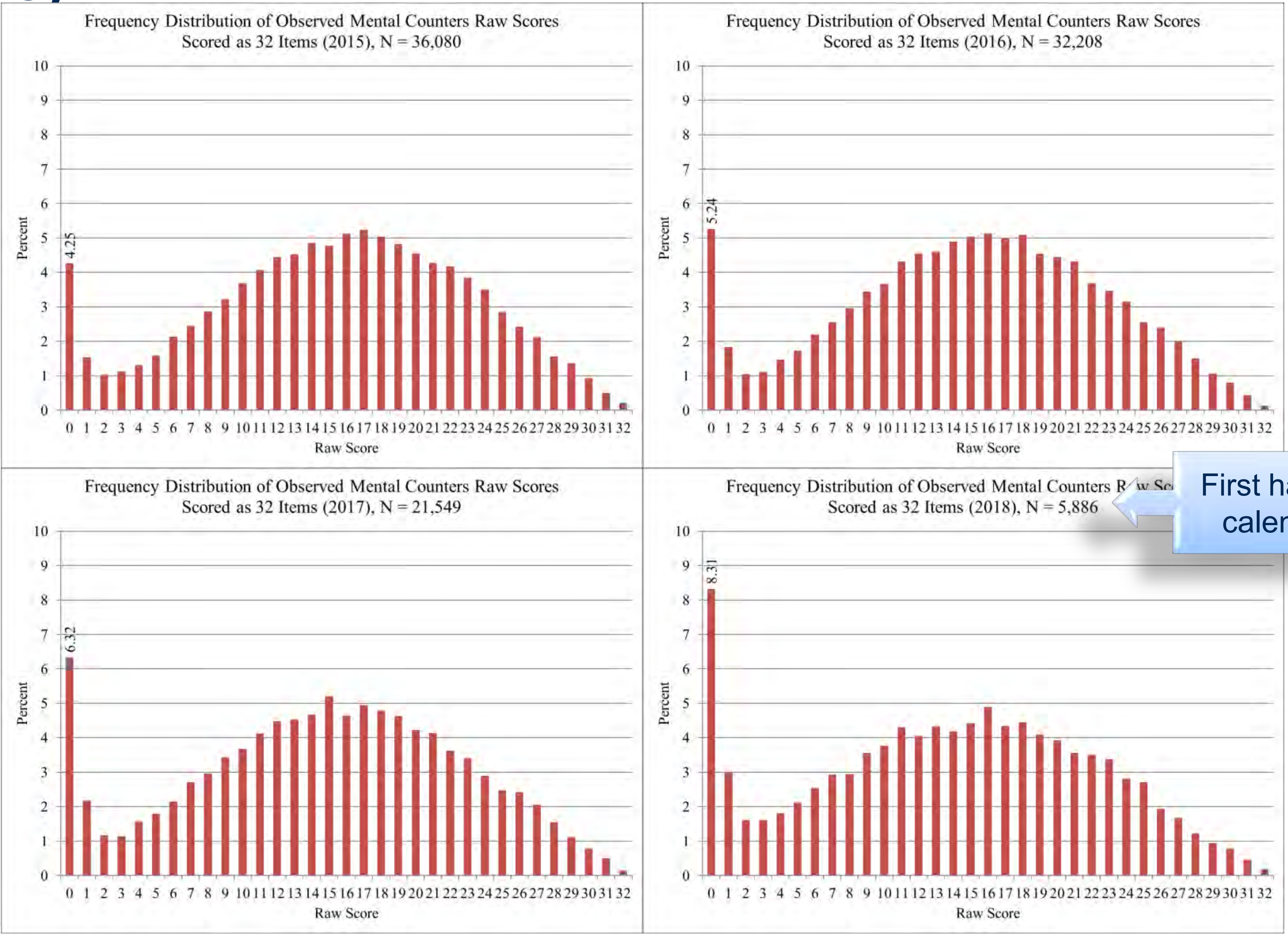

Frequency Distribution of Observed Mental Counters Raw Scores
Scored as 32 Items, N = 25,300

Very good distribution with smaller floor effect

Version 3.0 (2014)

Minor clarification of instructions to emphasize that the counter starts at 5.

FLOOR EFFECT OVER TIME: VERSION 3.0 (2015–2018)


Frequency Distribution of Observed Mental Counters Raw Scores Scored as 32 Items (2015), N = 36,080

Frequency Distribution of Observed Mental Counters Raw Scores Scored as 32 Items (2016), N = 32,208

Frequency Distribution of Observed Mental Counters Raw Scores Scored as 32 Items (2017), N = 21,549

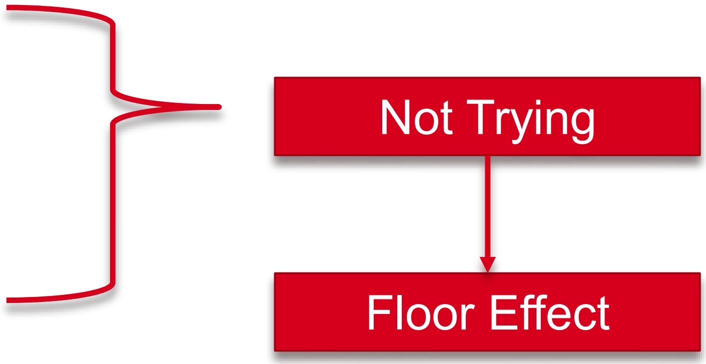
Frequency Distribution of Observed Mental Counters Raw Scores Scored as 32 Items (2018), N = 5,886

First half of 2018 calendar year

Purpose of the Study

POSSIBLE FACTORS CONTRIBUTING TO THE FLOOR EFFECT

- **Previously presented at the August 2018 MAPWG meeting:**
  1. Not able to understand the task (instruction)
  2. Lack of motivation
  3. Too difficult
  4. Fatigue and/or frustration
  5. Combinations of various factors above



Not Trying

Floor Effect

- **Operational definition of “not trying”:**
  - At the test level:
    - Spends less time over all items compared to those who are making an effort
  - At the item level:
    - An observable pattern of spending less time on items as item number increases (sequential or order effect)
    - Is independent of item design (i.e., the number of adjustments and delay)

OFFICE OF PEOPLE ANALYTICS

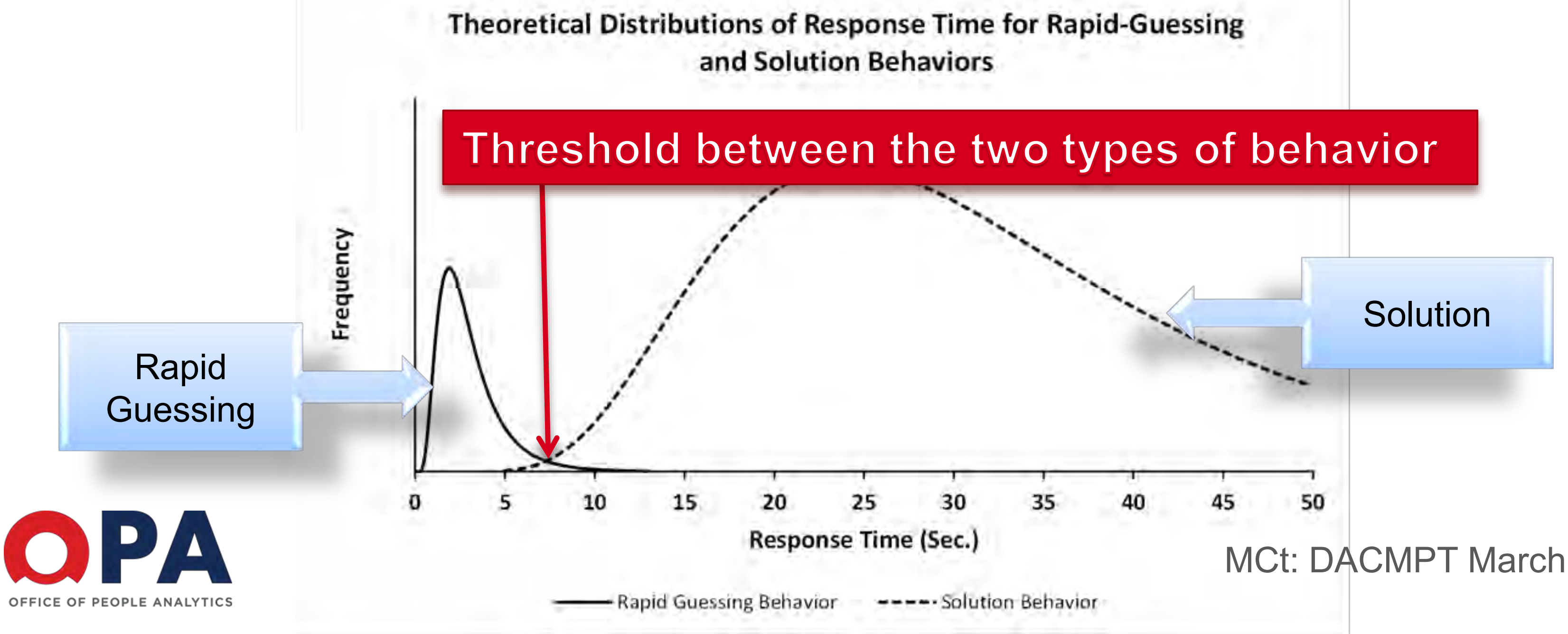
RESEARCH QUESTIONS

1. **Is it feasible to identify examinees who are “not trying” on MCt, using the response time distribution?**
2. **Is it feasible to identify examinees who are “not trying” on MCt, using an index at the test level?**
   - Examine the floor effect after removal
   - Examine the correlation coefficient between ASVAB (subtest and AFQT) and MCt total score after removal
3. **Is it feasible to identify examinees who are “not trying” on MCt by examining the item-level sequential effect?**
   - Is there any observable pattern of the sequential effect?

Review of Previous Research

REVIEW OF PREVIOUS RESEARCH

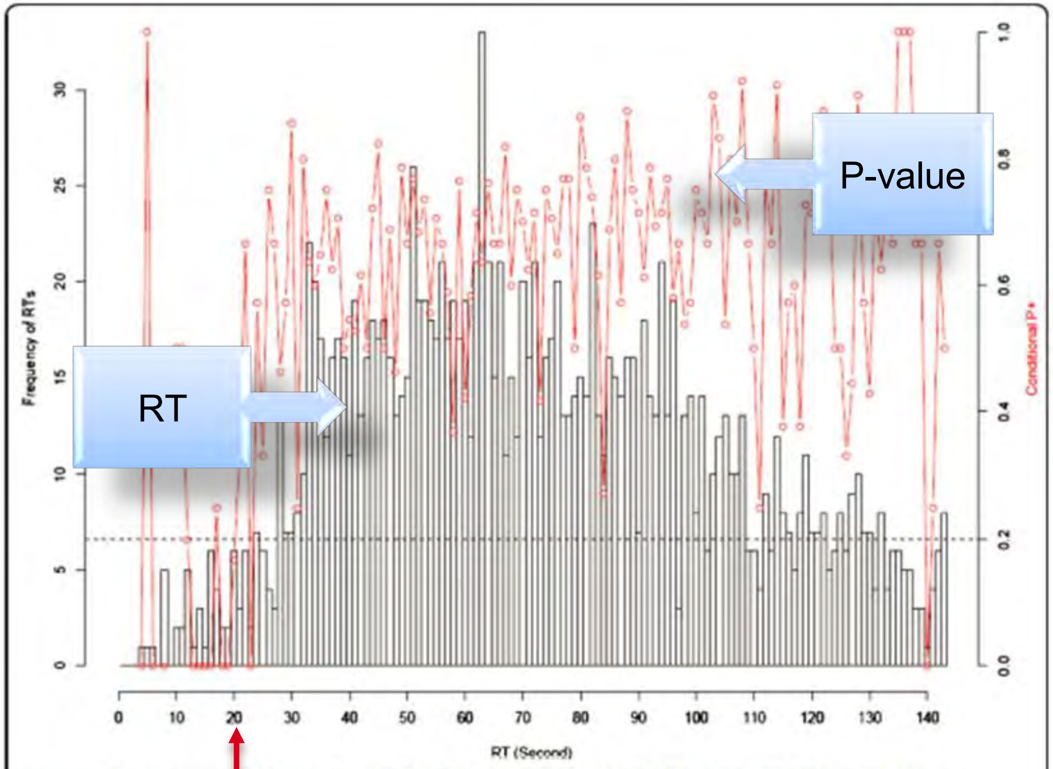
- **Research is largely based on achievement tests**
- **Response Time (RT)**
  - Schnipke (1995) noted that the RT distribution for the incorrect answers often had a sharp spike during the first few seconds (rapid-guessing behavior), and the RT distribution for the correct answers had a broader distribution with a smaller peak (solution behavior)
  - In **<u>theory</u>**, the combined RT distribution tends to be bimodal and positively skewed



OPA
OFFICE OF PEOPLE ANALYTICS

REVIEW OF PREVIOUS RESEARCH

- **RT and Response Accuracy Distributions**
  - Lee & Jia (2014) noted that for multiple-choice items, the conditional p-value associated with rapid guessing is expected to be near the chance level
  - In the example, RT distribution is somewhat bimodal, but the two modes are not distinct
  - The conditional p-values fluctuate widely above and below the chance level (0.2) until RT reached about 20 seconds



RT

P-value

Figure 4 Histogram of RTs and conditional P+ for an item exhibiting identifiable patterns (truncated at the 90th percentile of RTs and based on the Control sample). The bars represent the frequency of RTs at every possible RT, the circles joined by line represent the conditional P+, and the horizontal dashed line at conditional P+ equal to 0.2 represents the chance level.

Estimated Threshold

REVIEW OF PREVIOUS RESEARCH

- **The test-level index**
  - Rationale: not all RT distributions show two distinct modes; visual inspection may not be reliable/feasible; RT can vary by item difficulty and complexity
  - The response-time effort (RTE) index was developed to identify examinees who are engaged in the solution behavior over all items
    1. Wise & Ma (2012) defined the threshold for an item as a percentage of the average response time (e.g., 10%). A solution behavior (SB) index is defined as:

$$SB_{ij} = \begin{Bmatrix} 1, if\ RT_{ij} \geq Threshold_j \\ 0, if\ RT_{ij} < Threshold_j \end{Bmatrix} \qquad (1)$$

       where *i* and *j* represent examinee and item, respectively.
    2. The response-time effort (RTE) index is obtained by aggregating SB values over all items
    3. An RTE of more than 0.85 is recommended for solution behavior

REVIEW OF PREVIOUS RESEARCH

- **Item-level RT sequential effect**
  - Not found in the literature
  - We will conduct exploratory analysis and evaluate whether such an approach is feasible

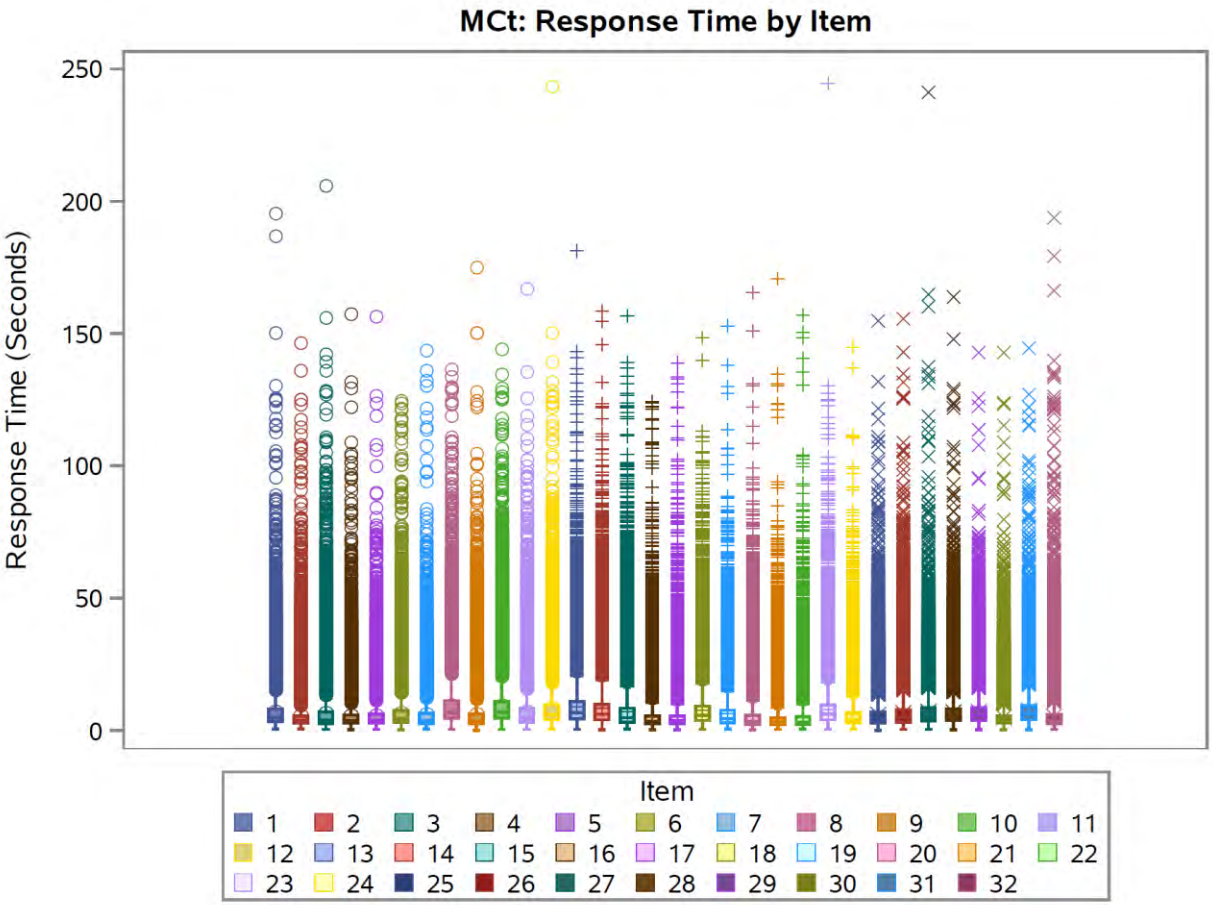
Analysis and Results

ANALYSIS AND RESULTS

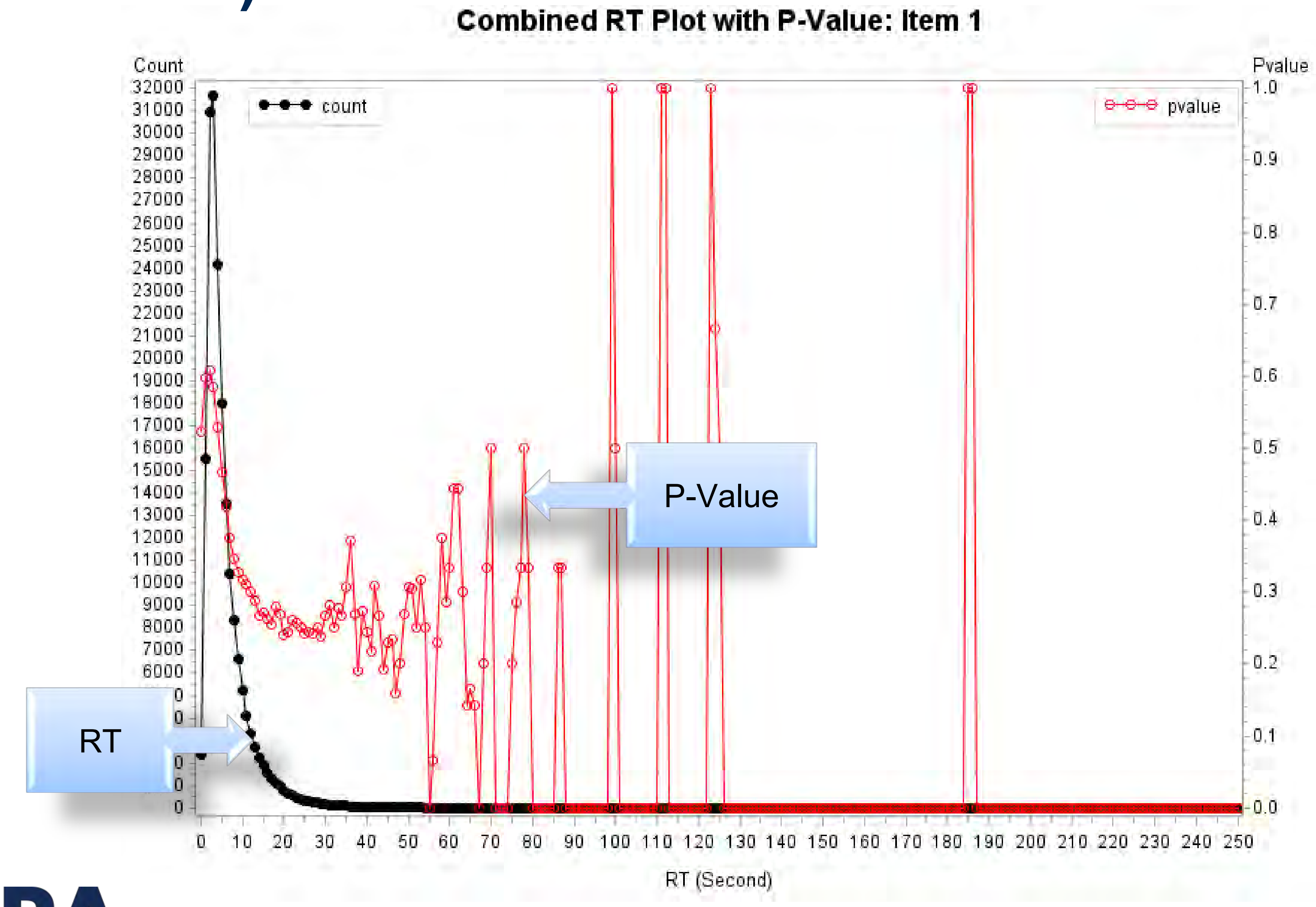
1. **Response time (RT) and RT/Accuracy distribution**
2. **Test-level aggregated index: RTE**
   - How is the floor effect after removal?
3. **Item-level sequential/order effect**
   - Item RT (ordered by item number) for two groups: is there any observable pattern?
     - MCt RS=0 (RS:raw score): those who are at the floor
     - MCt RS>0: above the floor
   - Item difficulty for two groups: how likely it is to answer an item correctly by guessing?
     - MCt RS=1
     - MCt RS>1

ANALYSIS #1: RT DISTRIBUTION FOR ALL ITEMS

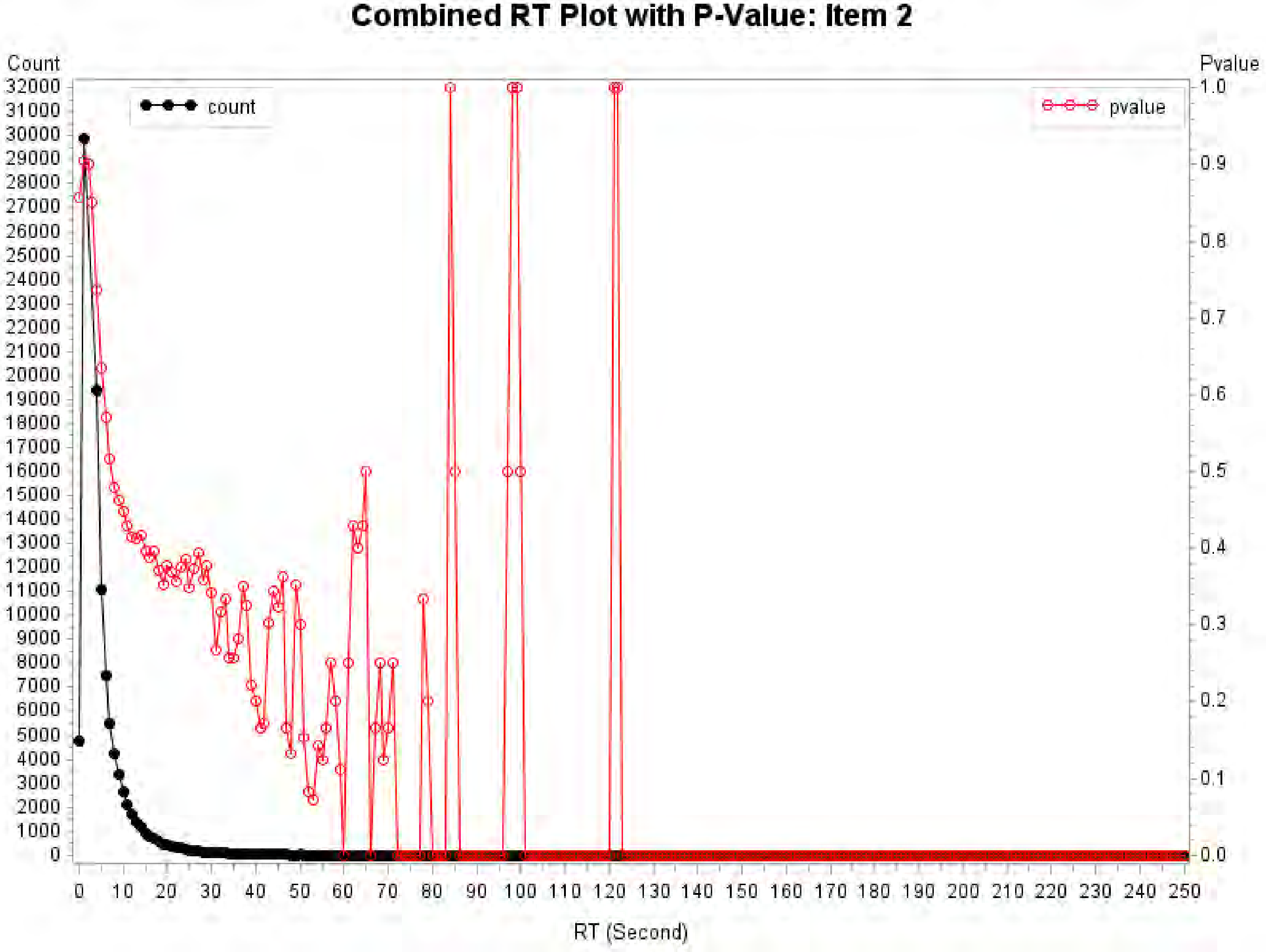
- For most items, the 75Th percentile of RT is less than 10 seconds, which indicates that most examinees spent less than 10 seconds on most items.
- Finding is consistent with the definition of working memory.



ANALYSIS #1: RT AND ACCURACY DISTRIBUTION (EXAMPLE)



ANALYSIS #1: RT AND ACCURACY DISTRIBUTION (EXAMPLE)



ANALYSIS #1: RT AND ACCURACY DISTRIBUTIONS

- **The RT distribution (black line with solid circle) is not bimodal. It is positively skewed with only one mode (the value that appears most often).**
- **For both example items**
  - RT peaks at about 5 seconds (mode=5 seconds)
  - The conditional p-value (red line with empty circle) also peaks at about RT=5 seconds, which means the accuracy is the highest when RT is about 5 seconds (where the mode is)
- **The conditional p-value declines after the first peak (with large fluctuations after 100 seconds), which means spending more time does not lead to more accurate answers**
  - Conditional p-value tends to be 0 when RT is greater than approximately 80 seconds
- **It is not feasible to identify the threshold visually**

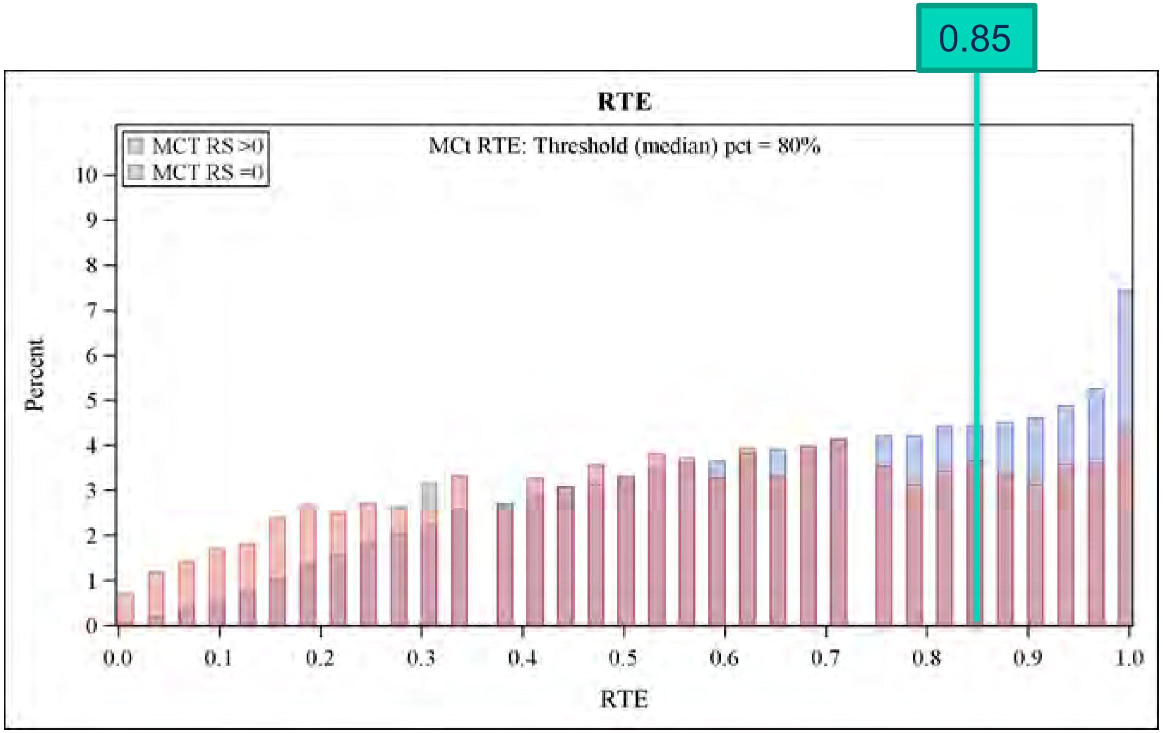
ANALYSIS #2: RTE INDEX

- **RTE is obtained by aggregating SB values over all items**
- **Threshold definition**
  - Mean (10%, 20%, 30%, …, 90%): 9 variations
  - Median (10%, 20%, 30%, …, 90%): 9 variations
- **Modified RTE**
  - Initial analysis showed that sometimes more than 50% of the examinees identified as "rapid guessing" based on the RTE
    - MCt is a WM test. RT is very short.
    - An examinee can spend less time on an item but still get the item correct
    - Spending more time on an item does not always lead to more accuracy
  - Modified RTE: SB = 0 if RT<threshold and response is incorrect
  - A total of 9+9=18 variations of modified RTE
  - Only selected results will be presented next

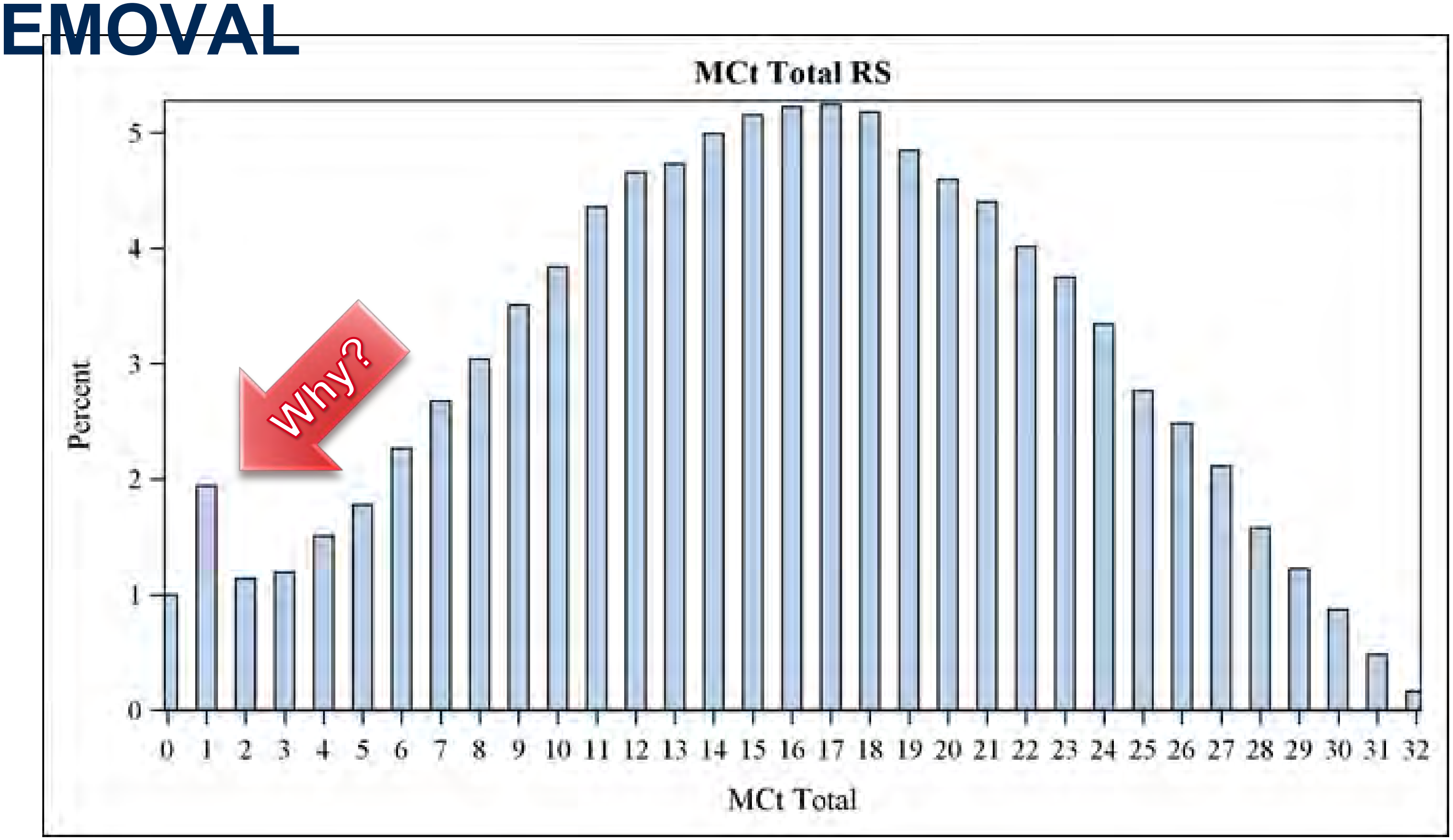
Threshold = 80% of Median RT with RTE≥0.85

ANALYSIS #2: RTE DISTRIBUTION

- **For RS=0 and RS>0, the value of RTE can be greater or less than 0.85.**
  - An examinee can be engaged in the solution behavior and still have a total RS of 0
  - An examinee can also “random guess” and receive a RS>0
- **The RTE distributions are somewhat similar for MCt RS=0 and RS>0 except for the tails**
  - More examinees with RS>0 when RTE≥0.85



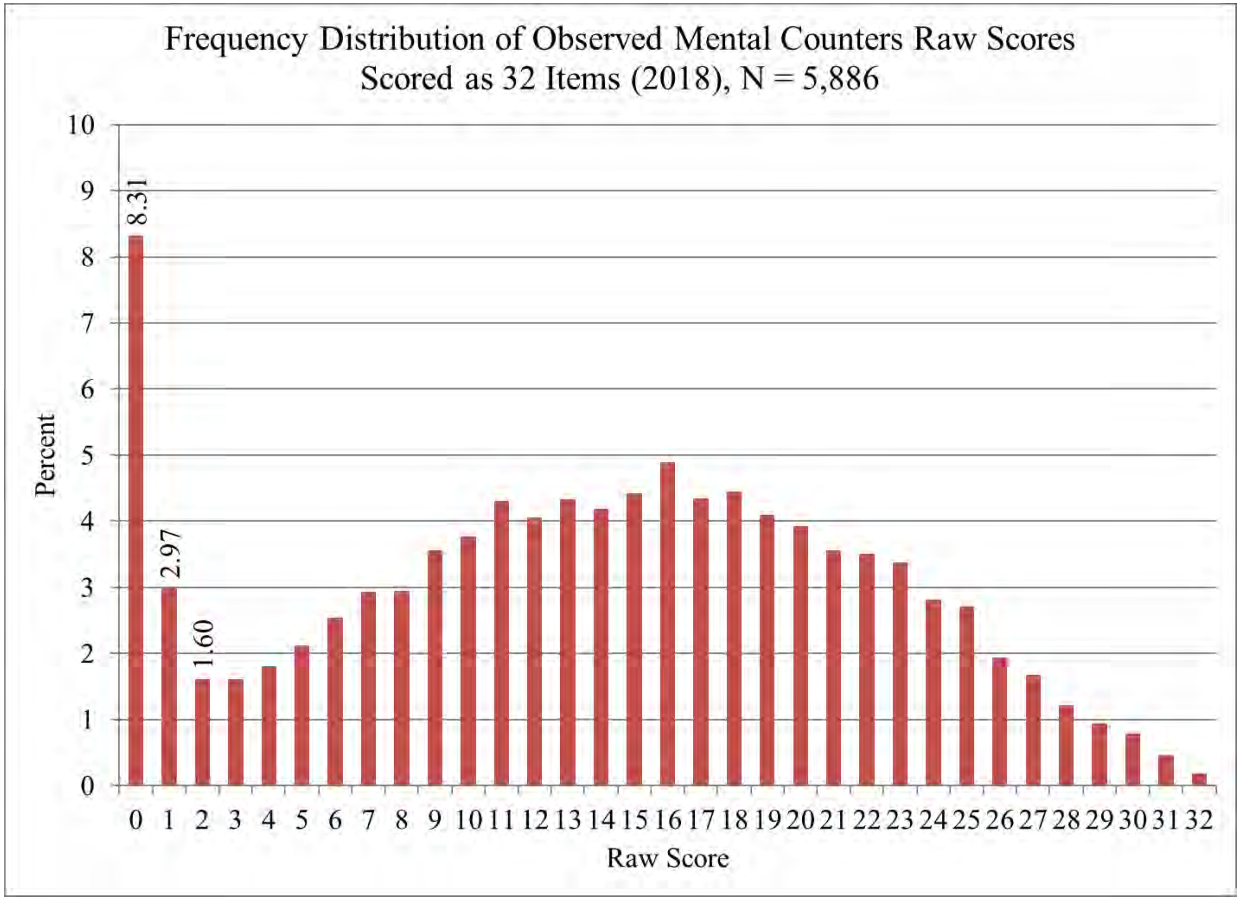
ANALYSIS #2: MCT RS DISTRIBUTION AFTER REMOVAL



- After removal, about 96% of the original data remains
- The MCt total raw score distribution looks relatively normal: floor effect for RS=0 is significantly reduced
- However, RS=1 seems to stand out (secondary floor)

ANALYSIS #2: WHY IS THERE A SPIKE AFTER REMOVAL?

- **The percent for RS=1 is the second highest in the original data. 2.97% of examinees answered only one item correctly.**
- **The modified RTE method reduced the number of examinees with RS=0.**
- **The aggregated RT index has some limitations.**
- **What else is going on?**



ANALYSIS #2: CORRELATION BETWEEN MCT RS AND ASVAB SCORES

	Historical (ECAT/v2.0)	Before Removal	After Removal
General Science (GS)	0.368	0.397	0.356
Arithmetic Reasoning (AR)	0.558	0.517	0.492
Word Knowledge (WK)	0.341	0.363	0.323
Paragraph Comprehension (PC)	0.353	0.401	0.356
Auto and Shop Information (AS)	0.209	0.279	0.249
Mathematics Knowledge (MK)	0.516	0.432	0.406
Mechanical Comprehension (MC)	0.426	0.483	0.447
Electronics Information (EI)	0.269	0.365	0.327
Assembling Objects (AO)	0.570	0.392	0.362
Armed Forces Qualification Test (AFQT)	0.44 (32) or 0.46 (96)	0.522	0.483

- **Historical**
  - subtest-MCt correlation is based on the 40-item ECAT
  - AFQT-MCt correlation is based on Version 2.0 (scored as 32 or 96 as the total score)

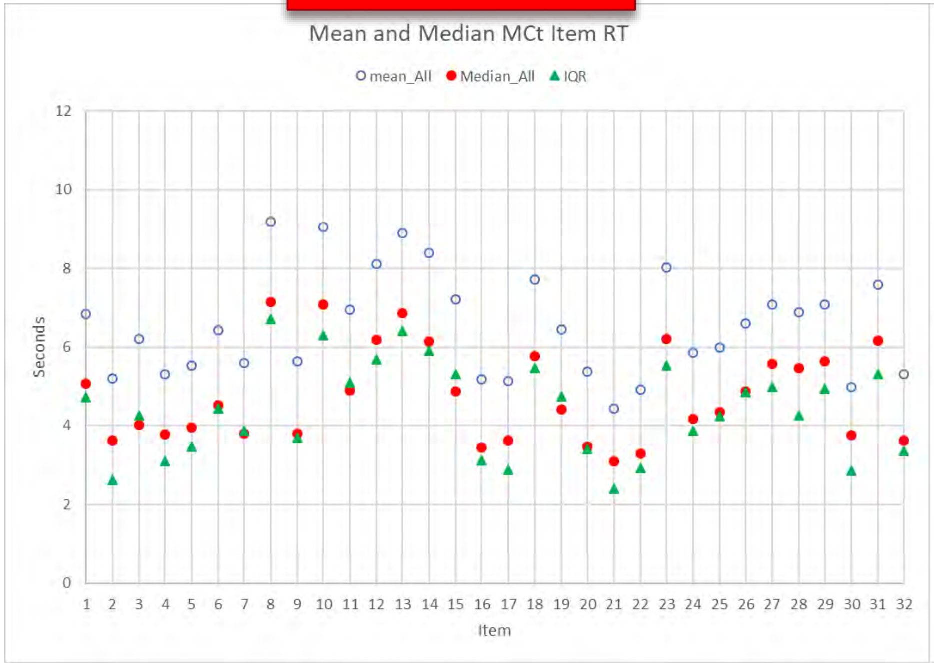
OFFICE OF PEOPLE ANALYTICS

ANALYSIS #2: CORRELATION BETWEEN MCT RS AND ASVAB SCORES

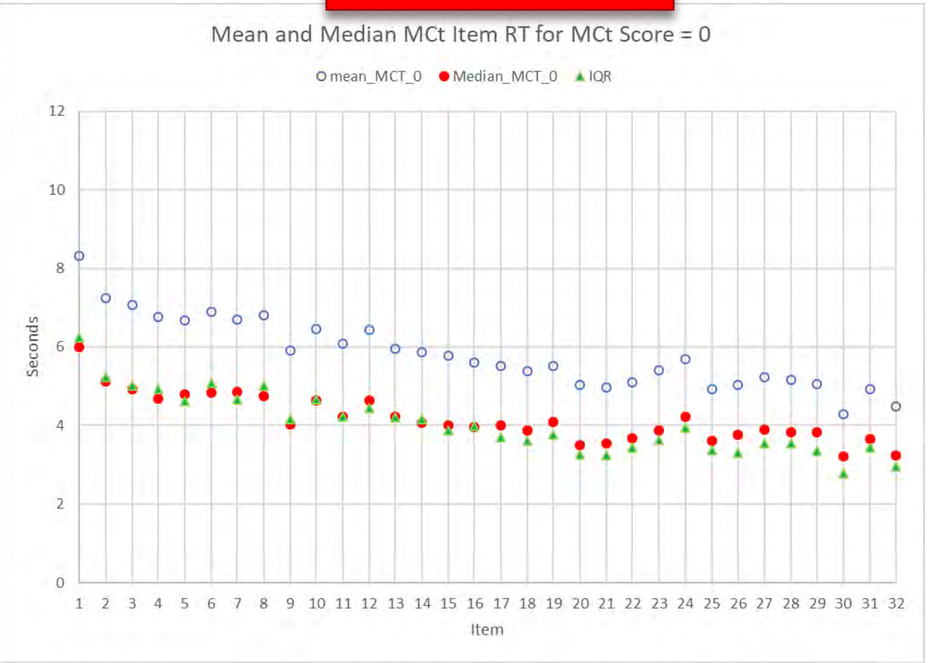
- **Correlation coefficients from this analysis is comparable to historical values**
- **MCt RS correlates moderately with AFQT scores**
  - Around 0.5
- **MCT RS correlates slightly lower with ASVAB subtests**
  - **Highest with Arithmetic Reasoning (AR)**
  - **Lowest with Auto and Shop Information (AS)**
- **Correlation reduced slightly after removal based on RTE**

ANALYSIS #3: ITEM-LEVEL RT FOR ALL AND RS=0

All Examinees



RS=0

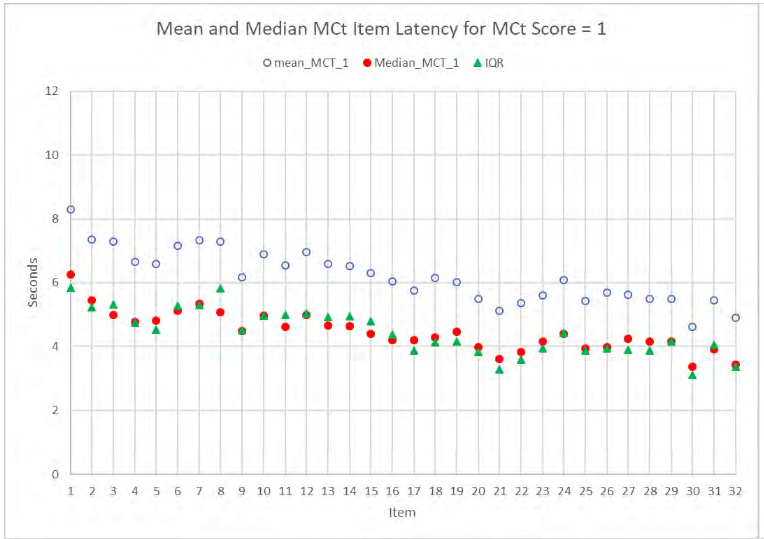


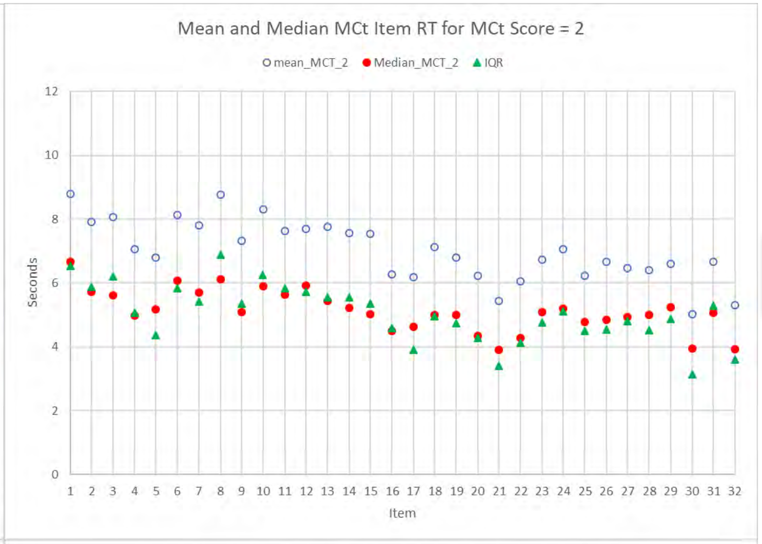
- **RT distribution (mean, median, IQR or interquantile range) for all examinees is largely random; mostly likely associated with the item design (delay and number of adjustments).**
- **For RS=0:**
  - Mean, median, IQR decrease as item number increases
  - Examinees answered items toward the end of the test more quickly than those items at the beginning of the test
  - A clear pattern indicates rushing toward the end

OPA
OFFICE OF PEOPLE ANALYTICS

ANALYSIS #3: ITEM-LEVEL RT FOR MCT RS=1, 2, 3, 10

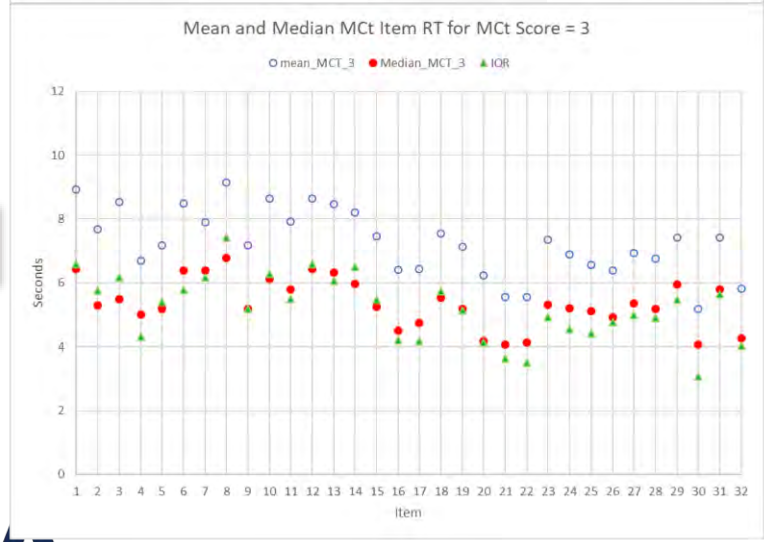
1

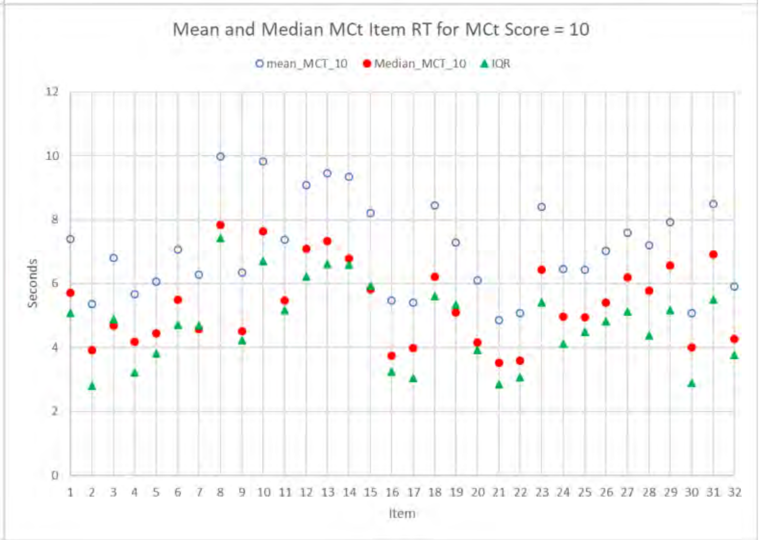

Mean and Median MCt Item Latency for MCt Score = 1


Mean and Median MCt Item RT for MCt Score = 2

2

3


Mean and Median MCt Item RT for MCt Score = 3

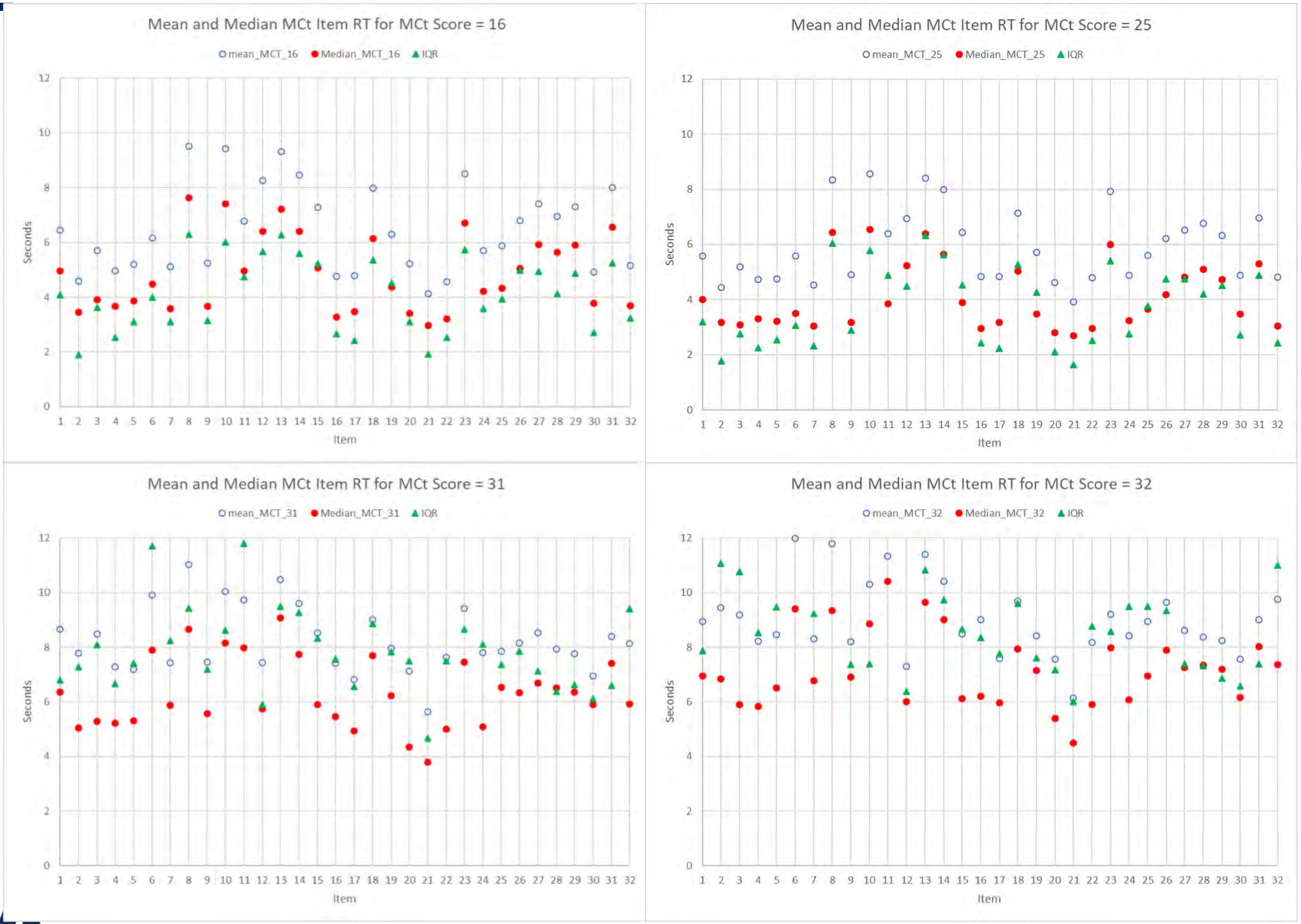

Mean and Median MCt Item RT for MCt Score = 10

10

ANALYSIS #3: ITEM-LEVEL RT FOR MCT RS=16, 25, 31, 32

16

25



31

32

OPA
OFFICE OF PEOPLE ANALYTICS

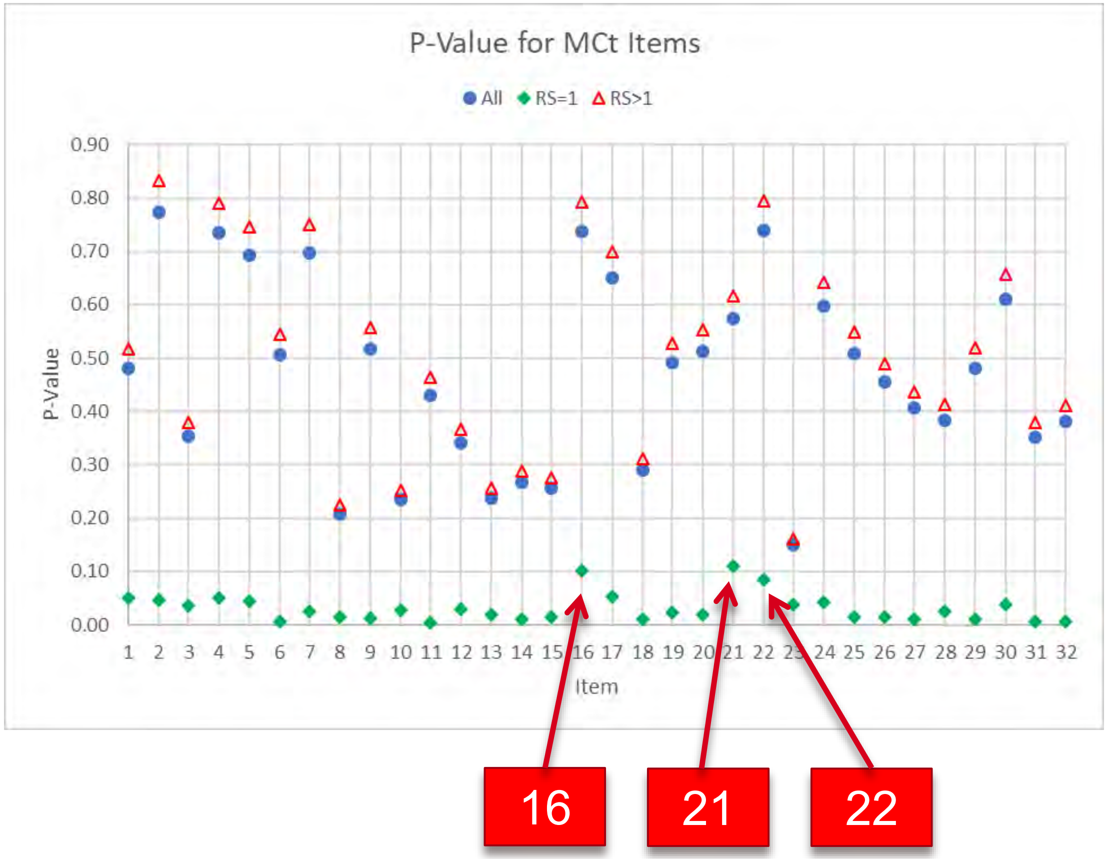
ANALYSIS #3: ITEM-LEVEL RT DISTRIBUTION

- **Item RT distribution should be mostly random (all examinees)**
  - Related to item design
- **There is a clear trend for RS=0, 1, 2, and probably 3**
  - As item number increases, the mean/median/IQR decreases
  - Are they running out of time?
    - For RS=0, mean total RT=**186** seconds; median total RT=**162** seconds
    - Very unlikely: MCt allows 30 minutes
- **For most RT conditional on RS, the RT statistics (mean, median, IQR) are between 2 and 10**
  - Median and IQR are somewhat similar
    - The **interquartile range** (**IQR**) is the difference between the upper (Q3) and lower (Q1) quartiles, and describes the middle 50% of values when ordered from lowest to highest.
- **For RS=31 and 32**
  - RT statistics (mean, median, IQR) are between 4 and 12
    - Examinees with higher MCt scores tend to spend slightly more time (2 seconds)
    - IQR is greater than median (more variability)

OFFICE OF PEOPLE ANALYTICS

ANALYSIS #3: ITEM-LEVEL P-VALUES FOR ALL, RS=1 AND RS>1

- **Item P-value for examinees with total RS=1 is very low (<0.1) compared to P-values for all examinees**
- **No obvious pattern for extremely easy items for RS=1**
- **For examinees with RS=1, items 16, 21, and 22 are the easiest (with P-values around 0.1)**
  - The three items have 830ms with 6 adjustments (moderately easy).



OFFICE OF PEOPLE ANALYTICS

ANALYSIS #3: TOP FIVE RESPONSE PATTERNS (RS=1)

	Item 16			Item 21			Item 22		
	Response	N	Percent	Response	N	Percent	Response	N	Percent
Key	456	181	10.16	777	196	11.00	456	150	8.42
	556	54	3.03	666	75	4.21	345	82	4.60
	567	46	2.58	888	36	2.02	455	79	4.43
	777	45	2.53	767	32	1.80	777	41	2.30
	455	40	2.24	555	32	1.80	567	39	2.19

- **Based on the key and response pattern, it is unlikely examinees can answer these items correctly by chance/guessing alone:**
  - The probability of answering a MCt item correctly by guessing alone is: (1/10)x(1/10)x(1/10)=0.001
- **However, it is possible that examinees guess on one or two of the numbers, which could increase this probability**
- **Next step: we will evaluate response patterns for signs of guessing—Does the same examinee use the same response pattern for all items?**

SUMMARY OF RESULTS

- **Is it feasible to identify examinees who are "not trying" on MCt using RT distribution? NO**
  - RT distribution for MCt is highly skewed with only one mode
  - RT is very short: the 75th percentile is usually less than 10 seconds
  - Cannot visually identify threshold
- **Is it feasible to identify examinees who are "not trying" on MCt using the aggregated index (RTE)? Somewhat, but not completely**
  - RTE needs to be modified:
    - If not adding the requirement of incorrect responses for RTE, more than 50% of examinees will be identified as "random guessing"
  - After removing examinees identified as "random guessing" with the modified RTE:
    - The floor effect is reduced; however, there is a "secondary" floor for RS=1

SUMMARY OF RESULTS (CONTINUED)

- **Is it feasible to identify examinees who are "not trying" on MCt by examining the item-level sequential effect? Highly likely, but it is difficult to implement at the individual level. An index to quantify the trend would be helpful.**
  - Item RT (ordered by item number)
    - There is a clear trend for RS=0, 1, 2, and probably 3: as item number increases, the mean/median/IQR decreases
    - The trend is not caused by the lack of time → Its very likely they are not trying!
      - ✓ Time allowed is 30 minutes
      - ✓ For RS=0, the average total latency for all 32 items is about 3 minutes.
  - Item difficulty for RS=1
    - It is very difficult to answer one MCt item correctly by guessing alone, but it is also possible that examinees guess on one or two of the numbers, which could increase this likelihood

OFFICE OF PEOPLE ANALYTICS

FUTURE RESEARCH

- **Evaluate response patterns**
  - Statistically model or predict responses based on responses to previous items
- **Continue to evaluate the sequential effect displayed in item-level RT**
  - Develop an index to quantify the sequential effect observed in item-level RT (e.g., a l_z-like statistic) and examine the floor effect after applying the new index
- **Additional research based on aggregated RTE**
  - Evaluate differential item information (DIF) of random guessing and solution behaviors
  - Additional modification of the RTE to remove the secondary floor
- **Evaluate whether instruction plays a role in the floor effect**
  - Results from this analysis suggest that guessing seems to play a role for those with very low MCt scores (RS=0, 1, 2, and possibly 3)
  - Next question: Can misunderstanding the MCt instructions play a role in low-scoring, but motivated examinees?

Next: Think-Aloud

Thank you!
Questions?
Comments?

OFFICE OF PEOPLE ANALYTICS

Tab J

OPA

OFFICE OF PEOPLE ANALYTICS

Mental Counters

Plans for Using the Think-Aloud Method to Evaluate Test Instructions

Presented to the DACMPT

Ping Yin, HumRRO

Gregory Manley, DPAC

Mary Pommerich, DPAC

March 28, 2019 | Carmel-By-The-Sea, CA

OVERVIEW

- Brief introduction
- What is think-aloud?
- Think-aloud research questions
- Study design
- Recommendations

BRIEF INTRODUCTION

- **Possible factors contributing to the floor effect:**
  1. Not able to understand the task
  2. Lack of motivation
  3. Too difficult
  4. Fatigue and/or frustration
  5. Combinations of various factors above
- **The previous presentation suggested that some examinees were “not trying”**
- **Some examinees may be eager to try, but still did poorly on the MCt test (at the floor or near floor)**
  - Can misunderstanding in MCt instructions play a role?
  - Think-aloud
    - We presented the idea of a think-aloud study during the 2018 MAPWG and DAC. DAC thought it was a good idea and was very supportive. Think-aloud will allow further investigation of test-takers’ understanding of MCt instructions.

WHAT IS THINK-ALOUD?

- Think-aloud is a research method that systematically collects validity evidence of response processes (Ericsson & Simon, 1980).
- In a typical think-aloud, participants speak aloud any thoughts in their mind as they complete a task.
- It is widely used in usability testing, education, and related fields to see how people approach tasks.
- Can be used to learn how participants think about tasks and identify common misconceptions.

THINK-ALOUD RESEARCH QUESTIONS

- **Are the MCt instructions clear?**
- **Are the MCt instructions easy to understand?**
- **Are the MCt instructions user-friendly?**
- **Is it possible to simplify and streamline the MCt instructions?**
- **Identify areas in MCt instructions that may potentially contribute to the floor effect:**
  - Confusion or misunderstanding
  - Lack of motivation
  - Too difficult
  - Mental fatigue
  - Frustration

STUDY DESIGN FOR THE THINK-ALOUD

- **The ideal study**
  - We are aware that there will be constraints, but rather than focusing on the constraints and artificially limit the scope, we will start with the ideal study
- **The study with constraints**
  - We will discuss the constraints we anticipate and implications of the constraints
- **A hybrid study**
  - Variations of study designs between the best and the worst
- **Next, we will focus on the ideal study and the study with constraints**

THE IDEAL STUDY

- **Random group design**
  - The most powerful design without introducing potential bias
  - Randomly equivalent groups of participants
  - Each group will take one of the two types of MCt instructions (current, updated)
- **Subjects**
  - Representative samples from the MCt target population (applicants)
  - Age, gender, race/ethnicity, level of education, and other relevant demographic characters
- **Sample size**
  - The "rule of thumb" for a boundary between small and large samples is between 25 and 30 (Hogg, Tanis, & Zimmerman, 2015)
  - A minimum sample size of 25 is required for each group

THE IDEAL STUDY (CONTINUED)

- **Think-aloud data collection**
  - Each participant will be scheduled for an individual session in a **quiet** setting (for talking and audiotaping)
  - Script and questionnaires will be provided (one questionnaire for each section, and a final/exit questionnaire for the overall experience); see draft
  - Steps and estimated amount of time for the think-aloud (between 1 and 1 ½ hours)

		Estimated Amount of Time in Minutes (Average)
Think-Aloud for MCt Instructions	1. Introduction/Housekeeping	5
	2. MCt Introduction Think-Aloud	10
	3. MCt Introduction Questionnaire	5
	4. MCt Demonstration Think-Aloud	10
	5. MCt Demonstration Questionnaire	5
	6. MCt Practice Items Think-Aloud	15
	7. MCt Practice Items Questionnaire	5
	8. MCt Overall Questionnaire	10
	Total	**65**

THE IDEAL STUDY

- **One step further: have another two random groups of participants take the actual MCt test, but with different MCt instructions**
  - Requires four randomly equivalent groups to avoid possible contamination/influence of the think-aloud on the participants

	Current MCt Instructions	Updated MCt Instructions
Think-Aloud Only	1	2
MCt Test Only	3	4

  - Observe test-takers' behavior during the test (rushing, motivation level, fatigue, etc.)
  - Possible hypotheses testing
    - Level of motivation and fatigue on MCt scores
    - Updated instruction is more effective → higher MCt score

THE STUDY WITH CONSTRAI🙁

- **Single-group design**
  - Each participant takes both MCt instructions
  - Problematic:
    - Order effect: the possibility that the order of current and updated instructions matters
    - Sequence effect: the possibility that either instruction will be affected by the instruction preceding it
  - Counter-balancing is often recommended to reduce these confounding effects
  - However, given the commonality between the two instructions (both for the same Mental Counters test), it is difficult to avoid such confounding even with counter-balancing
  - Will negatively impact the interpretations of the results
  - Is not recommended

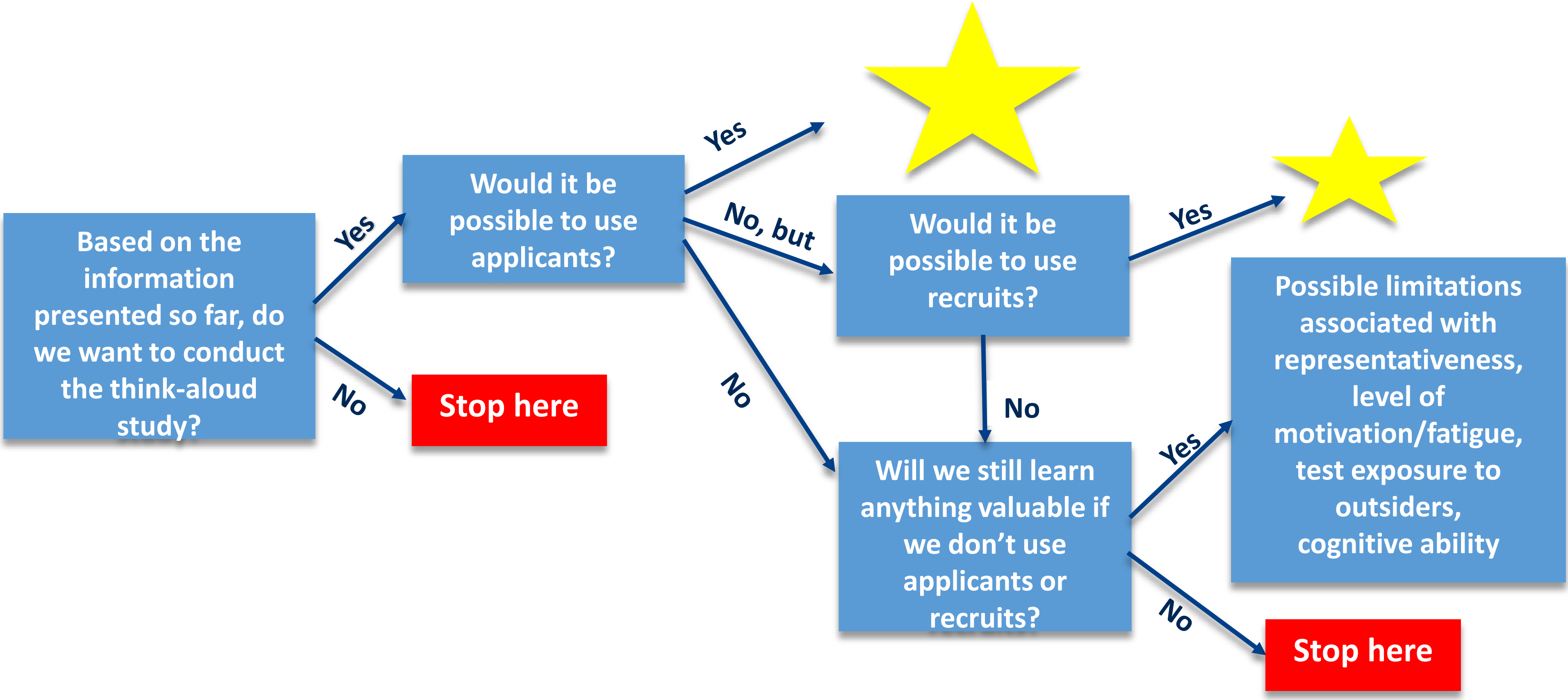
THE STUDY WITH CONSTRAINTS (CONTINUED)

- **Subject**
  - Not representative, but convenient sample (HumRRO? PERSEREC? OPA?)
  - May differ in age, gender, race/ethnicity, level of education, cognitive ability, familiarity with MCt, and other relevant characters, which will likely bias or interfere with their think-aloud
  - Motivation will be different
  - Results cannot be generalized to the MCt target population
- **Sample size**
  - Small sample size will impact the statistical power (random-group design)
  - In the extreme scenario of very few participants, focusing on the updated instruction will be the most beneficial (recommended at 2018 DAC)
    - Qualitative data
- **Data collection**
  - Single-group design will require more time per participant

THE STUDY WITH CONSTRAINTS (CONTINUED)

- **Platform requirement and location**
  - The current MCt instructions are provided only on a DOD desktop (only available at MEPS or CATLAB)
  - The think-aloud requires a **quiet** environment
- **Sequencing for the updated instructions**
  - Practice items for the MCt are available only on a DOD desktop and not integrated with the updated instructions
  - The updated sequence (e.g., looping back to demonstration after failing both easy practice items) has not been implemented
- **Recommend easier practice items for the updated instructions**
  - Not yet implemented on the laptop

THINK-ALOUD STUDY DESIGN POSSIBILITIES



Based on the information presented so far, do we want to conduct the think-aloud study?

Stop here

Would it be possible to use applicants?

Would it be possible to use recruits?

Will we still learn anything valuable if we don't use applicants or recruits?

Possible limitations associated with representativeness, level of motivation/fatigue, test exposure to outsiders, cognitive ability

Stop here

RECOMMENDATIONS

- **We have gained a lot more insight into the floor effect through the analysis of item-level response times. We will continue to refine the techniques for identifying test-takers who are not trying, but we also recommend a think-aloud study.**
  - We plan to develop an item-level index to quantify the observed pattern in item response time for those who were "not trying" and evaluate the floor effect after implementing this new index
  - Given the constraints, it seems more feasible to conduct a **small-scale, in-house convenience-sample** think-aloud study (pilot), with a focus on collecting **qualitative** data
    - We are fully aware of the limitations and will make careful decisions on how to interpret the findings
    - The pilot think-aloud can still provide useful information on test-takers' understanding of the MCt instructions despite the limitations
  - We will then re-evaluate the need for additional think-aloud studies based on the new findings, and may consider the possibility of using applicants and recruits for a more "ideal" study in the future

Thank you!
Questions?
Comments?

Tab K

OFFICE OF PEOPLE ANALYTICS

Evaluation of Air Force Cyber Test CAT Pools

Presented to the DAC

Furong Gao, HumRRO
Mary Pommerich, DPAC
March 28, 2019 | Carmel-by-the-Sea, CA

OUTLINE

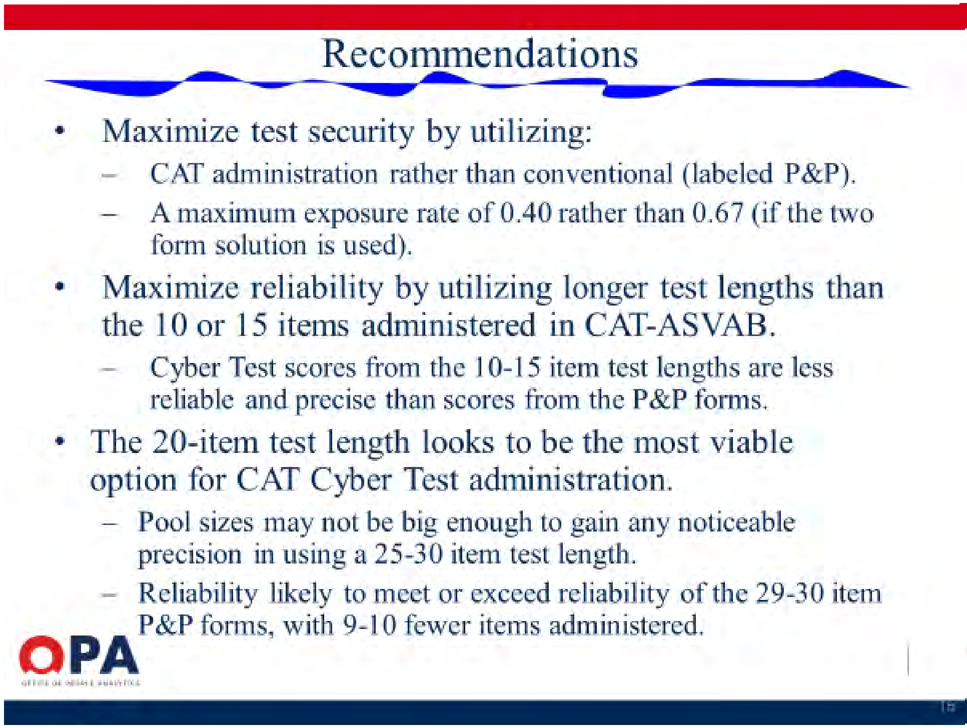
- **Background**
- **Dimensionality assessment**
  - Confirmatory analysis of uni-dimensionality
- **CAT simulation**
  - Score information function
  - Pool item usage
  - Test-retest reliability
- **Discussion and recommendations**

BACKGROUND

- **To evaluate the feasibility of administering the Cyber Test in a CAT framework**
- **Currently (since 2011), two operational 29-item static forms administered on computer**
- **Last evaluation: 2016**
  - CAT pools constructed using automated test assembly, with 166 items selected from 58 items (from the operational forms) + 190 then newly tried-out and calibrated items
  - A two-parallel (roughly) form/pool solution and a three-pool solution were evaluated by simulation
    - SIF, item usage, test-retest reliability
  - Resulted in decision to develop more items that target the low and middle range of the ability distribution
- **New CAT pools: 2018**
  - New two-pool and three-pool solutions to be evaluated, with items selected from 166 prior pool of items + 242 newly calibrated items
    - Two-pool: each pool contains 130 unique items
    - Three-pool: each pool contains 87 unique items
- **Item enemies contained across pools but not within pools**

RECOMMENDATIONS FROM PREVIOUS EVALUATION (2016)

- **Two-pool (83 items/pool) and three-pool (55 items/pool) solutions**

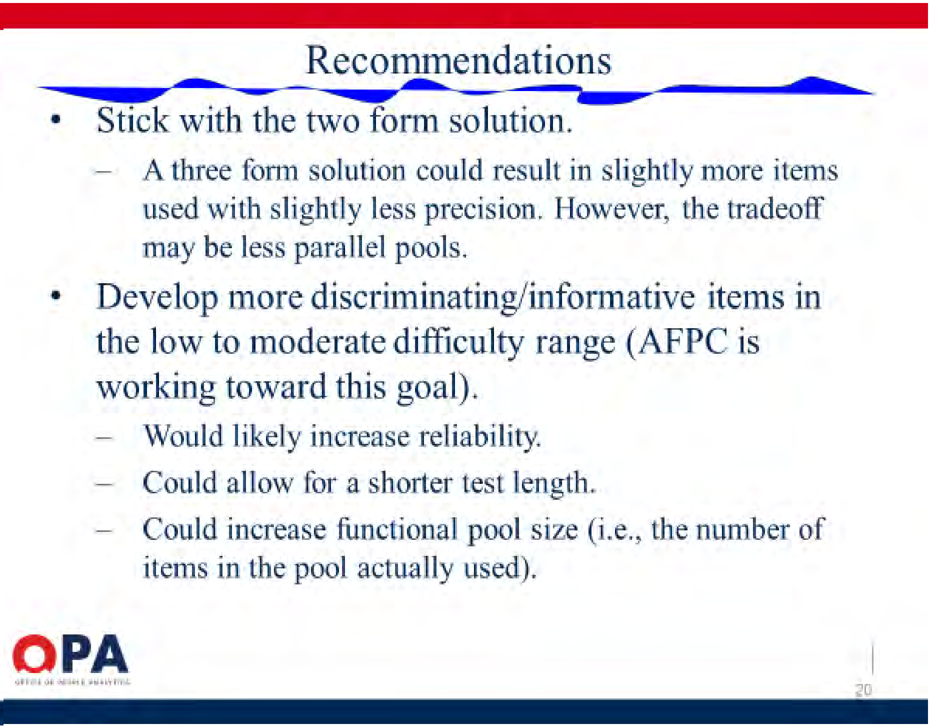


Recommendations

- Maximize test security by utilizing:
  - CAT administration rather than conventional (labeled P&P).
  - A maximum exposure rate of 0.40 rather than 0.67 (if the two form solution is used).
- Maximize reliability by utilizing longer test lengths than the 10 or 15 items administered in CAT-ASVAB.
  - Cyber Test scores from the 10-15 item test lengths are less reliable and precise than scores from the P&P forms.
- The 20-item test length looks to be the most viable option for CAT Cyber Test administration.
  - Pool sizes may not be big enough to gain any noticeable precision in using a 25-30 item test length.
  - Reliability likely to meet or exceed reliability of the 29-30 item P&P forms, with 9-10 fewer items administered.

OPA OFFICE OF PEOPLE ANALYTICS

15



Recommendations

- Stick with the two form solution.
  - A three form solution could result in slightly more items used with slightly less precision. However, the tradeoff may be less parallel pools.
- Develop more discriminating/informative items in the low to moderate difficulty range (AFPC is working toward this goal).
  - Would likely increase reliability.
  - Could allow for a shorter test length.
  - Could increase functional pool size (i.e., the number of items in the pool actually used).

OPA OFFICE OF PEOPLE ANALYTICS

20

*see ASVAB Tech#1 pages 2-29/30

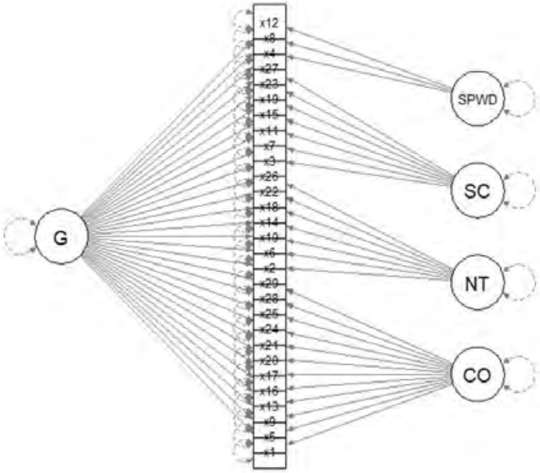
DIMENSIONALITY ASSESSMENT

- **IRT model-based item factor analysis**
  - Software: iFACT (Segall, 2002)
    - MCMC
- **Assumptions**
  - Test is designed to be uni-dimensional: measure a single construct but with broad content coverages that may introduce minor additional unintended dimensions to the test data
  - Items are rendered so that the “missingness” in the response data is missing completely at random (MCAR) or missing at random (MAR)
    - Both the CAT-ASVAB and the currently seeded item design produce MCAR data
- **Confirmatory analyses**
  - Data will be fit with both a one-factor model and a bi-factor model
  - Bi-factor model
    - One general factor (dimension) that all items have loadings on (G)
    - Group (secondary) factors, one for each of the content sub-domains
    - All factors are independent of each other

CYBER TEST

BI-FACTOR MODEL

- **A test of information and communications technology literacy**
  - General factor (G)
- **Four broad content areas—secondary factors**
  - Computer operations (CO)
  - Networks and telecommunications (NT)
  - Security and compliance (SC)
  - Software programming & Web development (SPWD)
- **All factors are independent of each other**
- **Item factor loading:**
  - SPWD items: $(\lambda_g, \lambda_{spwd}, 0, 0, 0)$
  - SC items: $(\lambda_g, \ 0, \ \lambda_{sc}, 0, 0)$
  - NT items: $(\lambda_g, 0, 0, \lambda_{nt}, 0)$
  - CO items: $(\lambda_g, \ 0, 0, 0, \lambda_{co})$



UNI-DIMENSIONALITY ASSESSMENT

- **One-factor and bi-factor comparison**
  - The G-factor loadings of the two models are compared
  - Small and negligible differences are expected
    - ❖ There is a small role of specific/group factors.
    - ❖ Specific factors don’t distort the meaning of the general factor that is measured generally by all the items on the test.

- **An indicator of essential uni-dimensionality: explained common variances (ECV)**
  - Calculated using the factor loading values of the G factor and the secondary factors of the bi-factor model

$$ECV = \frac{\sum \lambda_g^2}{\sum \lambda_g^2 + \sum \lambda_{s1}^2 + \ \ldots + \sum \lambda_{sk}^2}$$

  - Value is between 0 and 1; strictly uni-dimensional: ECV = 1
  - The larger the ECV, the stronger the uni-dimensionality
    - ❖ 0.9 < ECV, essentially uni-dimensional
    - ❖ 0.7 <= ECV <= 0.9, additional information should be used (subscore, etc.)
    - ❖ ECV < 0.7, evidence of multi-dimensionality
  - To adjust for the standard error of estimates

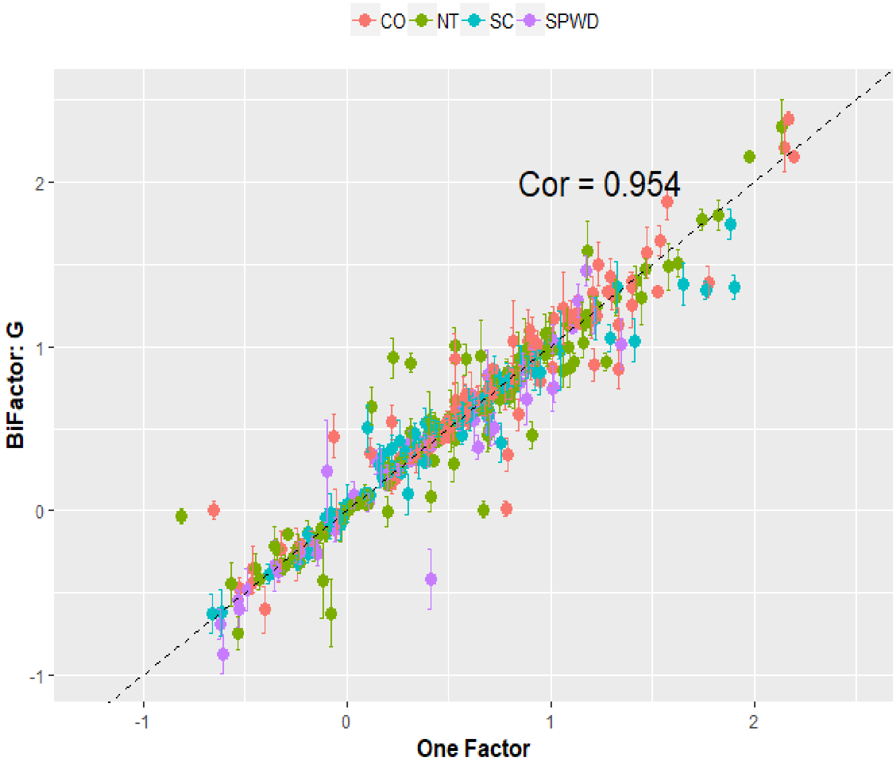
$$ECV_{adj} = \frac{\sum(\lambda_g^2 - e_{\lambda_g}^2)}{\sum(\lambda_g^2 - e_{\lambda_g}^2) + \sum(\lambda_{s1}^2 - e_{\lambda_{s1}}^2) + \ \ldots + \sum(\lambda_{sk}^2 - e_{\lambda_{sk}}^2)}$$

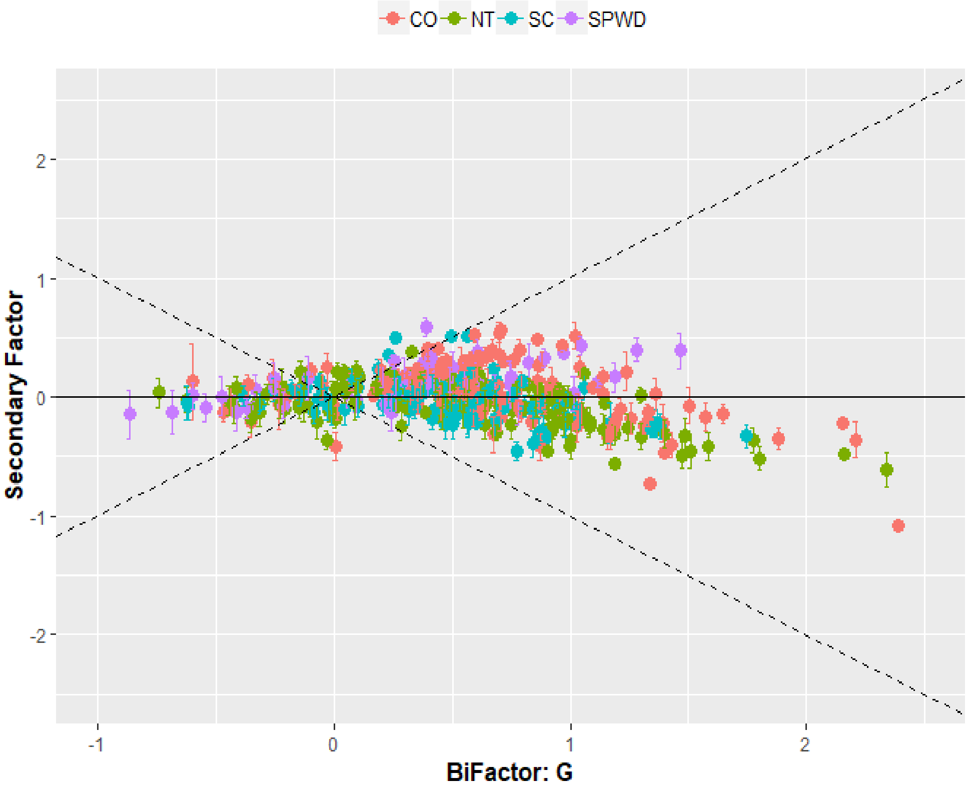
UNI-DIMENSIONALITY ASSESSMENT: CYBER ITEMS

- **Item response data**
  - Operational items
  - Previously seeded items
  - New seeded items
  - Number of items in each content area:
    - CO 142
    - NT 127
    - SC 100
    - SPWD 48
- **108,292 test-takers on Form1; 112,221 on Form2**
- **Case counts on previously seeded items range from 6,493 to 7,678**
- **Case counts on new seeded items range from 3,174 to 3,895**

IFACT RESULTS: 417 ITEMS

- **ECV = 0.925; ECV.adj = 0.938**





CYBER TEST CAT POOLS

- **Item source**
  - Two 29-item operational forms, scaled in 2011 and serve as baseline scale
  - Seed190: developed/scaled/equated in ~2015
  - Seed242: developed/scaled/equated in ~2018
- **Notation (consistent with what waw used in the 2016 evaluation):**
  - Two fixed operational forms: 02A, 03A
  - Two-form pools: 01Z, 02Z
  - Three-form pools: 01Y, 02Y, 03Y
- **Constructed CAT pools:**

CAT Pool Item Source

	Operational			
CAT Form	**02A/03A**	**Seed190**	**Seed242**	**Total**
01Z	19	47	64	**130**
02Z	26	55	49	**130**
Total	***45***	***102***	***113***	***260***
01Y	10	39	38	**87**
02Y	15	37	35	**87**
03Y	19	28	40	**87**
total	***44***	***104***	***113***	***261***

CONTENT DISTRIBUTIONS

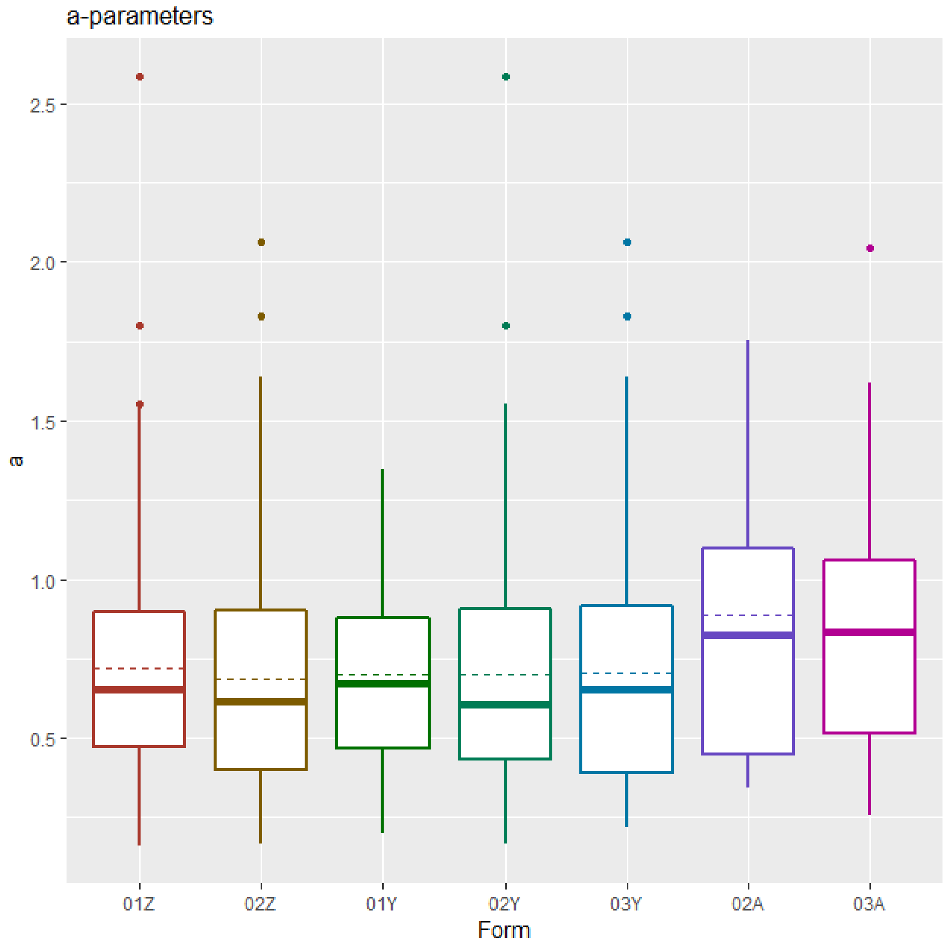
- **Content summary:**

Content Distribution

Content	OP-form	01Z/02Z	01Y	02Y	03Y
CO	12 (41.4%)	49 (37.7%)	33 (37.9%)	32 (36.8%)	34 (39.1%)
NT	7 (24.1%)	34 (26.2%)	23 (26.4%)	23 (26.4%)	23 (26.4%)
SC	7 (24.1%)	33 (25.4%)	22 (25.3%)	22 (25.3%)	21 (24.1%)
SPWD	3 (10.3%)	14 (10.8%)	9 (10.3%)	10 (11.5%)	9 (10.3%)
Total	**29**	**130**	**87**	**87**	**87**

ITEM PARAMETER DISTRIBUTIONS: DISCRIMINATION

- The dashed lines are means
- a-parameters in the CAT pool forms are generally lower than those in the two 29-item operational forms

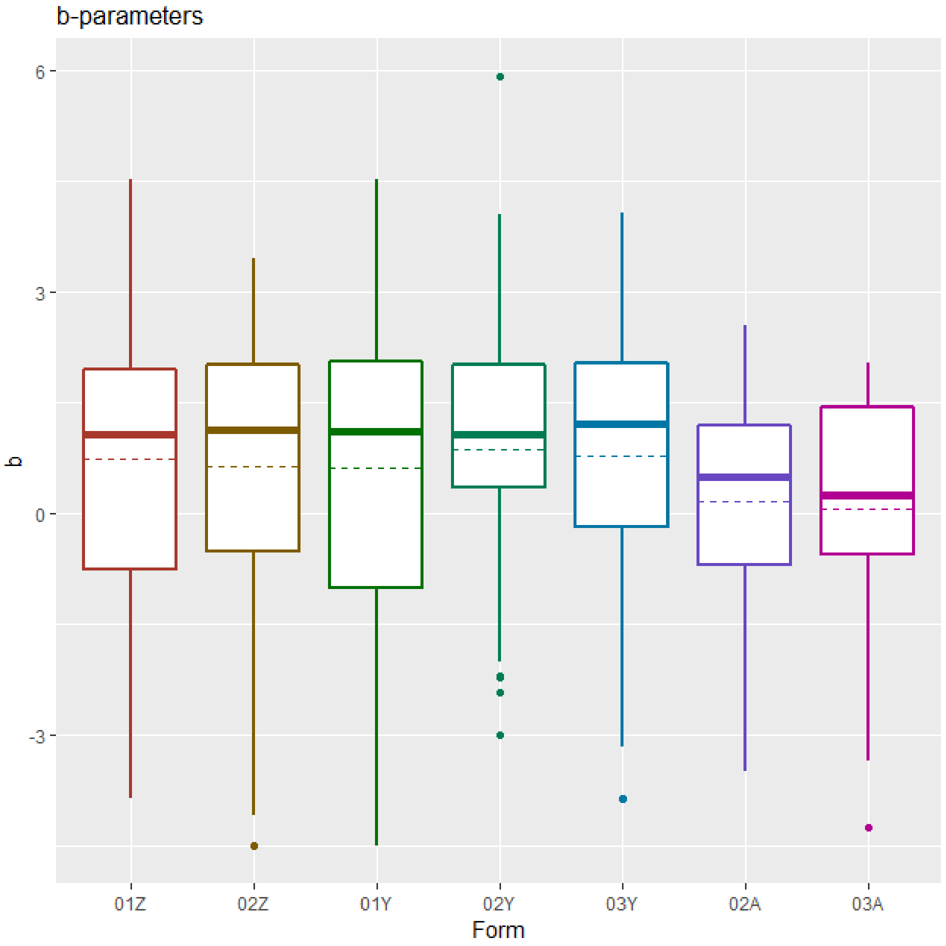


OFFICE OF PEOPLE ANALYTICS

ITEM PARAMETER DISTRIBUTIONS: DIFFICULTY

- b-parameters in the CAT pool forms are generally higher than those in the two 29-item operational forms

- Except for one of the forms in the three-form CAT pool solution, all the difficulty distributions of the CAT pool forms show more spread and have larger IQR (interquartile ranges) than the operational forms

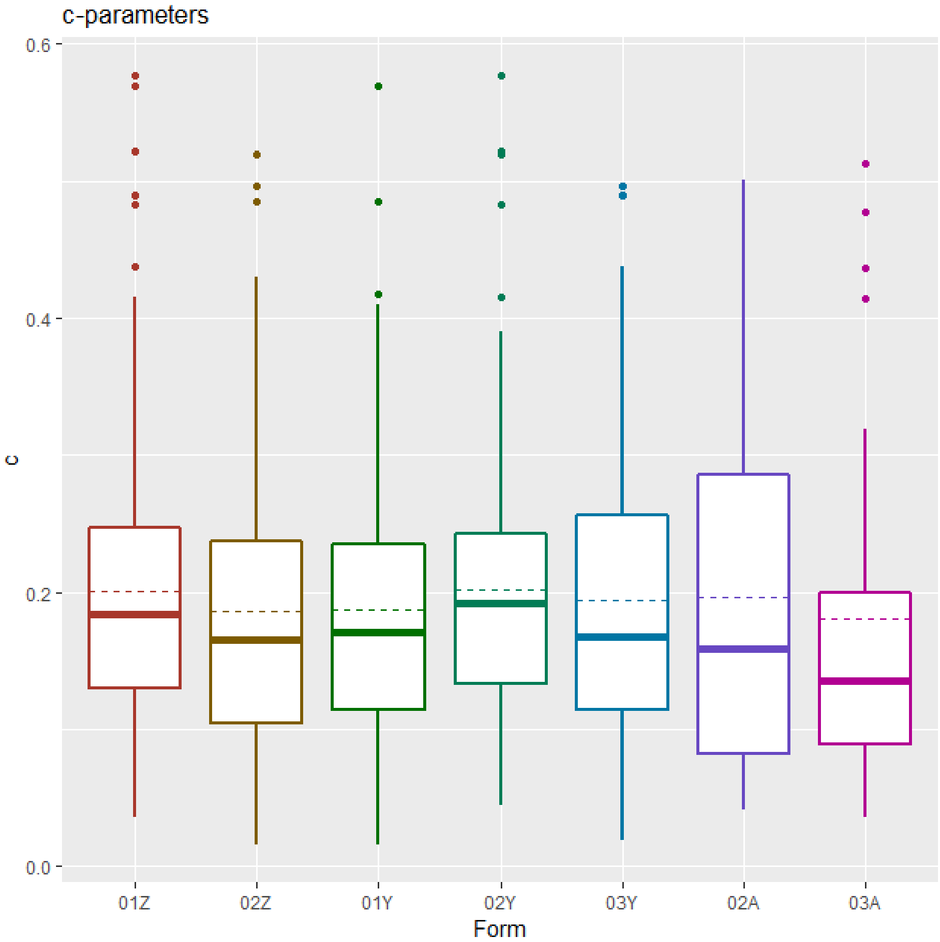


OPA OFFICE OF PEOPLE ANALYTICS

ITEM PARAMETER DISTRIBUTIONS: GUESSING

- Mean values of the c-parameters in the CAT pool forms are similar to those in the two operational forms

- However, the median values are much closer to the means in the CAT pool forms than in the operational forms, indicating less skewness of the distributions



OFFICE OF PEOPLE ANALYTICS

EVALUATION APPROACH

- **Evaluation by simulation using the item pools under CAT-ASVAB administration conditions**
  - Test lengths of 10, 15, 20, 25, and 30 items were used
  - No content constraints were used
    - Supported by the dimensionality assessment findings
  - Target maximum exposure rate of 0.67 was used
    - The target maximum exposure rate of 0.67 (2/3) was selected to match the current maximum exposure rate for all subtests on CAT-ASVAB Forms 5–9
  - Score precision and item usage were evaluated

SCORE INFORMATION FUNCTION

- **CAT pool precision was evaluated using score information function (SIF):**

$$I(\theta) = \frac{\left[\frac{\partial}{\partial\theta}\mu(\hat{\theta}|\theta)\right]^2}{\sigma^2(\hat{\theta}|\theta)}$$

- **SIF was calculated using simulated data***
  - 500 examinees at each of the 31 equally spaced θ values in [-3,3]
  - At each θ, the mean and variance of the 500 scores were calculated, and *I*(θ) was approximated using these results (Lord, 1980, eq. 10–7).

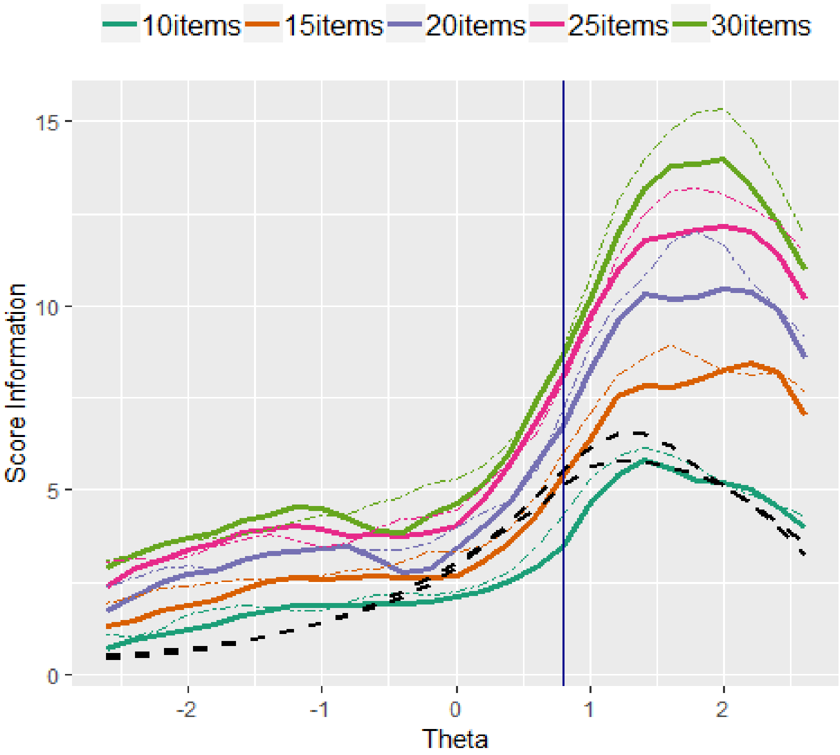
* ASVAB Technical Bulletin #1 pages 2-29/30

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

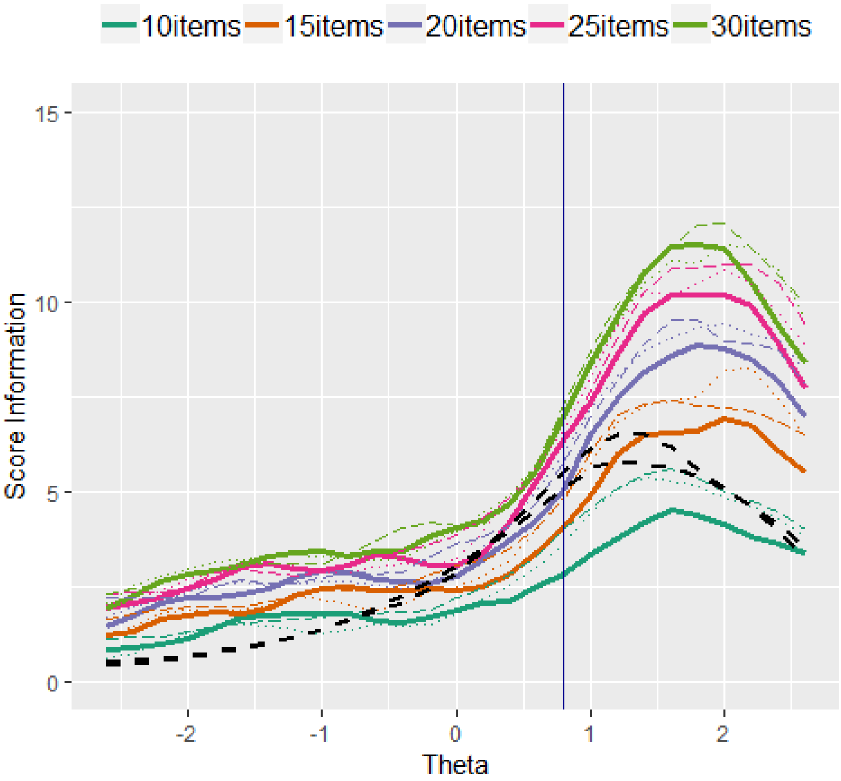
SCORE INFORMATION FUNCTION

- Higher score precision from the two-form solution
- Higher score precision than the pools used in the 2016 evaluation

Two-Form pools

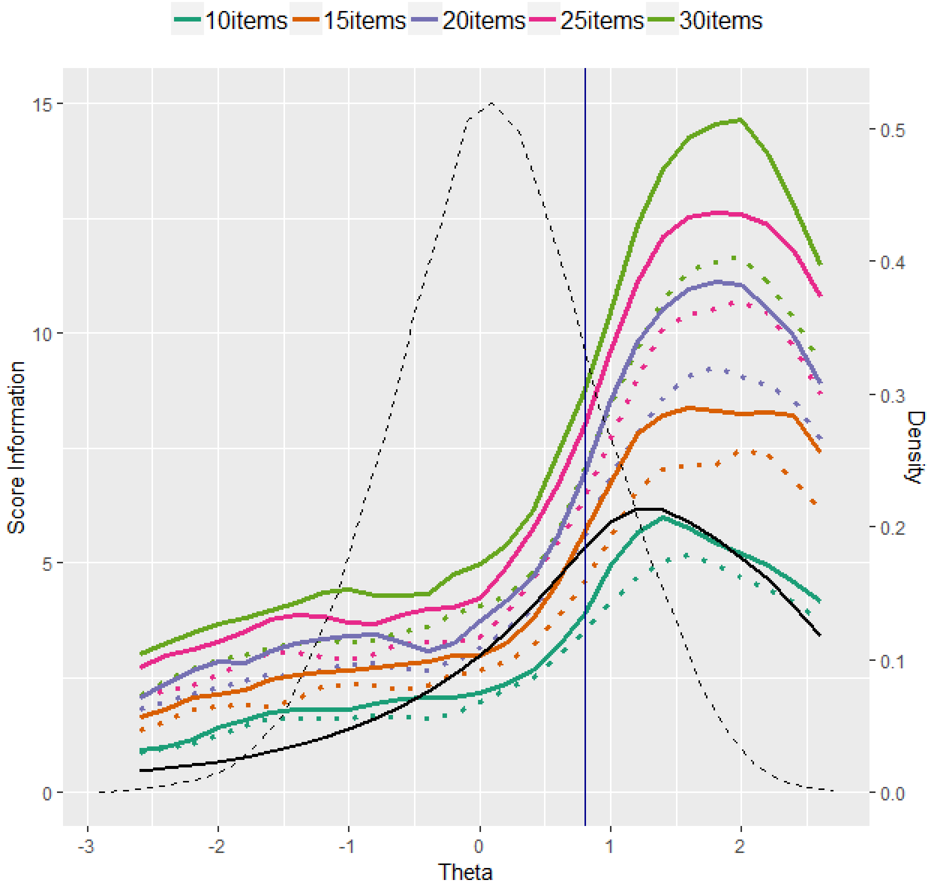


Three-Form pools



OPA
OFFICE OF PEOPLE ANALYTICS

AVERAGE SIF COMPARISON



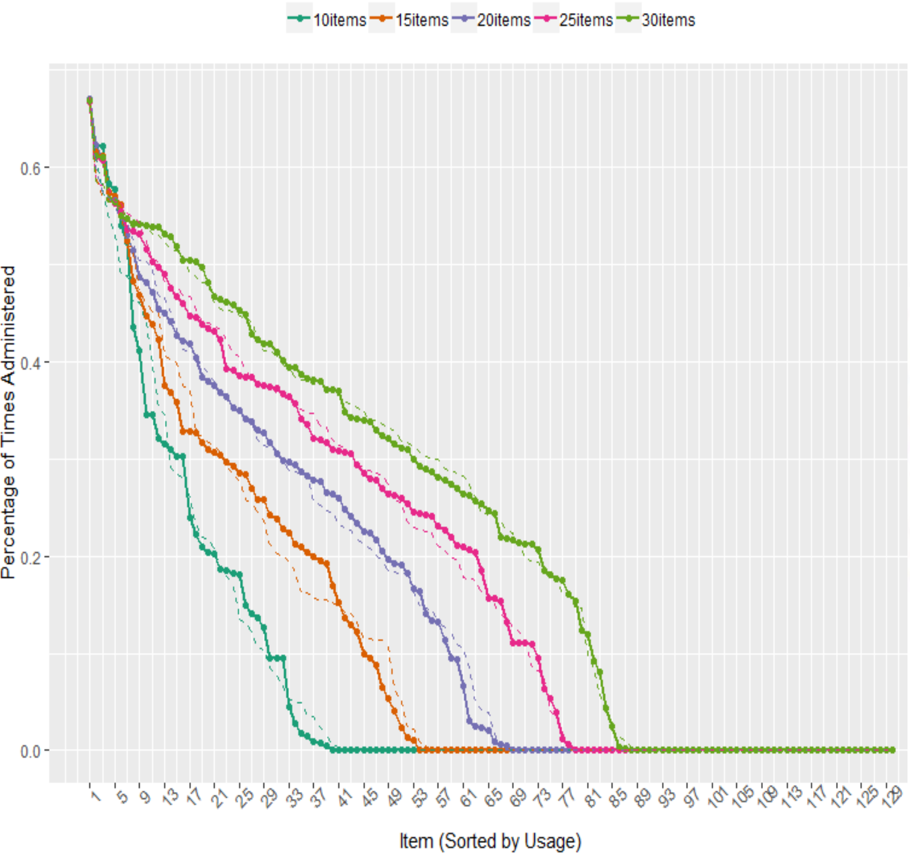
- **Averaged across the simulated tests within the two-form or three-form pool solution**
- **Dotted colored curves are the average SIFs from the three-form pool solution**
  - Test scores from the three-form pools would have lower precision
- **Black solid curve is the average test information function of the two operational forms (02A and 03A)**
  - Dashed black curve is the empirical theta score distribution of 85,294 examinees from 08/2016 to 04/2018

mean	sd	median
0.132	0.791	0.128

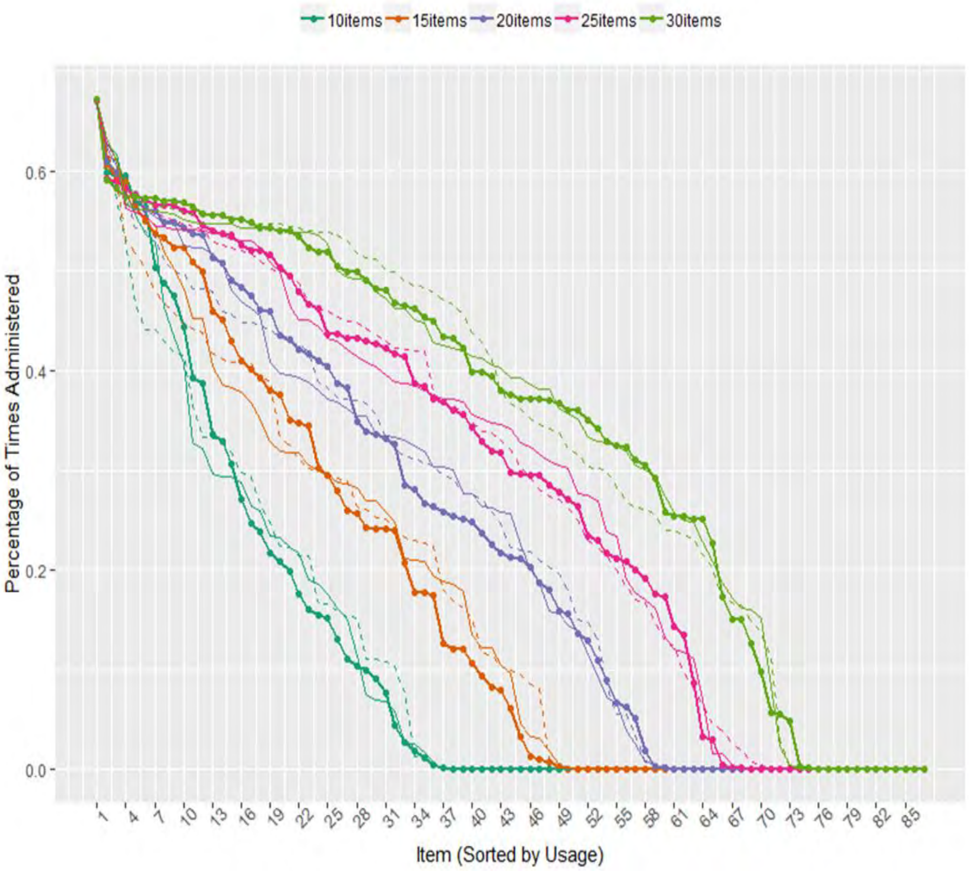
ITEM USAGE

- **Item usage is better across the three-form solution vs. the two-form solution**

Two-Form pools



Three-Form pools



ITEM USAGE

- **Two-form solution shows more not-used items than the three-form pools**
  - Even with the 30-item test, about 30% of the items in the pool were not used

Number of Items Used

CAT Form	Pool Size	10-item	15-item	20-item	25-item	30-item
01Z	130	44	55	71	82	90
02Z	130	45	58	70	82	92
01Y	87	38	51	62	71	75
02Y	87	37	48	62	71	77
03Y	87	37	48	62	71	77

SIMULATED CAT TEST-RETEST RELIABILITY

	Two-Form Pools				Three-Form Pools				
Test Length	01Z	02Z	Ave.	2016Eval. Avg	01Y	02Y	03Y	Ave.	2016Eval. Avg
10	0.72	0.73	0.73	0.73	0.68	0.70	0.72	0.72	0.70
15	0.77	0.79	0.78	0.78	0.75	0.76	0.77	0.76	0.75
20	0.81	0.82	0.82	0.81	0.78	0.79	0.80	0.80	0.78
25	0.84	0.84	0.84	0.83	0.80	0.81	0.82	0.81	0.79
30	0.85	0.86	0.86	0.84	0.82	0.83	0.83	0.83	0.78

- **Slightly higher than what was reported in the 2016 evaluation**
- **The average reliability from the two 29-item operational forms is 0.78**
- **The average simulated test-retest reliabilities of the CAT-ASVAB tests across forms 5–9 are generally higher and in the high .80s**

ASVAB Test	# Items	Avg Test-Retest Reliability
GS	15	0.87
AR	15	0.91
PC	10	0.86
SI	10	0.85

OPA
OFFICE OF PEOPLE ANALYTICS

DISCUSSION AND RECOMMENDATIONS

- **With the additional 242 items, the constructed CAT pools showed higher test score precision and test-retest reliability than previously evaluated pools**
- **Many low-discriminating items in the pool were not used in the simulated tests; these items will likely be dropped from the pools**
- **Recommendations**
  - Use the 15-item test length for CAT administration
    - Score precision higher than the two operational forms mostly in the ability range
  - Use the three-form solution
    - Higher score precision than the two static operational forms in most of the ability ranges
    - Use two forms for operational CAT to replace the two static 29-item forms
    - Reserving one as reference form for future new item scaling/equating
  - Start to develop two additional CAT forms/pools
    - Targeting more discriminating/informative items in the low-to-moderate difficulty range
- **A note: at the Feb. 20, MAPWG meeting, the Services' representatives voted unanimously in favor of a CAT test transition (at a future date) with the recommended 15-item test length and the three-form solution.**

OPA

OFFICE OF PEOPLE ANALYTICS

Backup Slides

THE MODEL

- The sampling distribution of item responses U on a *d*-dimensional test, given latent factor vector Θ

$$P(U|\Theta) = \prod_{a=1}^{N} \prod_{i=1}^{n} P_i(\theta_a)^{u_{ia}}[1 - P_i(\theta_a)]^{1-u_{ia}}$$

$$P_i(\theta_a) = c_i + (1 - c_i)\,\Psi_i(\tau_i + \lambda'_i\theta_a)$$

Where:

$\Psi(\cdot)$ is the distribution function of N(0, 1)

c_i, τ_i are the guessing, intercept parameter for the *i-th* item

λ_i is the slope parameter vector for the item

θ_a is the *d*-dimensional latent vector of examinee *a*

- Item factor analysis
  - θ latent factors
  - λ factor loadings

OFFICE OF PEOPLE ANALYTICS

FORM1 + FORM2 + NEW SEEDED: 300 ITEMS

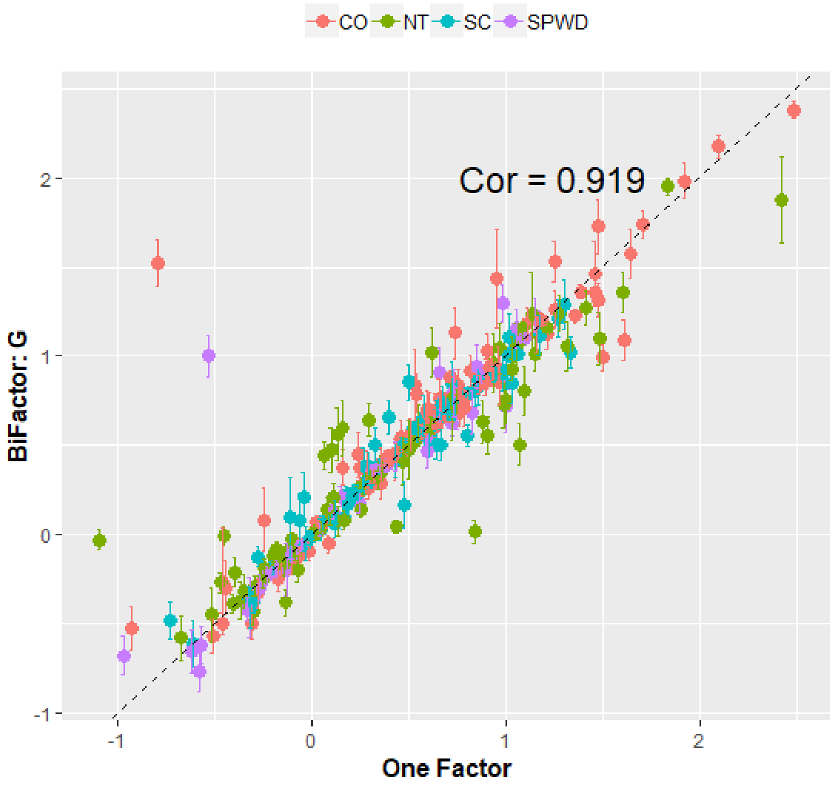
- **Total 300 items**
  - 29 Form1 items
  - 29 Form2 items
  - 242 seeded items
  - Number of items from each content area:
    - CO 95
    - NT 89
    - SC 76
    - SPWD 40

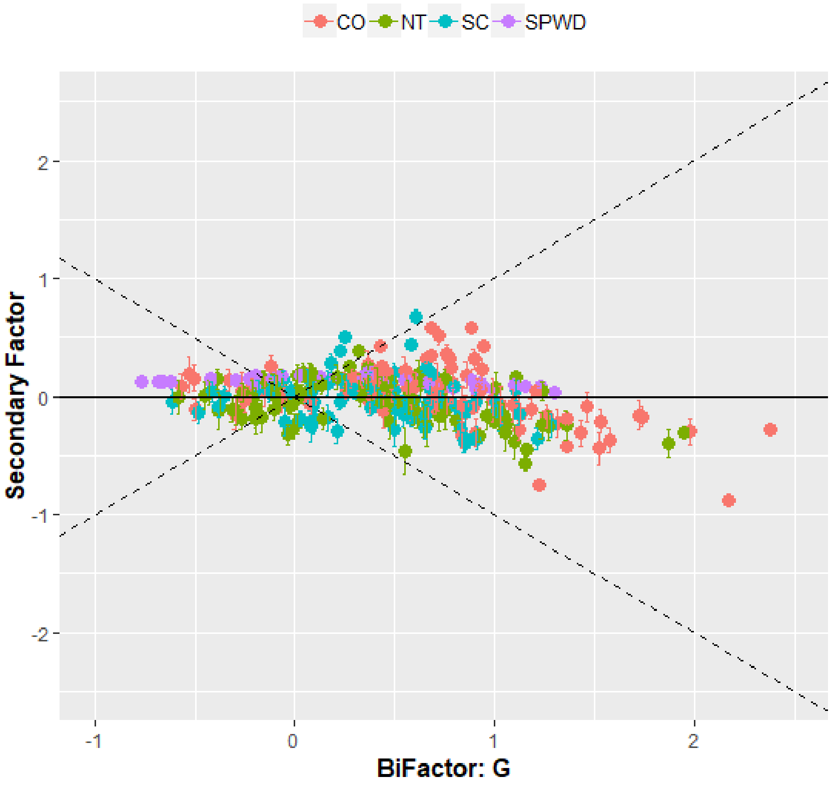
- **42,009 test-takers on Form1; 43,295 on Form2**

- **Case counts on new seeded items range from 3,174 to 3,895**

IFACT RESULTS (NEW 242 ITEMS RESULTS)

- Form1 + Form2 + 242 seeded items: 300 items
- ECV = 0.927; ECV.adj = 0.938

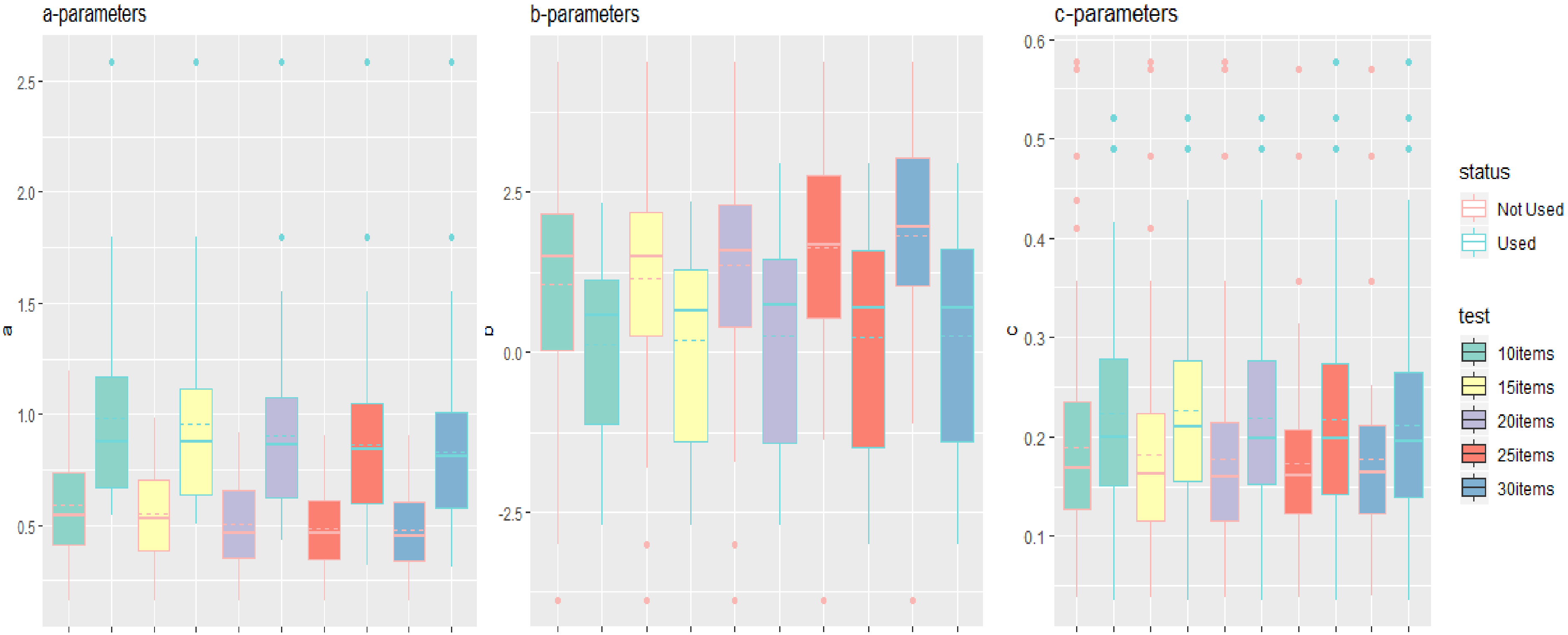




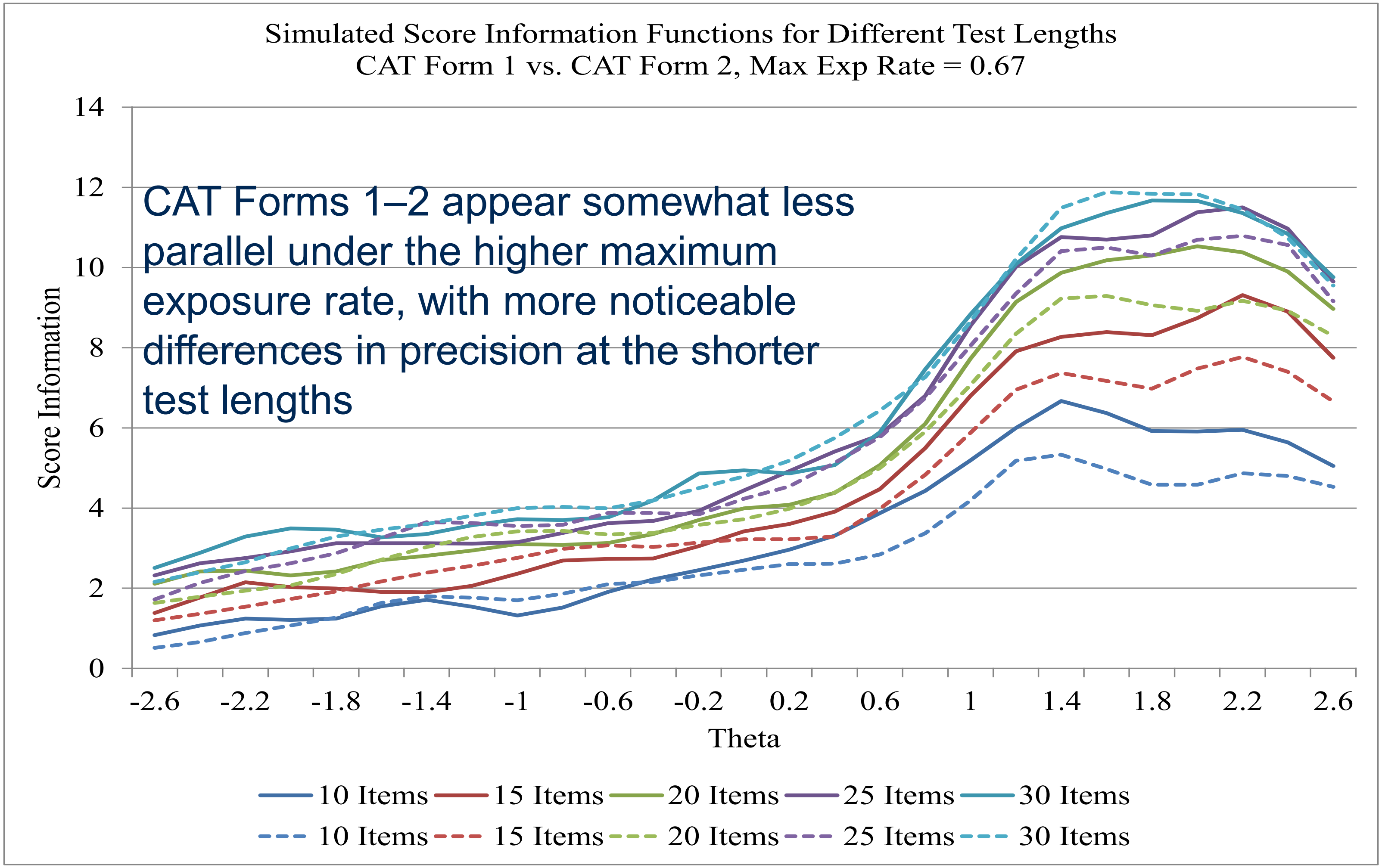
USED AND NOT-USED ITEM COMPARISON

IN 01Z (ONE OF THE TWO-FORM POOLS)

- **Not-used items: low-discriminating, difficult items**



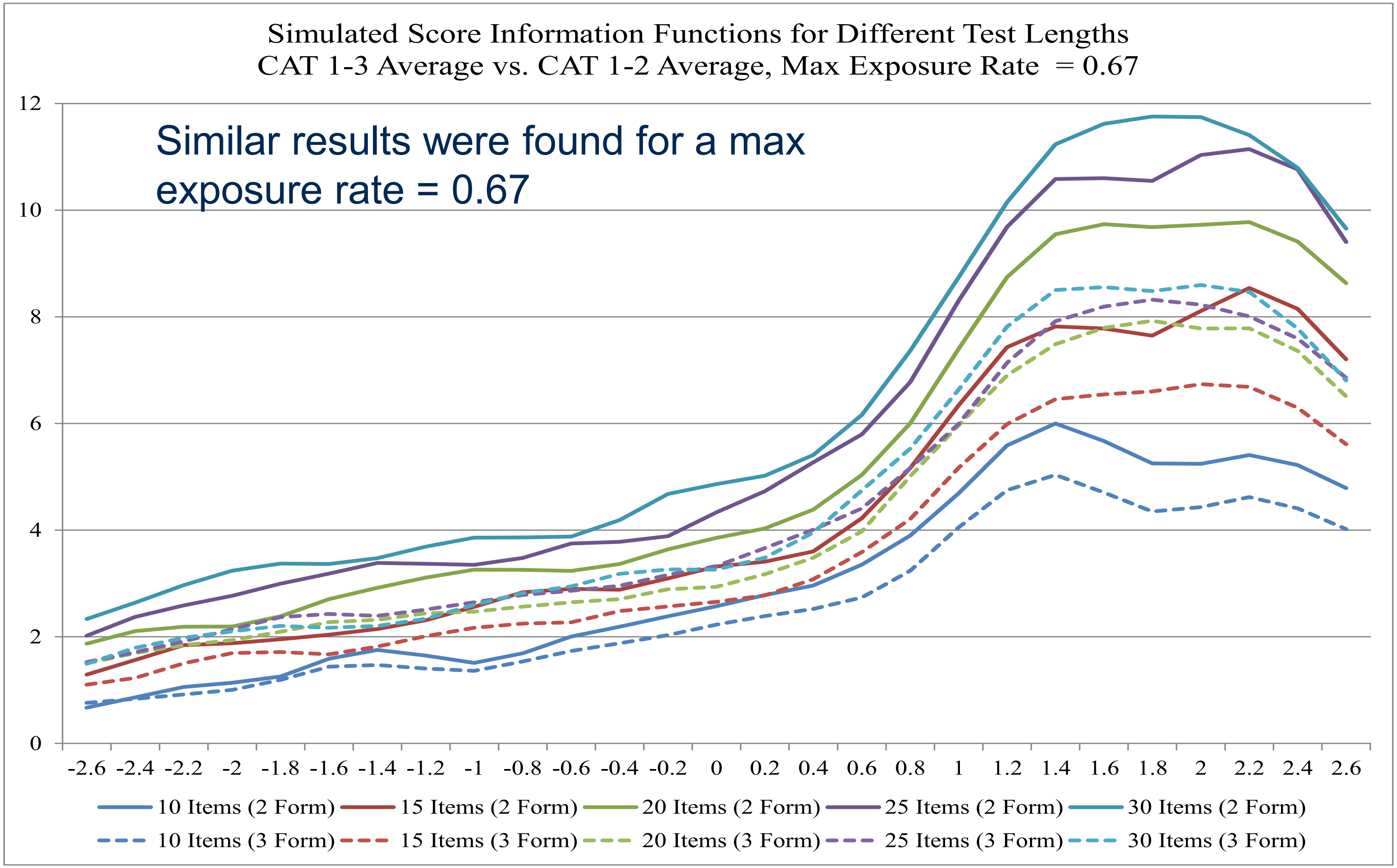
2016: CAT Results—2-Form Solution



CAT Forms 1–2 appear somewhat less parallel under the higher maximum exposure rate, with more noticeable differences in precision at the shorter test lengths

OFFICE OF PEOPLE ANALYTICS

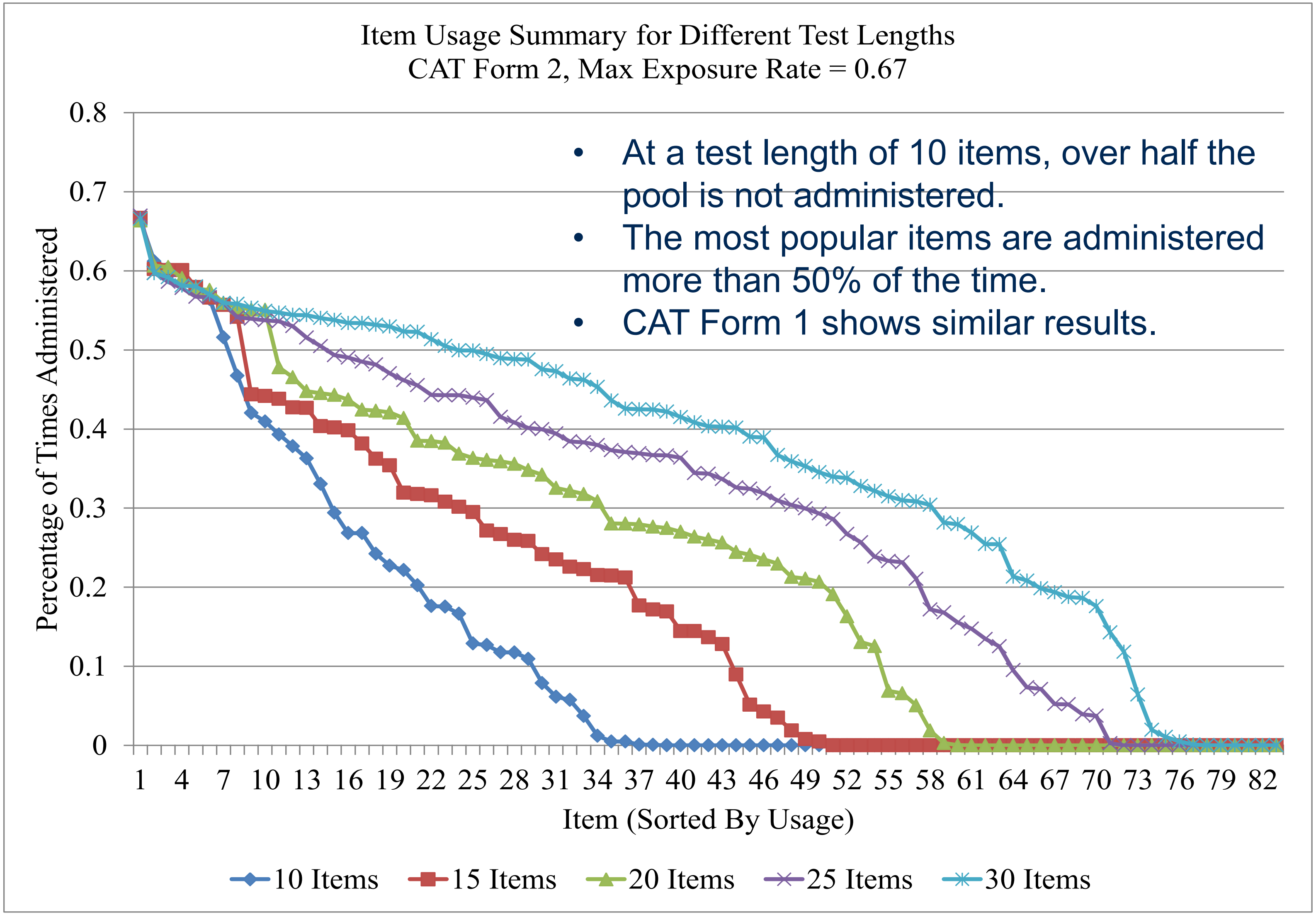
2016: CAT Results—3-Form Solution



Similar results were found for a max exposure rate = 0.67

OFFICE OF PEOPLE ANALYTICS

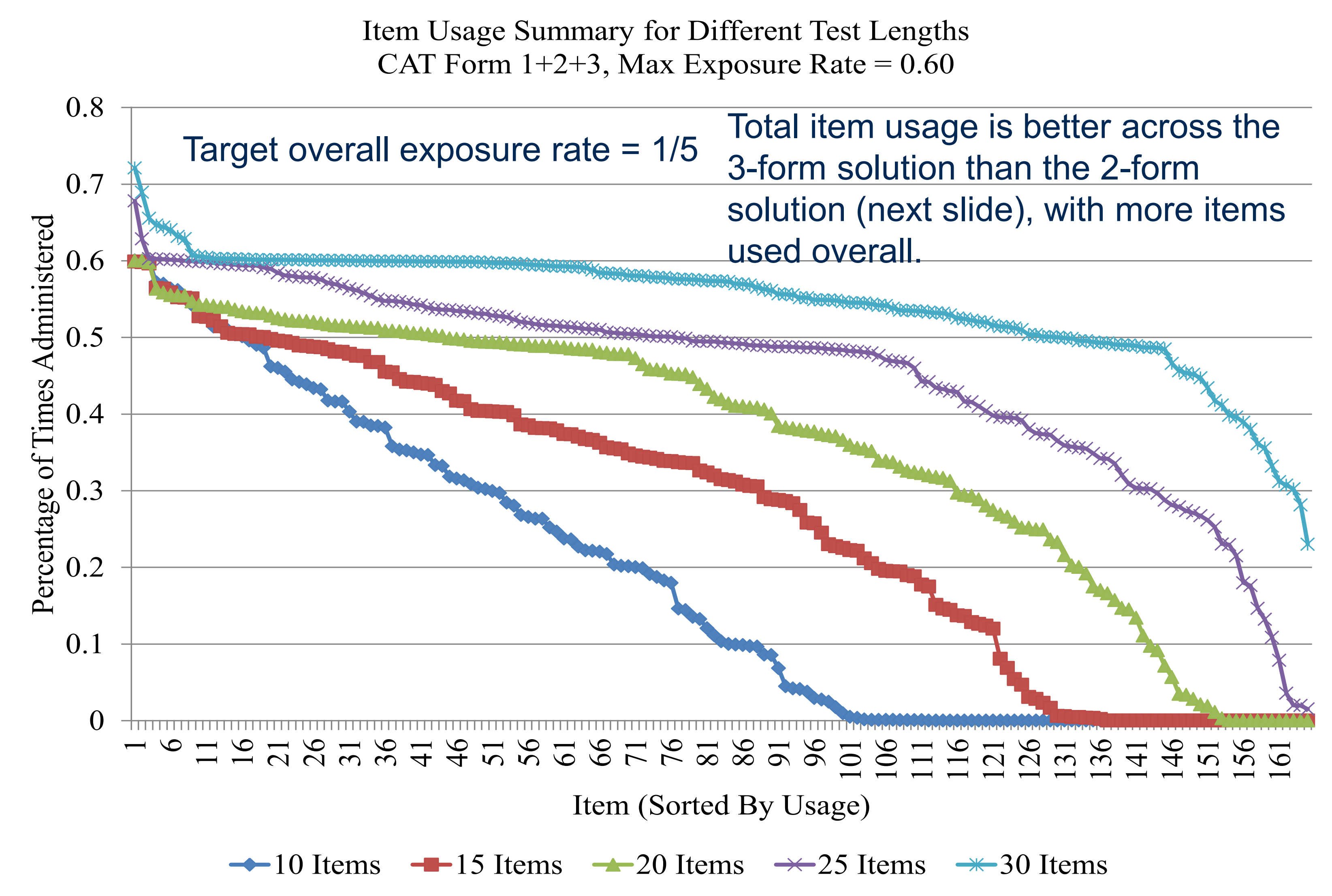
CAT Results



- At a test length of 10 items, over half the pool is not administered.
- The most popular items are administered more than 50% of the time.
- CAT Form 1 shows similar results.

OFFICE OF PEOPLE ANALYTICS

CAT Results—3-Form Solution



Target overall exposure rate = 1/5

Total item usage is better across the 3-form solution than the 2-form solution (next slide), with more items used overall.

OFFICE OF PEOPLE ANALYTICS

DESCRIPTIVE STATISTICS

Form	Type	Parameter	N	Mean	SD	Min	Max
01Z	CAT	a	130	0.722	0.355	0.163	2.584
		b		0.745	1.725	-3.869	4.515
		c		0.201	0.108	0.036	0.577
02Z	CAT	a	130	0.686	0.360	0.166	2.064
		b		0.634	1.821	-4.498	3.450
		c		0.186	0.108	0.016	0.519
01Y	CAT	a	87	0.700	0.279	0.200	1.347
		b		0.611	1.900	-4.498	4.515
		c		0.187	0.105	0.016	0.569
02Y	CAT	a	87	0.700	0.402	0.166	2.584
		b		0.856	1.726	-2.999	5.906
		c		0.202	0.109	0.045	0.577
03Y	CAT	a	87	0.705	0.383	0.220	2.064
		b		0.769	1.711	-3.869	4.071
		c		0.194	0.112	0.019	0.496
02A	Static	a	29	0.900	0.449	0.343	1.753
		b		0.152	1.468	-3.500	2.540
		c		0.196	0.138	0.041	0.501
03A	Static	a	29	0.832	0.403	0.256	2.047
		b		0.052	1.655	-4.260	2.033
		c		0.180	0.132	0.036	0.513

Tab L

OFFICE OF PEOPLE ANALYTICS

Potential for Adverse Impact of the ASVAB Platform Special Tests Findings for Fiscal Year 2017 Applicants

Gregory Manley
Richard Riemer
Ping Yin
Mary Pommerich
DPAC

DAC-MPT
3.29.2019 | Carmel-By-The-Sea, CA

POTENTIAL FOR ADVERSE IMPACT

- Adverse impact (AI) is the unintended discrimination of a protected class that is the result of a selection procedure (Uniform Guidelines, 1978).
- AI is not a property of a test per se. However, AI may occur when a test's scores are used as the bases for selection.
- A test may contribute to the occurrence of AI when it shows sizable mean test score differences between a majority group and a protected class (minority).
- Effect sizes of the standardized mean difference gives us an index to examine a test's potential for AI.

POTENTIAL FOR ADVERSE IMPACT

- Effect sizes (i.e., standardized mean differences) provide a method of evaluating potential for adverse impact across individual ASVAB and Special Tests, where no direct selection occurs.
- Effect sizes are computed for all group comparisons as:

$$ES = \frac{\mu_R - \mu_F}{\sigma_p}$$

where:

μ_R is the mean score in the Reference (Majority) group.

μ_F is the mean score in the Focal (Minority) group.

σ_p is the pooled standard deviation across the two groups.

Note. Positive values are the direction of minority impact

OFFICE OF PEOPLE ANALYTICS

CONFIDENCE INTERVALS ABOUT EFFECT SIZES

- A 95% confidence interval (δ_L, δ_U) for the effect size (ES) is computed as (Hedges & Olkin, 1985):

$$\delta_L = ES - 1.96\hat{\sigma}(ES) \qquad \delta_U = ES + 1.96\hat{\sigma}(ES)$$

where

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

- Effect sizes can be plotted and classified with respect to Cohen’s (1988) standards of evaluation.
  - Small effect sizes start at 0.20.
  - Moderate effect sizes start at 0.50.
  - Large effect sizes start at 0.80.

OFFICE OF PEOPLE ANALYTICS

WHO IS AFFECTED BY ADVERSE IMPACT?

- The ASVAB testing program evaluates comparisons for the following pairs of groups:

Pair	Reference Group	Focal Group
1	Males	Females
2	Non-Hispanic Whites	Hispanic Whites
3	Non-Hispanic Whites	Non-Hispanic Blacks
4	Non-Hispanic Whites	Non-Hispanic Asians

- The focal group is potentially disadvantaged relative to the reference group.
- Pairs 1–3 are the same groups that are used in evaluating DIF. Pair 4 is also included because Non-Hispanic Asians now represent >2% of the applicant population.

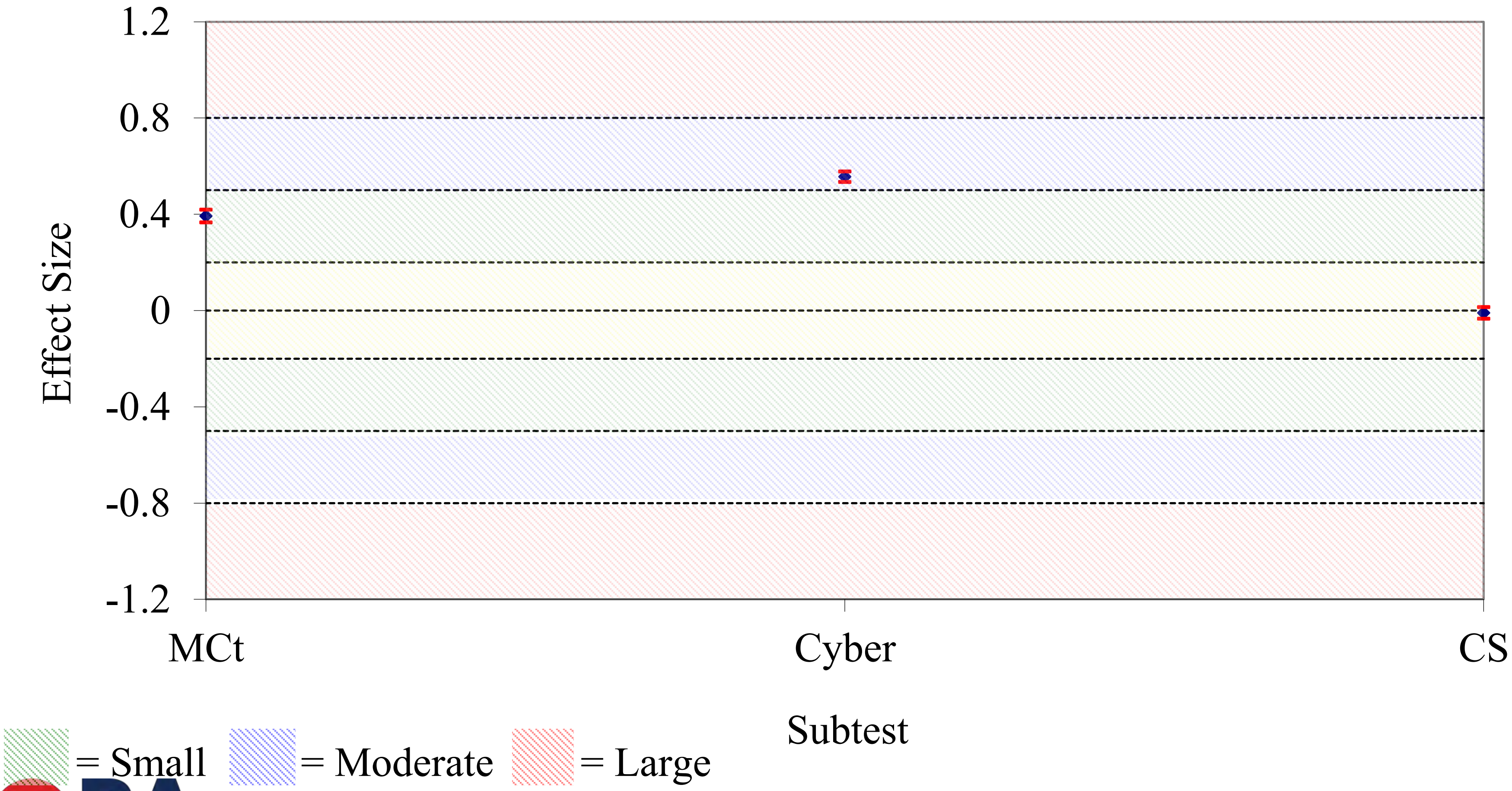
SPECIAL TESTS ON ASVAB PLATFORM

- **Mental Counters (MCt)**: A counting test of working memory (Navy only)

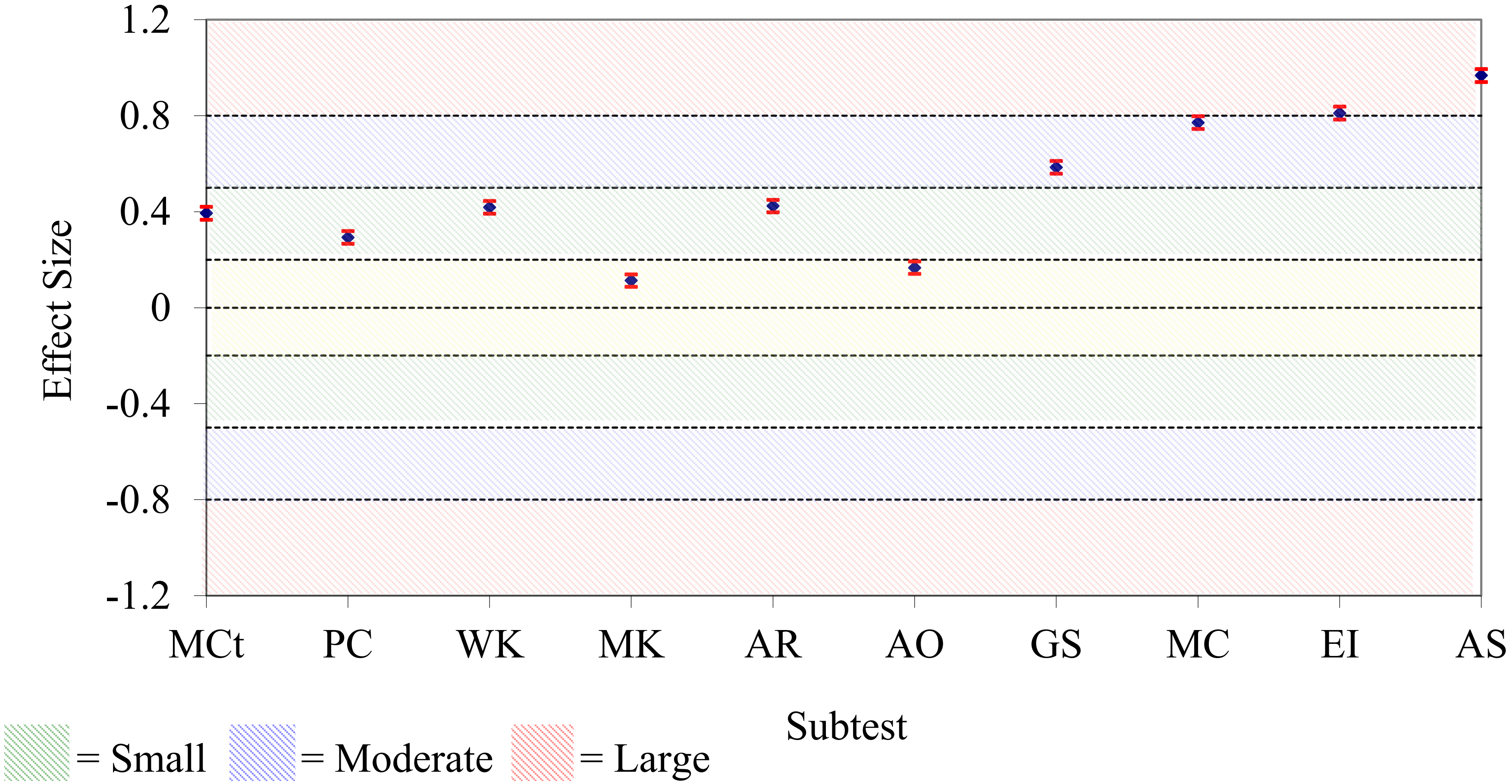
- **Cyber Test (Cyber):** Test of basic computer and information systems knowledge (All Services)

- **Coding Speed (CS):** A speeded test of assigning code numbers to words (Navy only)

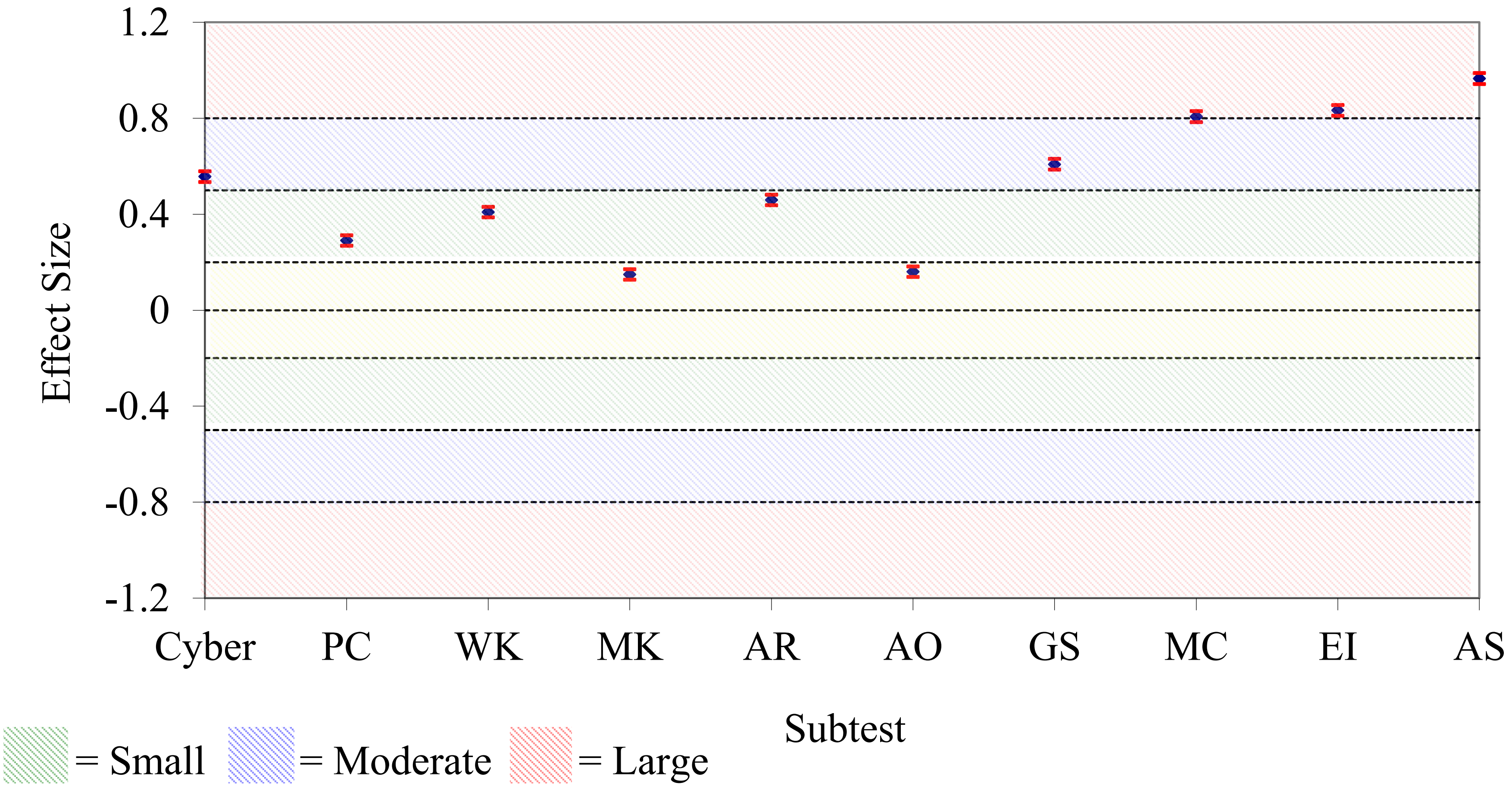
Effect Sizes (and 95% Confidence Interval) for Special Tests Scores
Males vs. Females
FY2017



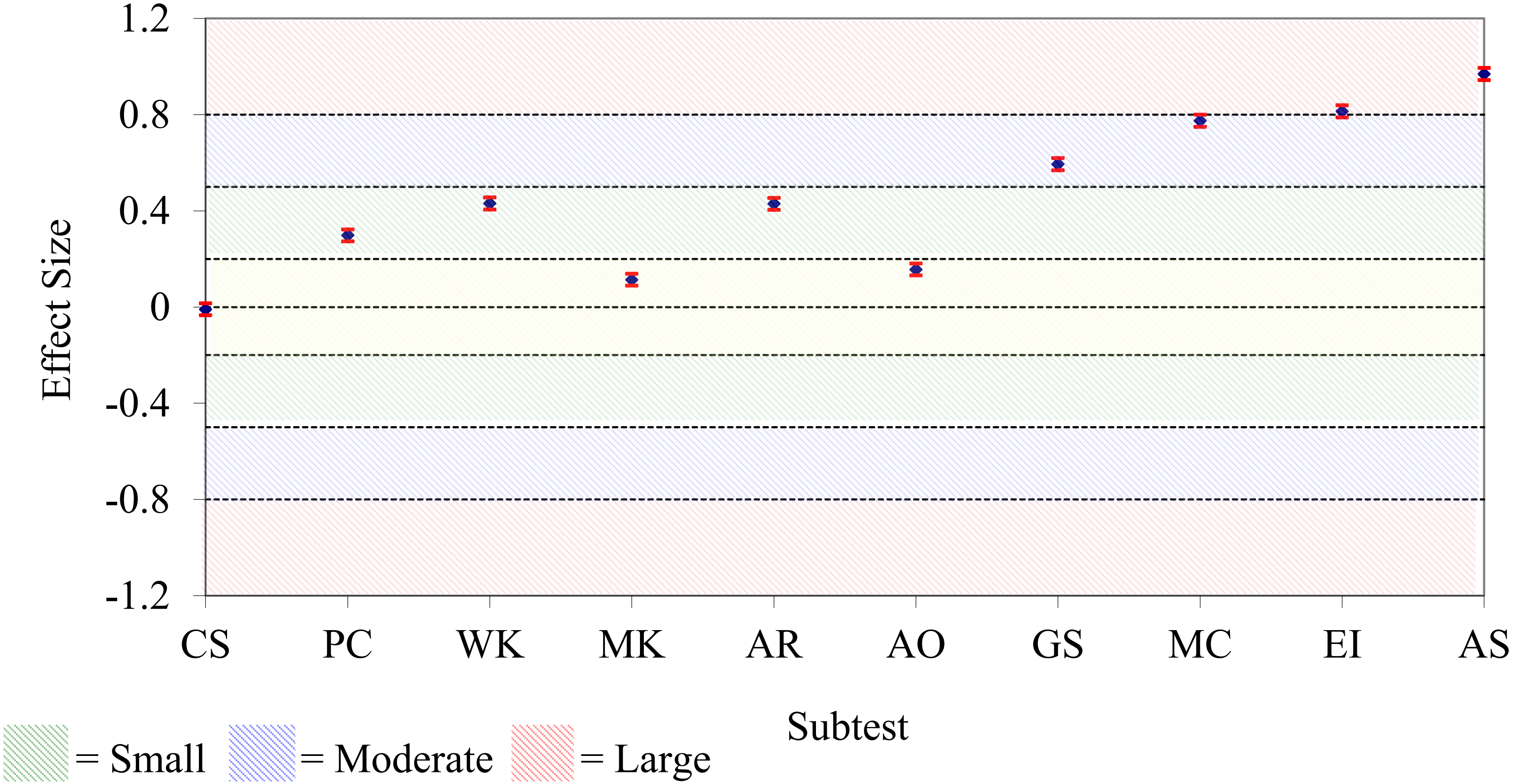
Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Males vs. Females for MCt Sample
FY2017



Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Males vs. Females for Cyber Test Sample
FY2017

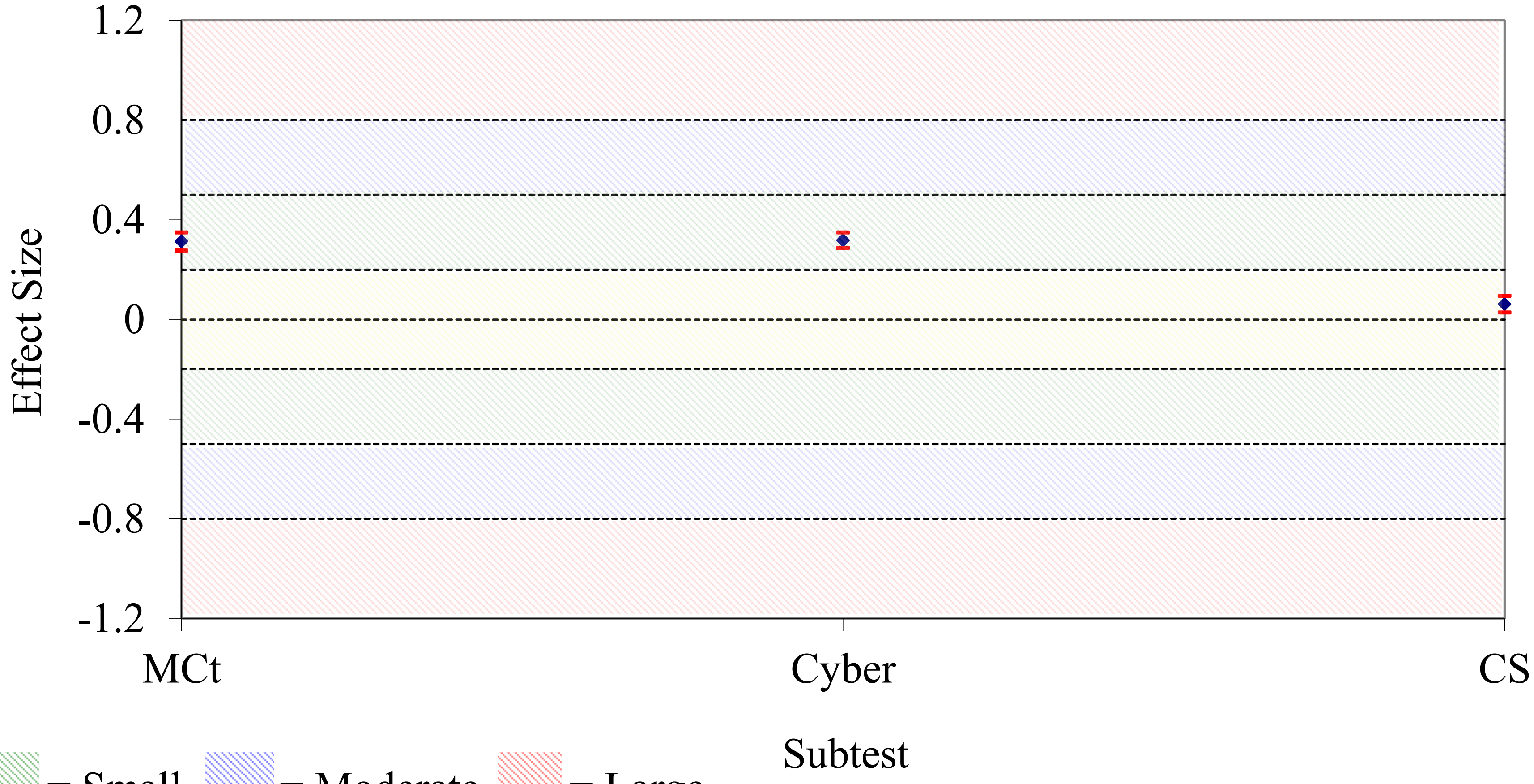


Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Males vs. Females for Coding Speed Sample
FY2017



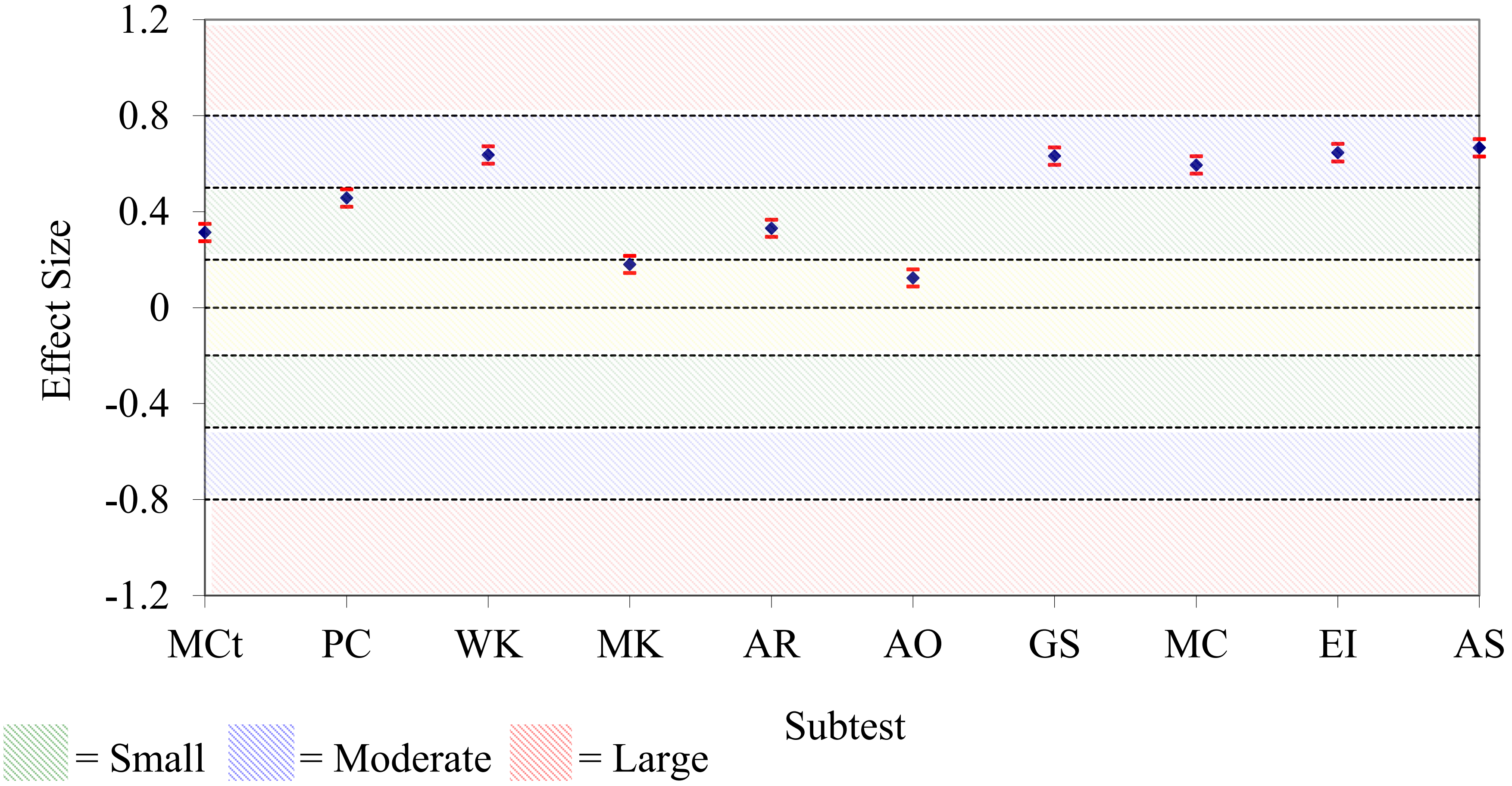
OPA OFFICE OF PEOPLE ANALYTICS

Effect Sizes (and 95% Confidence Interval) for Special Tests Scores
Non-Hispanic Whites vs. Hispanic Whites
FY2017

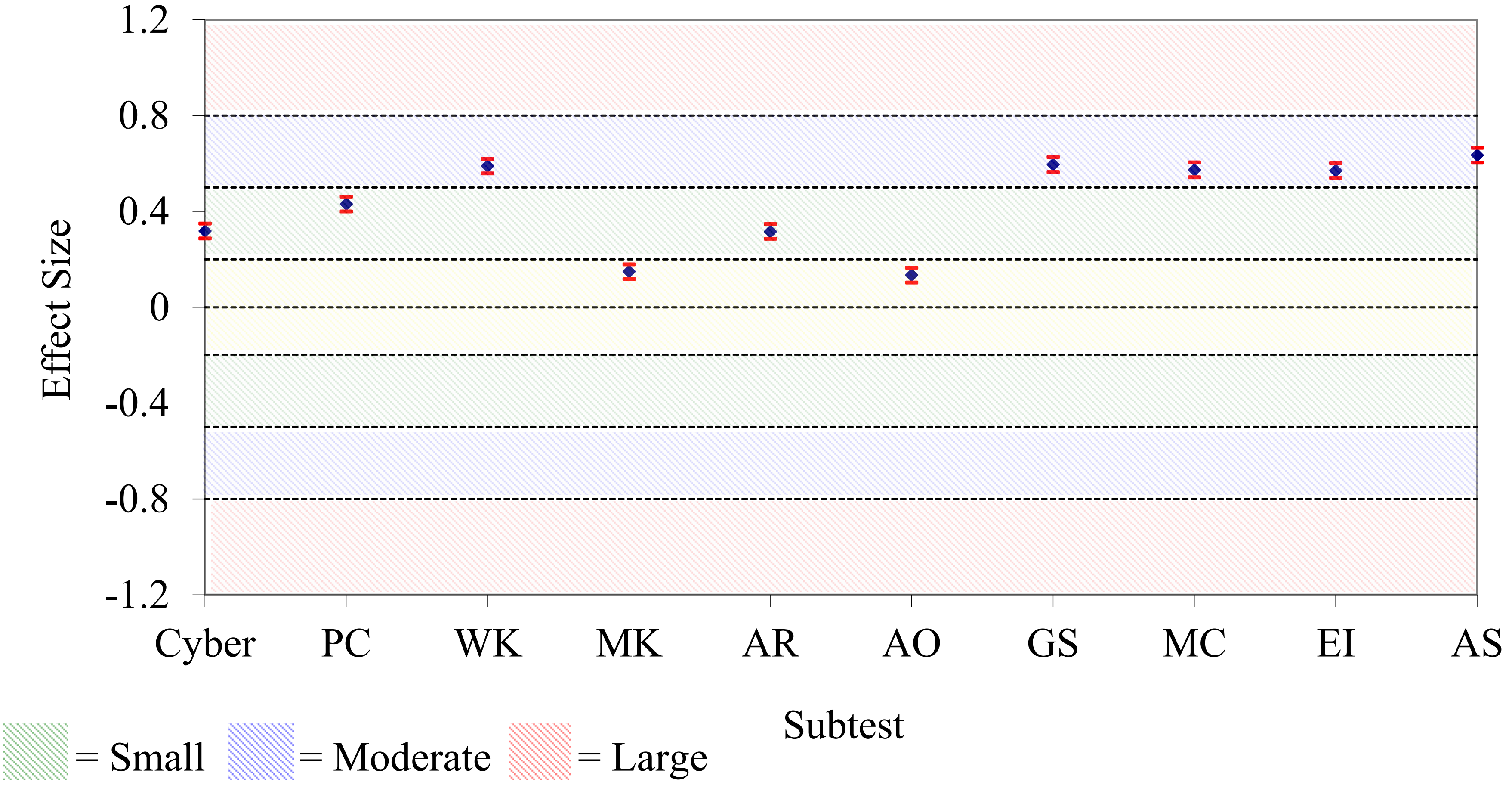


▨ = Small ▨ = Moderate ▨ = Large

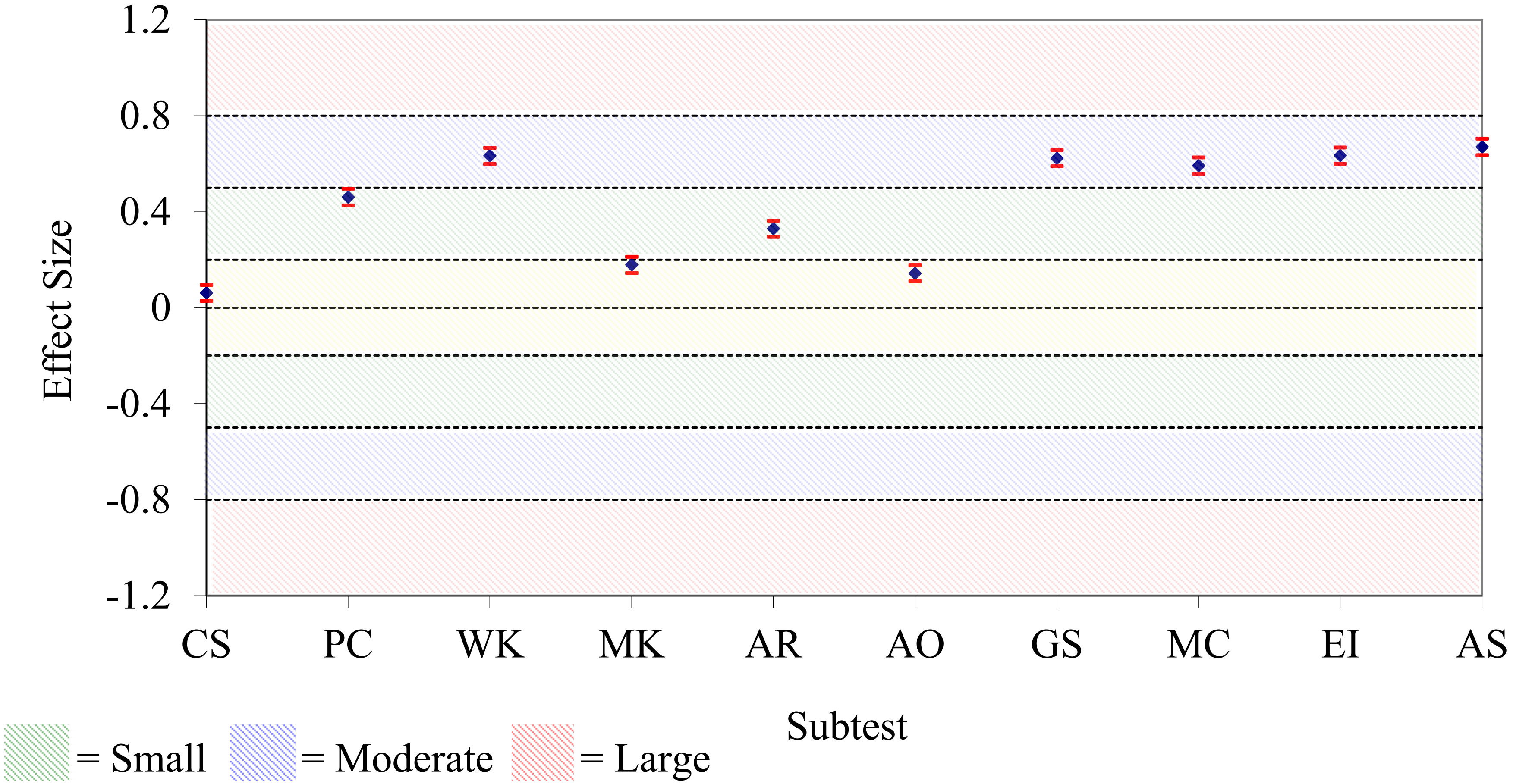
Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Non-Hispanic Whites vs. Hispanic Whites
FY2017 for Mental Counters Sample



Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Non-Hispanic Whites vs. Hispanic Whites
FY2017 for Cyber Test Sample

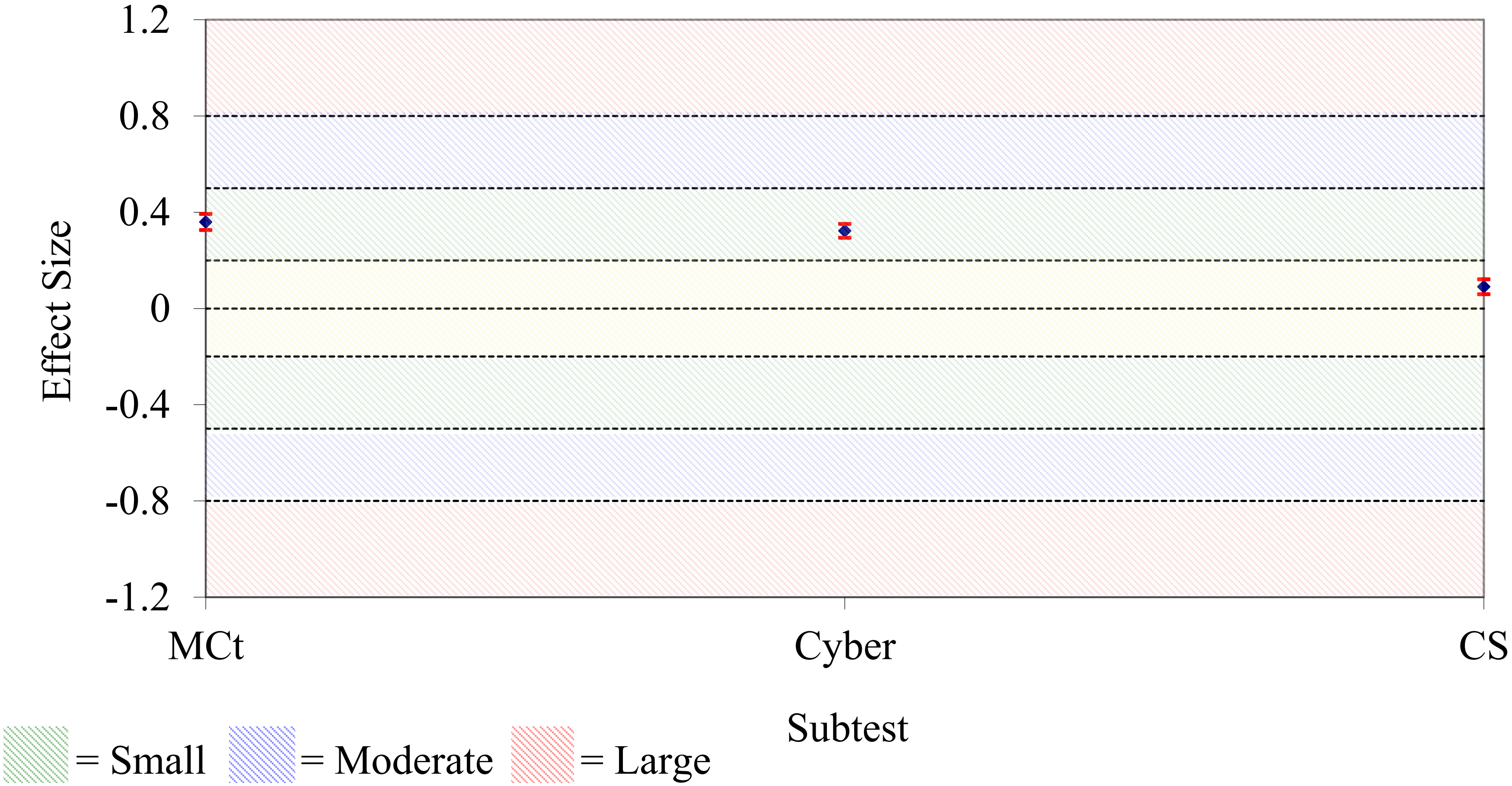


Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Non-Hispanic Whites vs. Hispanic Whites
FY2017 for Coding Speed Sample

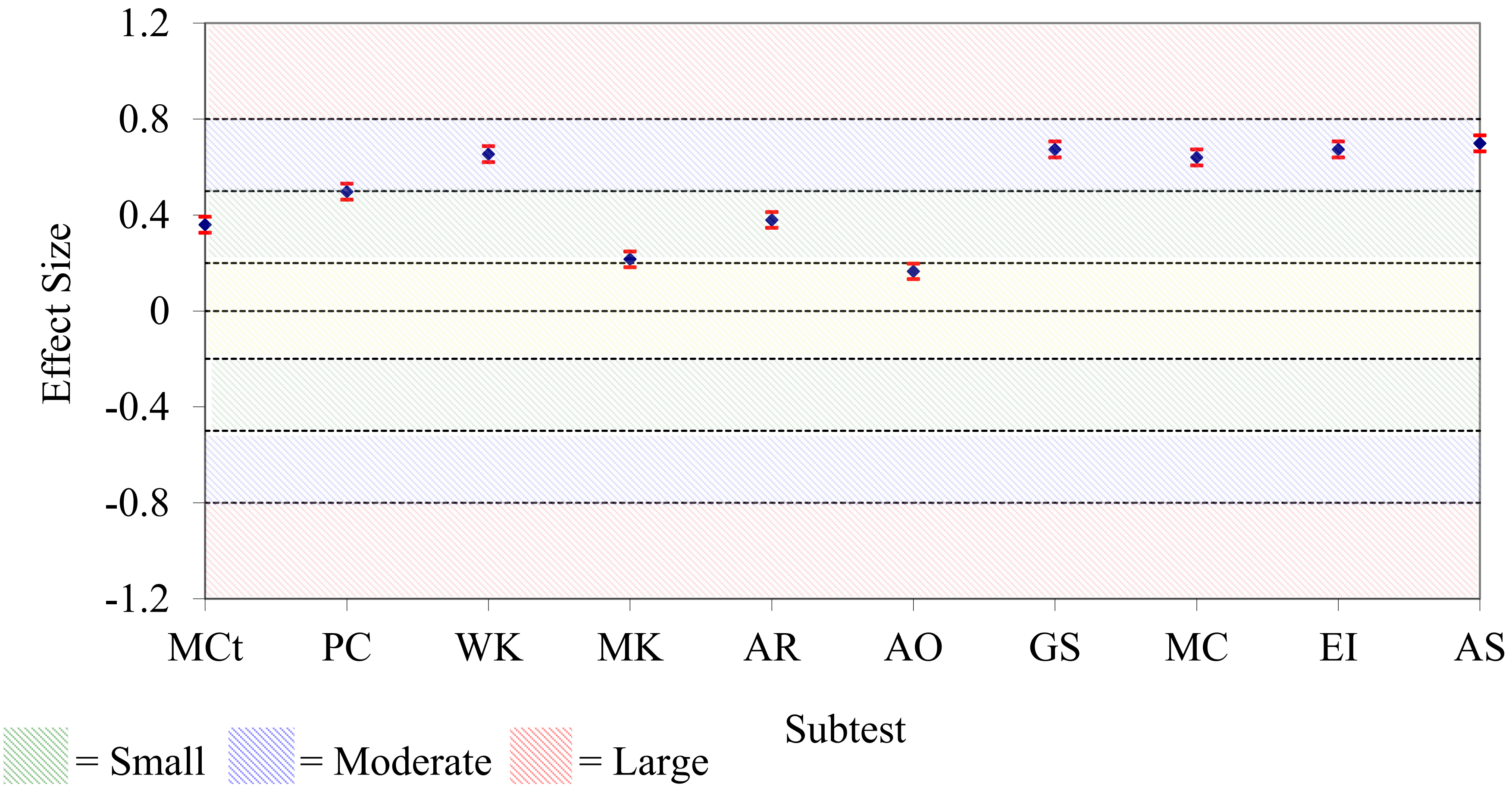


OFFICE OF PEOPLE ANALYTICS

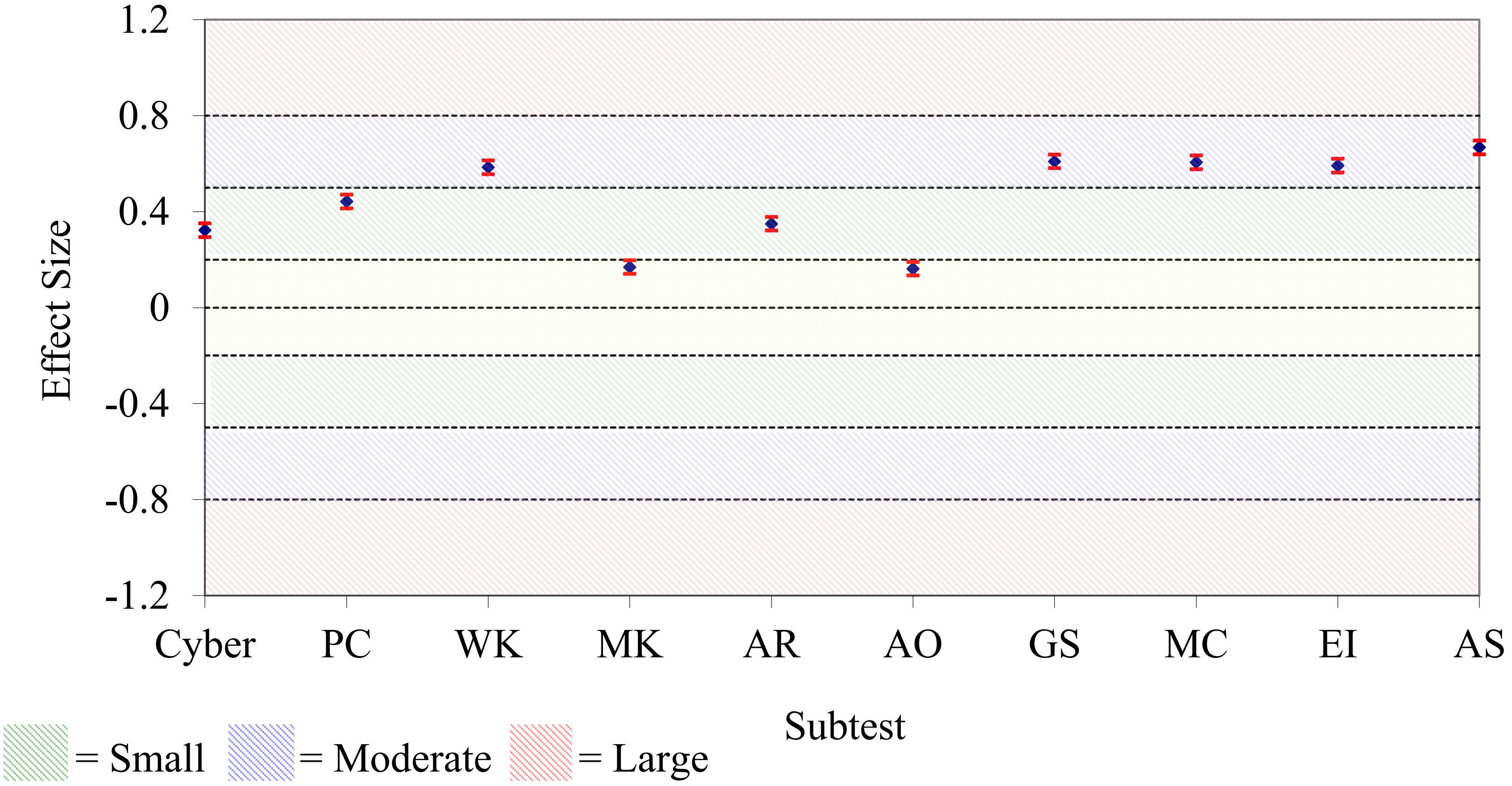
Effect Sizes (and 95% Confidence Interval) for Special Tests Scores
Non-Hispanic Whites vs. Hispanics*
FY2017



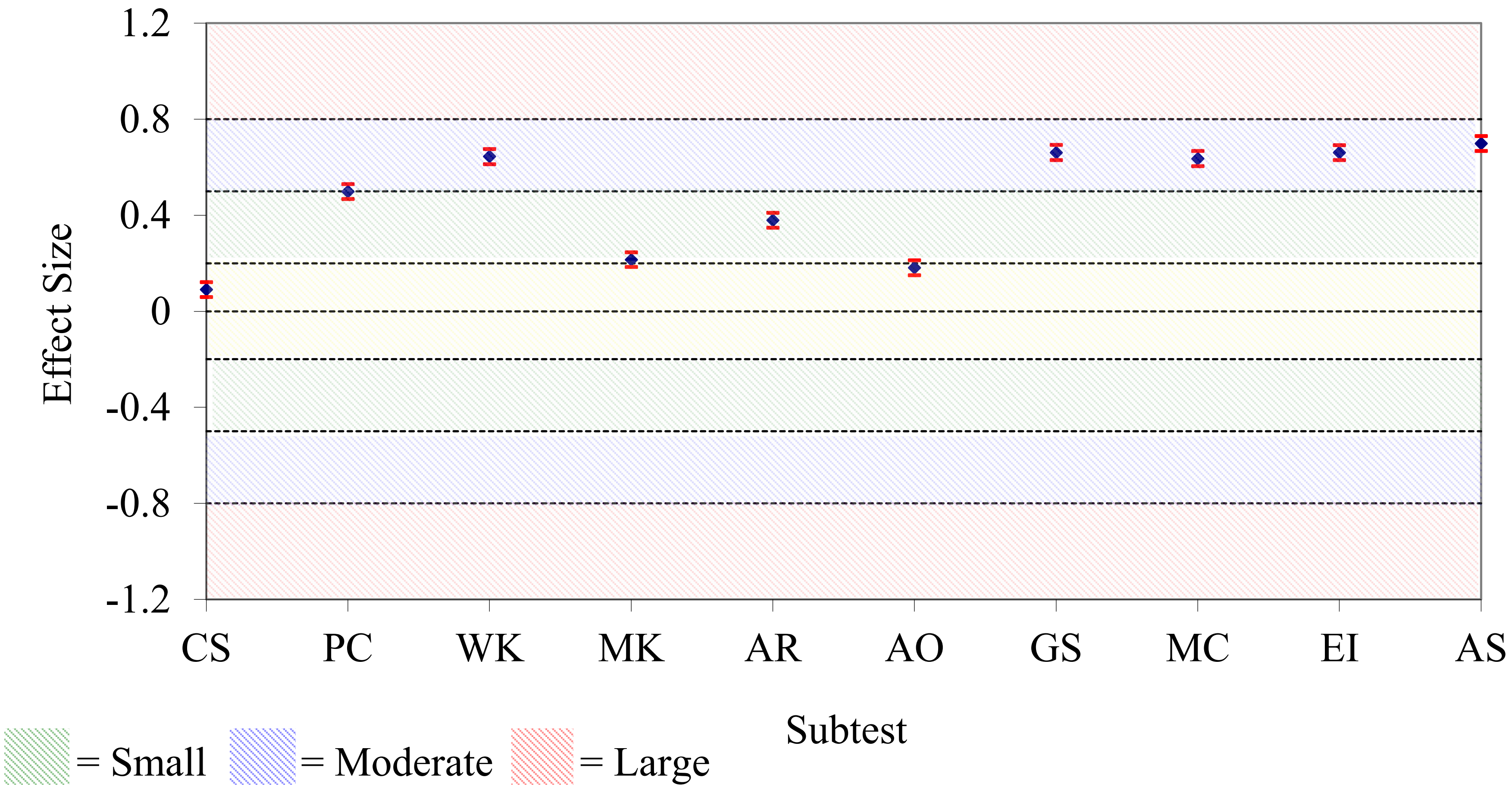
Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Non-Hispanic Whites vs. Hispanics*
FY2017 for Mental Counters Sample



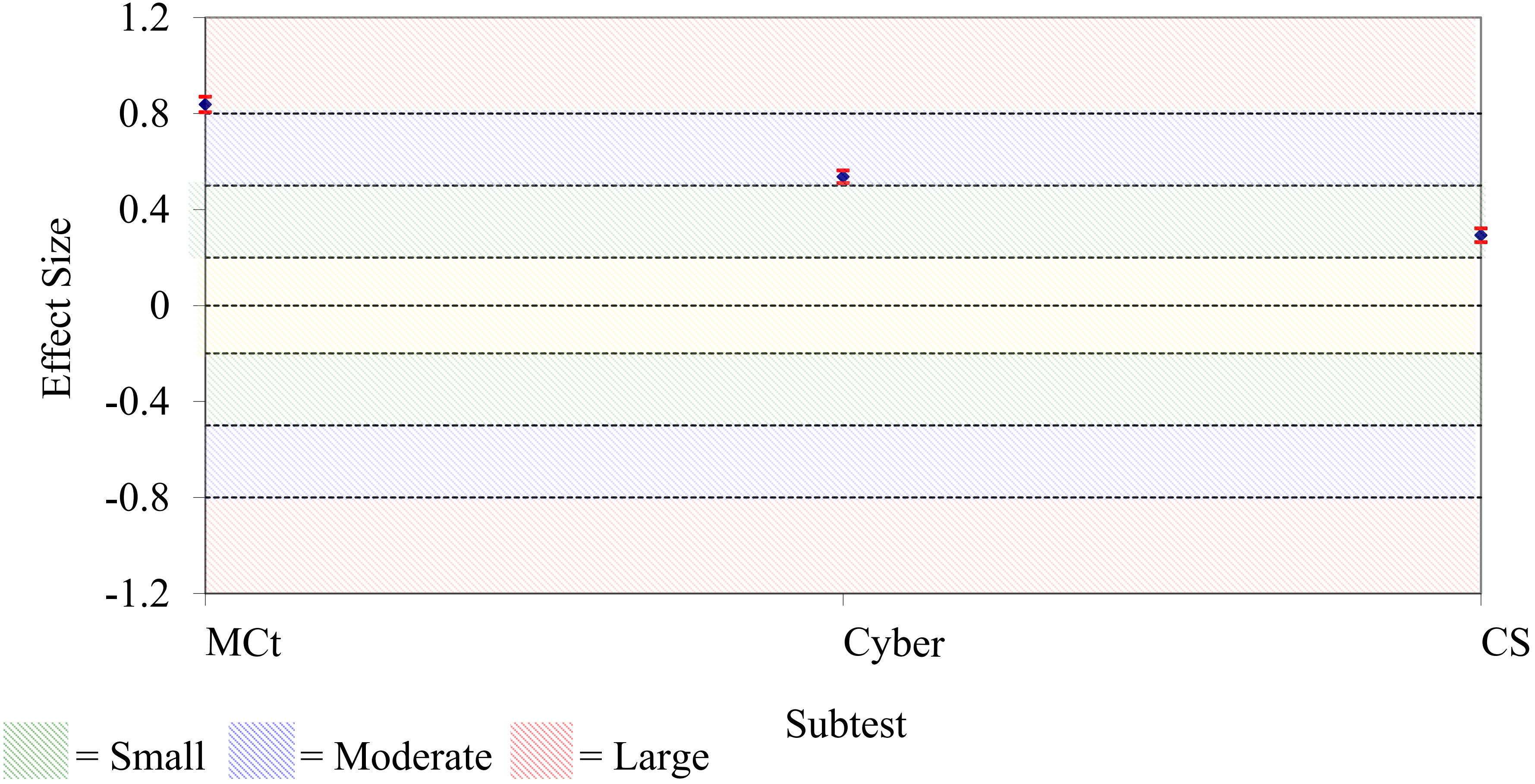
Effect Sizes (and 95% Confidence Interval) for ASVAB Scores Non-Hispanic Whites vs. Hispanics* FY2017 for Cyber Test Sample



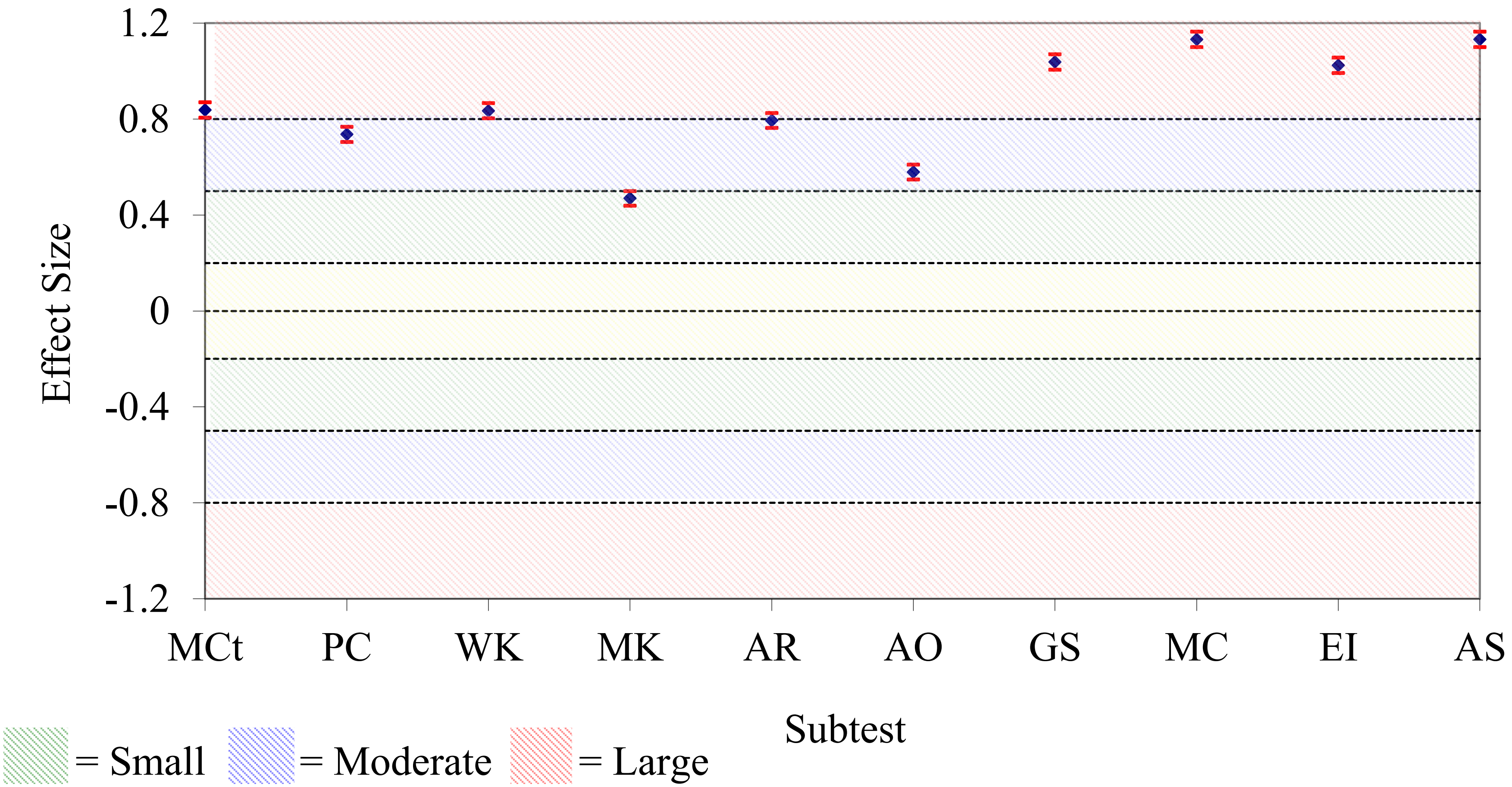
Effect Sizes (and 95% Confidence Interval) for ASVAB Scores Non-Hispanic Whites vs. Hispanics* FY2017 for Coding Speed Sample



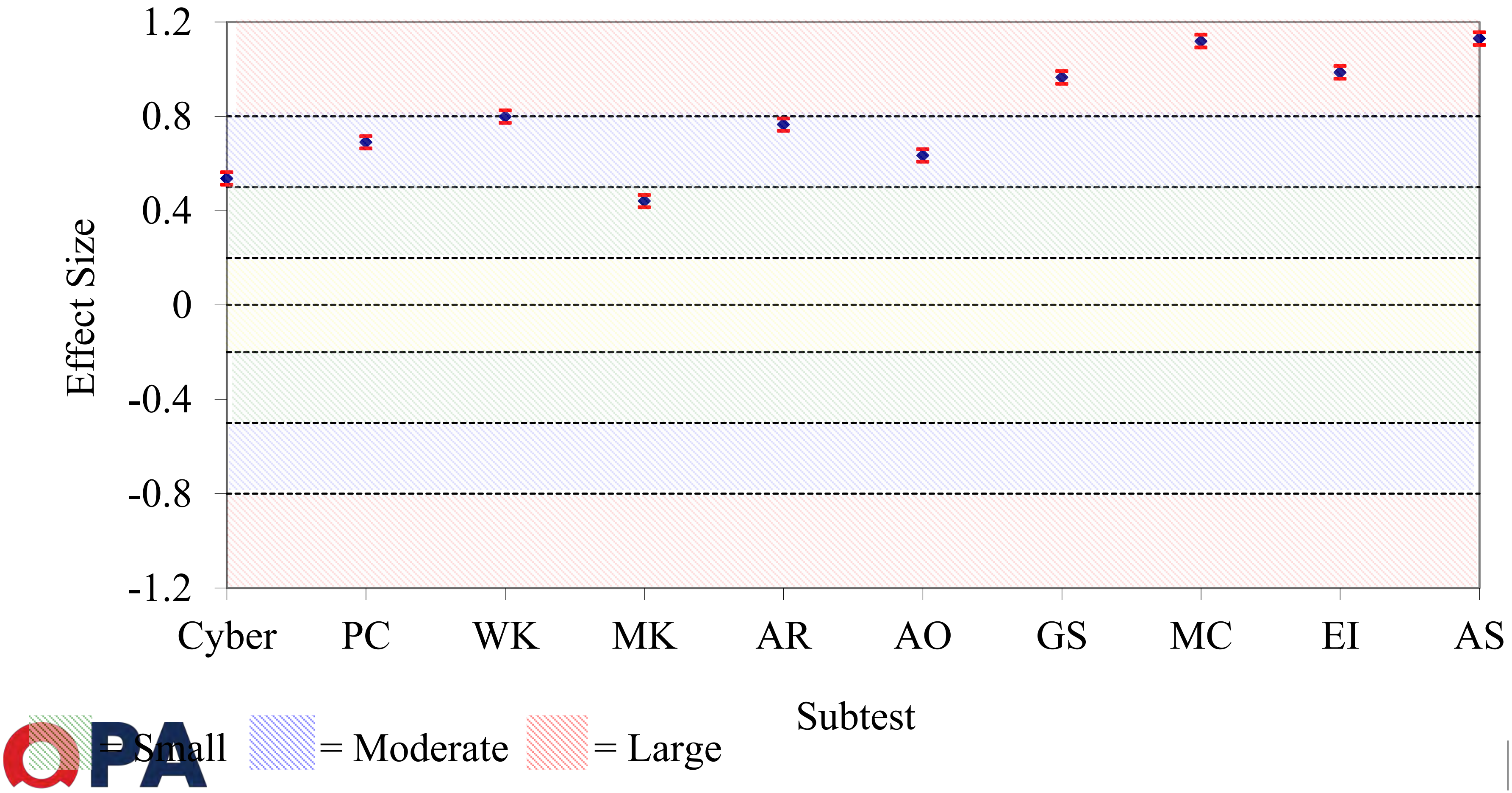
Effect Sizes (and 95% Confidence Interval) for Special Tests Scores
Non-Hispanic Whites vs. Non-Hispanic Blacks
FY2017



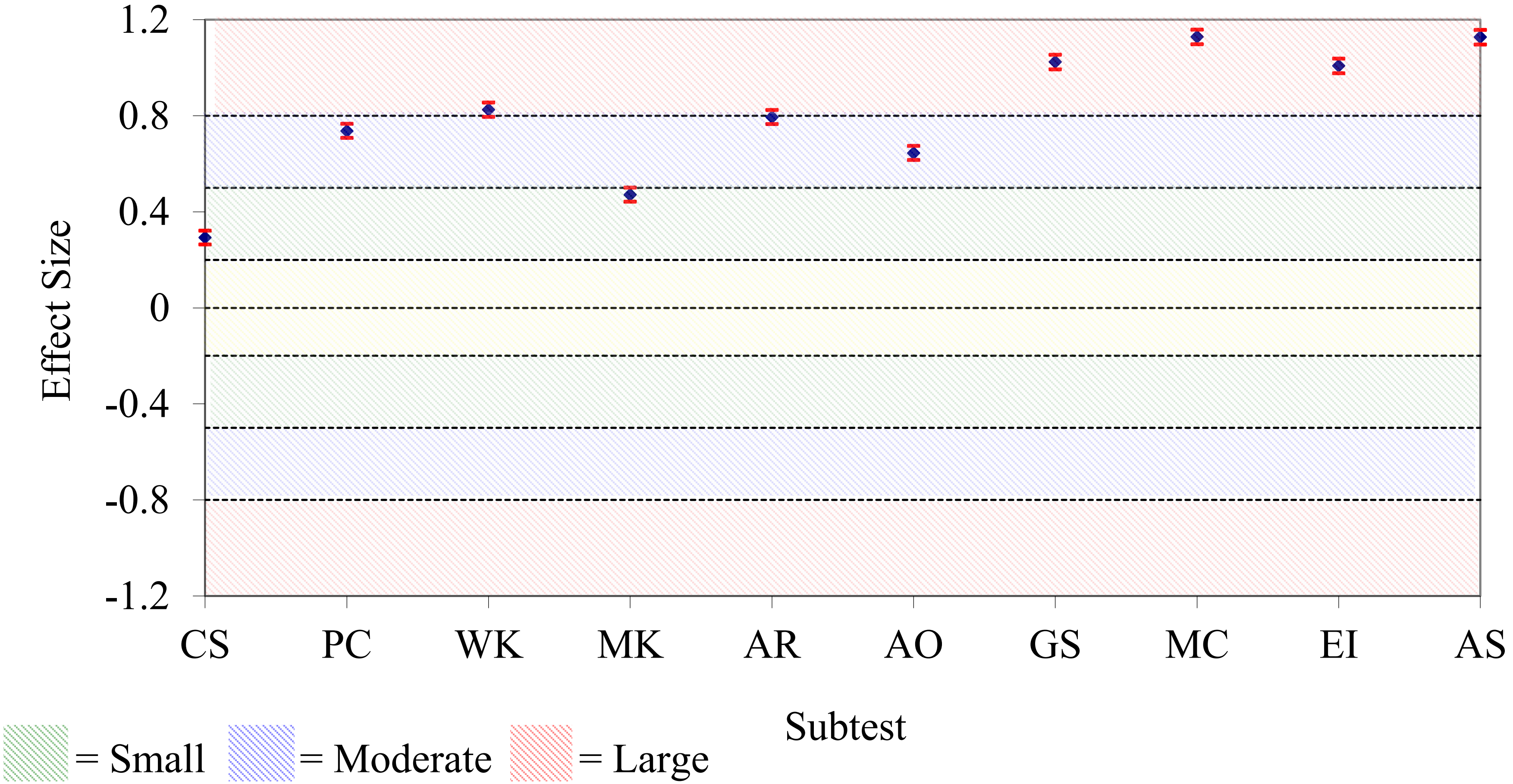
Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Non-Hispanic Whites vs. Non-Hispanic Blacks
FY2017 for Mental Counters Sample



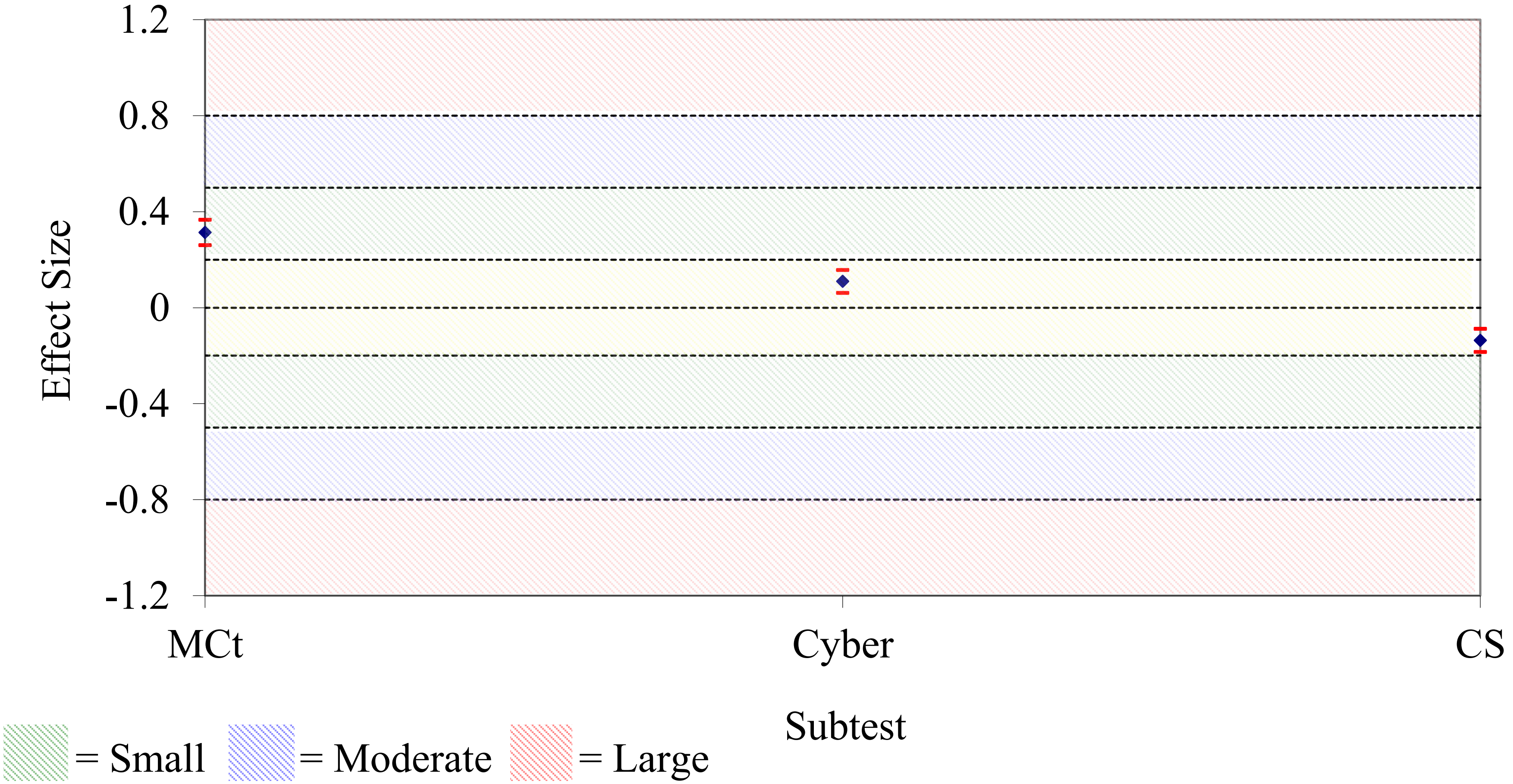
Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Non-Hispanic Whites vs. Non-Hispanic Blacks
FY2017 for Cyber Test Sample



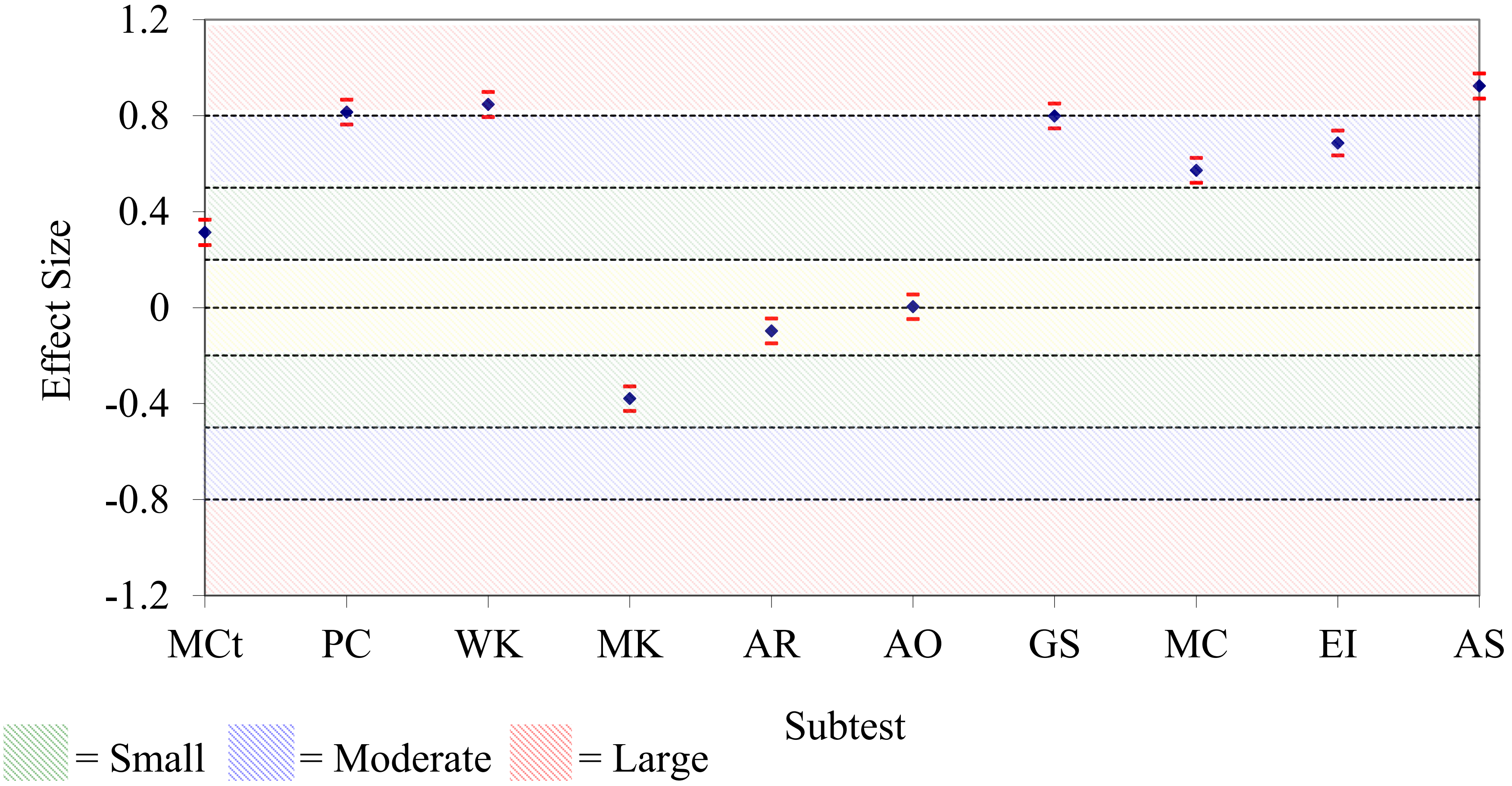
Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Non-Hispanic Whites vs. Non-Hispanic Blacks
FY2017 for Coding Speed Sample



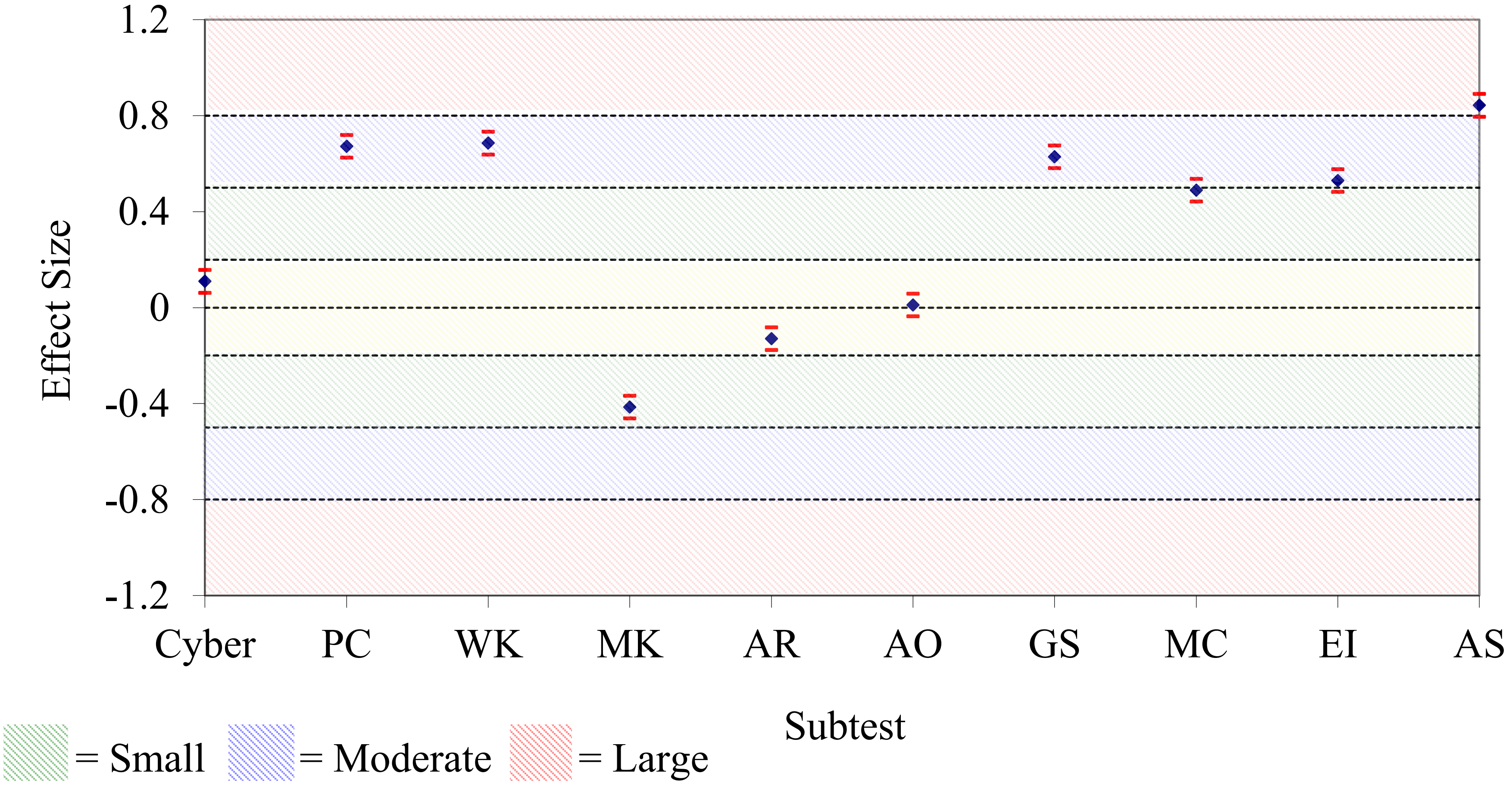
Effect Sizes (and 95% Confidence Interval) for Special Tests Scores
Non-Hispanic Whites vs. Non-Hispanic Asians
FY2017



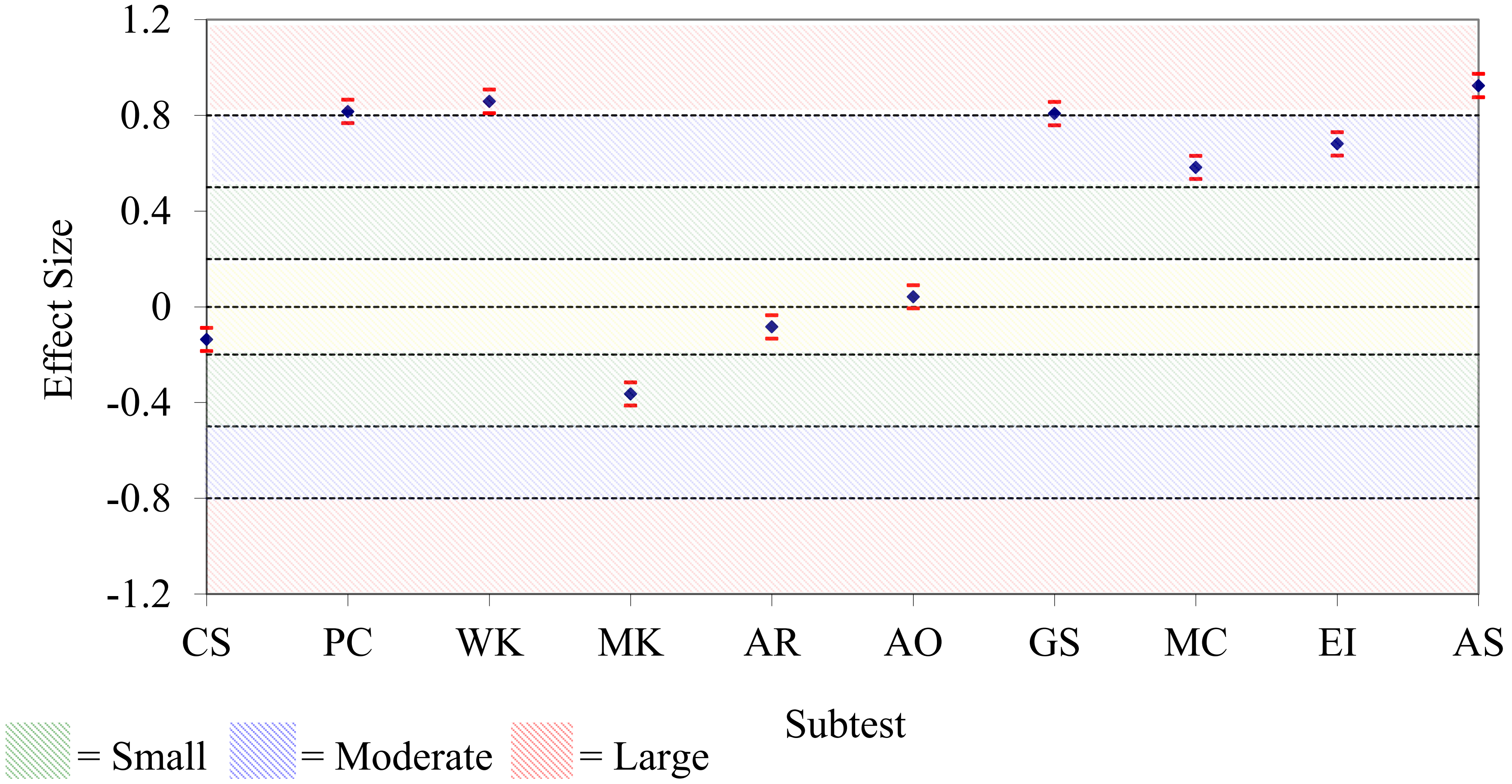
Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Non-Hispanic Whites vs. Non-Hispanic Asians
FY2017 for Mental Counters Sample



Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Non-Hispanic Whites vs. Non-Hispanic Asians
FY2017 for Cyber Test Sample



Effect Sizes (and 95% Confidence Interval) for ASVAB Scores Non-Hispanic Whites vs. Non-Hispanic Asians FY2017 for Coding Speed Sample



CONCLUSIONS

- The three special tests (MCt, Cyber, CS) generally exhibited small to moderate effects and were usually as low or lower than most ASVAB tests.
- White-Black comparisons were generally larger for MCt than for the other group comparisons.
- Coding Speed usually had very small effects (near 0), BUT, this test suffers from other issues, for example:
  - Affected by lag time in internet delivery (speeded test)
  - Known to be affected by test delivery device
  - Suffers from coachability, and susceptibility to invalid strategies that result in high scores
- Potential for adverse impact is not the only consideration for making changes to the ASVAB.

BACKUP SLIDES

SAMPLE SIZES

MCt sample	Test	N		Cyber sample	Test	N		CS sample	Test	N
Males	ASVAB	21422		Males	ASVAB	29496		Males	ASVAB	24126
	MCt	20781			Cyber	29757			CS	24126
Females	ASVAB	7743		Females	ASVAB	11128		Females	ASVAB	8551
	MCt	7442			Cyber	11147			CS	8551
NHW	ASVAB	13882		NHW	ASVAB	21591		NHW	ASVAB	15504
	MCt	13323			Cyber	21502			CS	15504
HW	ASVAB	3859		HW	ASVAB	5074		HW	ASVAB	4292
	MCt	3757			Cyber	5036			CS	4292
HispanicALL	ASVAB	4889		HispanicALL	ASVAB	6215		HispanicAL	ASVAB	5478
	MCt	4767			Cyber	6217			CS	5478
NHB	ASVAB	5987		NHB	ASVAB	8101		NHB	ASVAB	6761
	MCt	5546			Cyber	7807			CS	6761
NHA	ASVAB	1635		NHA	ASVAB	1896		NHA	ASVAB	1835
	MCt	1547			Cyber	1856			CS	1835

Tab M

OFFICE OF PEOPLE ANALYTICS

DPAC Device Evaluation for the ASVAB

DAC

03.29.2019 | Seaside, CA

Tia Fechter

DISCUSSION TOPICS

- Goals & Impact
- Existing Research Update
- DAC Role
- Device Evaluation Questions
- Evaluation Design Updates
- Pilot Preliminary Feedback
- Analysis Plan
- Discussion

GOALS & IMPACT

- Facilitate device expansion of the ASVAB iCAT and PiCAT by evaluating examinee performance differences among electronic devices (e.g., tablets, smart phones).
- Allow for more flexibility for ASVAB administration to reduce time spent in MEPS, increase number of enlistees, and increase schools' participation in CEP.
- Make a recommendation for which types of electronic devices should be approved or prohibited for ASVAB administration.
- Inform a Next Generation user interface that incorporates a Responsive Design approach, which automatically formats the test display to alternative devices.

EXISTING RESEARCH UPDATE

- **Laurie Davis (personal communication, Oct. 2018)**
  - Found performance across math, reading, and science high school exams to be similar between tablet and computer conditions
    - For reading, a small device effect favoring tablets was found for middle to lower parts of the score distribution (males tended to perform better using tablets); Davis, et.al., 2017
  - Response time is longer on tablets
  - Allow for modifications to test layout to best fit device (i.e., responsive app design)
  - Smartphones not tried but possible—don't eliminate condition
  - Optimistic that we will see comparable performance

- **USMC Observations (email communication, Sep. 2018)**
  - Delivers APT and PiCAT using tablets at recruiting commands
    - 8″ & 10″ Samsung Tab Active tablets
    - Currently not experiencing image display issues
    - No other known issues
    - DON'T KNOW IMPACT ON PERFORMANCE (i.e., scores)

EXISTING RESEARCH UPDATE

- **Additional I/O Findings of Interest**
  - Score differences between mobile and non-mobile devices were not found for personality job selection assessments
    - It did take longer for examinees to complete the tests on mobile devices
  - Score differences between mobile and non-mobile devices were found for cognitive job selection assessments
  - Measurement invariance held up for all tests administered
  - Minority groups tend to have access to internet primarily through smartphones
  - Summary of research shows that it takes longer to take tests on mobile devices
  - Job applicants report more positive reactions to taking tests on mobile devices when the delivery application is specifically designed to support mobile administration
  - More likely to encounter distractions and interruptions when taking tests on mobile devices

DAC ROLE

- Identify any barriers foreseen for implementing current evaluation design.
- Offer any feedback based on preliminary pilot.
- Offer recommendations to strengthen or support analysis plans.

DEVICE EVALUATION QUESTIONS

- Does device differentially impact examinee performance (score; response time) on ASVAB subtests?
- Does device familiarity differentially impact examinee performance on ASVAB subtests?
- Does device differentially impact item difficulty?
- Are there item features (e.g., inclusion of graphic) that interact with the device that increase the probability that item difficulty is differentially impacted?

EVALUATION DESIGN UPDATES

- **Sampling Plan**
- **Methods**

EVALUATION DESIGN—SAMPLING PLAN

- **Participants**
  - Recruits (2,330)
    - Air Force: Lackland AFB (4 SATs OCT)
    - Army: Fort Drum (25–29MAR)

Marine Corps:

  - Fort Leonard Wood (23–25APR)
  - MCCSSS (16–17APR)

- Navy: NAT Center (TBD)

- Applicants (7,010)
  - 15 medium-volume MEPS (3 months at each site; begin in May)

Examinee Group	Form ID Assignments[a]	ASVAB Subtest[b]							Test Time (minutes)[c]	Number of Items[c]	Number of Subjects
		GS	AR	WK	PC	MK	MC	AO			
1	F01/F02		X						30	12	1750
2	F03/F04		X						30	12	585
3	F05/F06						X		30	24	585
4	F07/F08							X	30	30	1750
5	F09/F10	X			X				30	30	585
6	F11/F12			X				X	30	40	585
7	F13/F14					X			30	24	585
8	F15/F16				X				28	14	585
9	**F17/F18**			**X**	**X**	**X**		**X**	**88**	**78**	**1165**
10	F19/F20	X	X		X		X		90	66	1165
TOTALS										186	9340

OFFICE OF PEOPLE ANALYTICS

EVALUATION DESIGN—METHODS

Device ID	Device Type	Model	Web Browser
1 CONTOL	**Notebook** CONDITION	**Dell XPS 13**	**Internet Explorer**
2	Notebook	Dell Chromebook 3380	Chrome
3	Notebook	Apple MacBook Pro	Safari
4	Tablet	Samsung Galaxy Tab A	Chrome
5	Tablet	Apple iPad Pro	Safari
6	Smart phone	Samsung Galaxy S9+	Chrome
7	Smart phone	Apple iPhone XS	Safari

EVALUATION DESIGN—METHODS

Device ID	**Device Type**	**Model**	**Operating System**
1 CONTOL	**Notebook** CONDITION	**Dell XPS 13**	**Windows**
2	Notebook	Dell Chromebook 3380	Chrome
3	Notebook	Apple MacBook Pro	MacOS
4	Tablet	Samsung Galaxy Tab A	Android
5	Tablet	Apple iPad Pro	iOS
6	Smart phone	Samsung Galaxy S9+	Android
7	Smart phone	Apple iPhone XS	iOS

EVALUATION DESIGN—METHODS

Device ID	Device Type	Model	Screen Size
1 CONTOL	**Notebook** CONDITION	**Dell XPS 13**	**13”**
2	Notebook	Dell Chromebook 3380	11.6”
3	Notebook	Apple MacBook Pro	13.3”
4	Tablet	Samsung Galaxy Tab A	8”
5	Tablet	Apple iPad Pro	11”
6	Smart phone	Samsung Galaxy S9+	6.2”
7	Smart phone	Apple iPhone XS	5.8”

PILOT PRELIMINARY FEEDBACK

- San Jose MEPS: 18-19 MAR
  - Monday: 18 people
  - Tuesday: 6 people
  - May not get 50/week as planned
- Chromebook issue with next button
- WiFi: successfully accessed in closed room once router was moved closer to door
- Recruiting strategy – 2 trialed
- Test Administrator Expectations
  - May not be familiar with a variety of devices to offer support to examinees
  - Many don’t have computers at home
  - Anticipate on-site training
- Use of TCOs – MEPS staff very eager and helpful

ANALYSIS PLAN

- Does device differentially impact examinee performance (score; response time) on ASVAB subtests?
  - Conduct MANOVA (after equating the two parallel forms) across all device conditions
    - Dependent Variables: Equated Subtest Score, Response Time
    - Independent Variable: Device
  - 7 MANOVAs – one for each subtest
    - If the F-test is not significant, no further analysis is needed
    - If the F-test is significant, post hoc analyses are needed to determine where the differences are

ANALYSIS PLAN

- Does device familiarity differentially impact examinee performance on ASVAB subtests?
  - Conduct t-test between subtest scores pooling across device conditions those reporting familiarity with the device used and those not reporting familiarity with the device used.
    - Independent Variable: Familiarity (self reported)
      - **Which electronic devices are you comfortable using? Please select all that apply.**
  - Repeat t-test between response times
  - 14 t-tests – one for each subtest and dependent variable
    - If the t-test is not significant, no further analysis is needed and familiarity groups can be pooled
    - If the t-test is significant, consider modifying the design to test device effect by adding a categorical covariate (e.g., MANCOVA)
  - Plan to conduct this analysis before conducting the MANOVAs

ANALYSIS PLAN

- Does device differentially impact item difficulty?
  - Conduct multi-group IRT calibration and compare item difficulty values (as would be done for a DIF analysis)
    - Groups: 7 device conditions
    - Note any items flagged for DIF

- Are there item features (e.g., inclusion of graphic) that interact with the device that increase the probability that item difficulty is differentially impacted?
  - Of the items noted for DIF, explore whether there are patterns based on item features that may explain the differences detected

DISCUSSION

- Foreseen barriers for implementing current evaluation design
- Feedback based on preliminary pilot
- Recommendations to strengthen or support analysis plans

Tab N

U.S. Army Research Institute
for the Behavioral and Social Sciences

Development of the Adaptive Vocational Interest Diagnostic (AVID)

Briefing for

Defense Advisory Committee on
Military Personnel Testing

29 March 2019

Dr. Cristina Kirkendall, cristina.d.kirkendall.civ@mail.mil, 703-545-2431
Dr. Tonia Heffner, tonia.s.heffner.civ@mail.mil, 703-545-4408

Vocational Interests for MOS Assignment

- The U.S. Army has approximately 140 entry-level Military Occupation Specialties (MOS)
  - For an interest assessment to be useful for classification, it would need to be applicable to all of these MOS
  - In a survey of over 24,000 U.S. Soldiers, “Perceived fit” with MOS was the top reason for selecting an MOS
- Potential benefits of a vocational interest inventory:
  - Provide recruits information about the MOS in which they will be most successful
  - Predict valued work outcomes across the Army
  - Minimize the negative effects of poor Soldier-MOS fit

Research Goal: Develop a new generation vocational interest assessment that incorporates recent research and advanced statistical techniques

Adaptive Vocational Interest Diagnostic (AVID)

AVID - New generation interest assessment

AVID Background

- Most major interest measures assess only six primary dimensions
  - Most Army jobs cluster into 1 or 2 or these dimensions
  - Broad interest dimensions may not be flexible enough to select and classify Soldiers across a wide range of jobs
- To increase assignment potential:
  - Develop an assessment of basic interests that may be more useful for differentiating across jobs
  - Focus on identifying a comprehensive list of basic interest dimensions that would be useful in the Army

AVID Characteristics

- Item response theory (IRT) based, computer-adaptive assessment
  - Forced-choice format
  - Will more accurately measure the range of interests
  - Easily customized to predict performance across a broad range of jobs
  - Reduced testing time

AVID Interest Dimensions

- Identification of basic interest scales to use in AVID
  - Based on:
    - Review of previous ARI work on interests
    - Review of literature on basic interest dimensions
- 20 Basic interest dimensions were identified for pretesting

- **Realistic**
  - Construction
  - Protection
  - Combat
  - Physical activity
  - Mechanical
  - Electronics
  - Outdoor
- **Artistic**
  - Writing
- **Investigative**
  - Medical services
  - Mathematics
  - Science
  - Information technology
- **Social**
  - Teaching
  - Personal service
- **Enterprising**
  - Leadership
  - Sales
  - Human relations
- **Conventional**
  - Office work
  - Finance
  - Food service

AVID Pretesting

- Item development and pretesting
  - Wrote approximately 1,000 test statements (~50 items per dimension)
  - Pretesting data was collected from approximately 3,300 enlisted Soldiers in Reception Battalions and BCT
    - Pretesting established item parameters and social desirability ratings
    - Examined the correlations between the AVID scales and another interest measure to establish construct validity
      - Compared scores to the O*Net Interest Profiler
      - Results generally confirmed the construct validity of the AVID dimensions

- Outcome: Developed both static and adaptive forms of AVID

AVID Initial Validation

- Static form of AVID includes 123 item pairs to assess 16 of the 20 AVID dimensions
  - Science, Personal Service, Finance, Sales were excluded to reduce total testing time
- Collected data from two samples for the validation study
  - Sample 1: Data collection focused on four high-density MOS:
    - Military Police (n = 287)
    - Combat Medics (n = 273)
    - Motor Transport Operators (n = 529)
    - Wheeled Vehicle Mechanics (n = 457)
    - Given the sample size, we also examined the other Healthcare MOS as a group (n = 116)
  - Sample 2: Data collected from 1,999 Soldiers as part of another project
    - Majority were E-3 (29%) or E-4 (47%)
    - Largest MOS was Infantry (N = 343)

AVID Initial Validation

- AVID and Army Life Questionnaire (ALQ) data were collected in both samples.
- Also collected Soldiers' ratings of their MOS:

Work Dimension	Rate your current MOS on the work dimension described below using the following seven-point scale. Ask yourself, "How descriptive is this dimension of my MOS?"
Construction	**Description:** Involves designing and/or building things, maintaining structures with one's hands, or using tools and materials. **Example Activities:** • Building roads and bridges • Designing and constructing buildings • Working on a construction site • Pouring concrete • Using power tools • Welding

75. How descriptive is this dimension of your current MOS?

Not at all descriptive			Moderately descriptive			Extremely descriptive
A	B	C	D	E	F	G

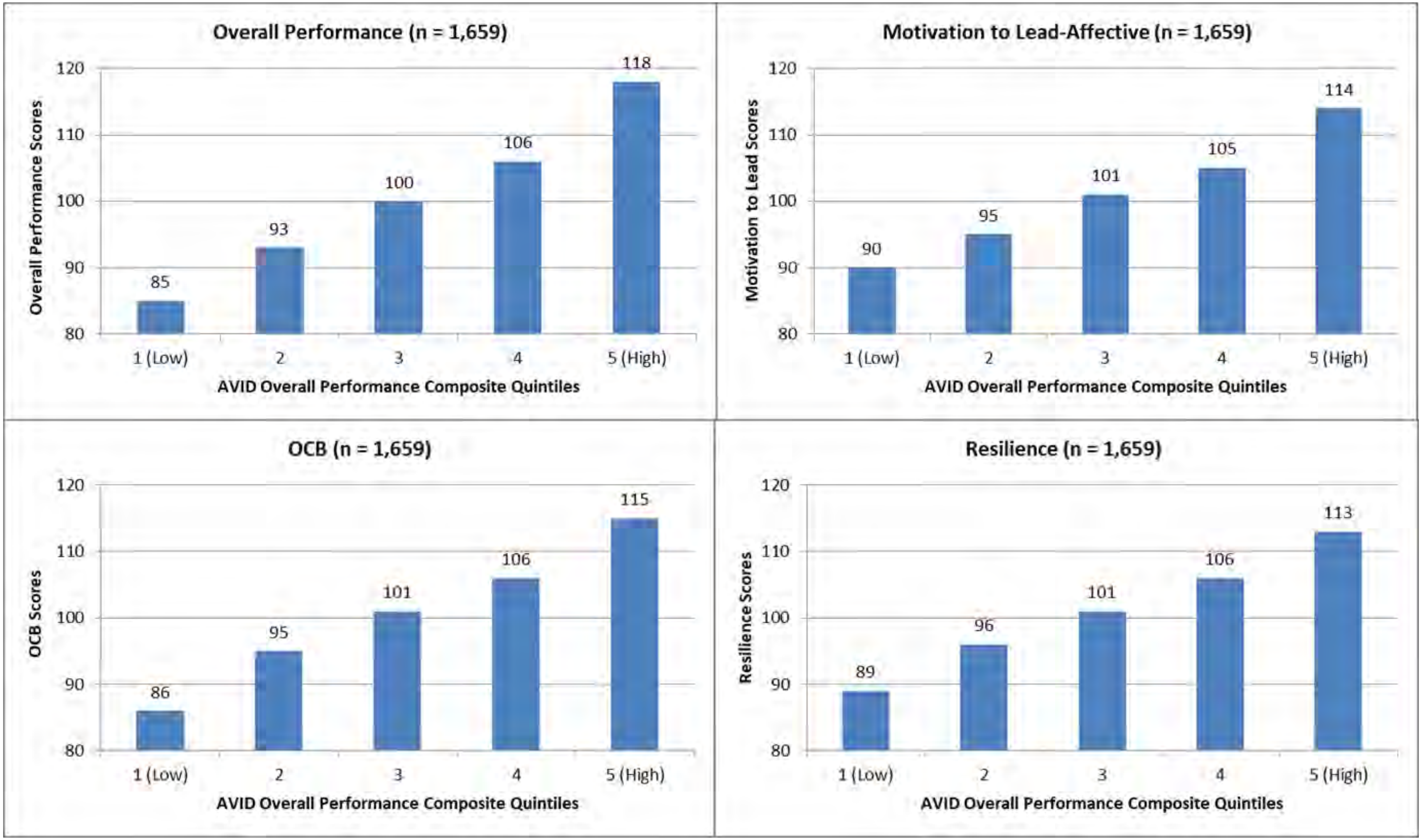
AVID Initial Validation

- Data were cleaned using items to detect unmotivated responding (e.g., “Select option B”)
  - Excluded 124 Soldiers in Sample 1 and 218 in Sample 2
- Analyzed the data using correlation and regression analyses
- The validity of vocational interests is highest when considering the match between individuals and their jobs (i.e., interest fit)
  - Analyses focused on identifying the validity of AVID using Soldiers’ fit with their MOS

AVID Initial Validation Results: Sample 1

	MOS Fit	Army Fit	Affect. Commit	OCB	Res-ilience	Reenlist Intent	Mot. to Lead	APFT	Overall Perf.
Combat		.08	.08	.09	.10				**.07**
Construction		-.07	-.06			-.06			
Electronics									
Food Service	.06								
Human Relations		.06		.09	.10				**.10**
Information Tech.	-.10	-.06	-.09			-.07			**-.08**
Management	.10	.13	.11	.22	.16	.16	.32		**.24**
Mathematics	.05							.09	
Mechanical	.18							-.09	
Medical Services	-.06				.07				
Office Work			.06						
Outdoors									
Physical Activity	.05	.11	.05	.06	.18	.09		.29	**.15**
Protection		.07	.09	.07		.07	.09		**.10**
Teaching	.07	.06	.06	.14	.09		.09		**.12**
Writing		-.08		-.13	-.11		-.12		**-.15**
Multiple R	.29	.31	.27	.40	.39	.28	.42	.31	**.47**
Adjusted R	.28	.29	.25	.39	.38	.26	.41	.29	**.46**

AVID Initial Validation Results: Sample 1



Overall Performance (n = 1,659)

Motivation to Lead-Affective (n = 1,659)

OCB (n = 1,659)

Resilience (n = 1,659)

AVID Initial Validation Results: Sample 1

	Military Police	Combat Medic	Health-care	Transport Operator	Mechanic
Combat	.13				
Construction					
Electronics					
Food Service					
Human Relations	.15			.11	.17
Information Tech.					-.10
Management	.32	.22	.25	.24	.20
Mathematics			.09		
Mechanical	-.12				.13
Medical Services		.16	.09		
Office Work					
Outdoors					
Physical Activity	.14			.11	.21
Protection		.15	.15		
Teaching		.12	.19		
Writing	-.13	-.13	-.17	-.14	
Multiple R	.57	.50	51	.43	.49
Adjusted R	.52	.44	.47	.38	.45

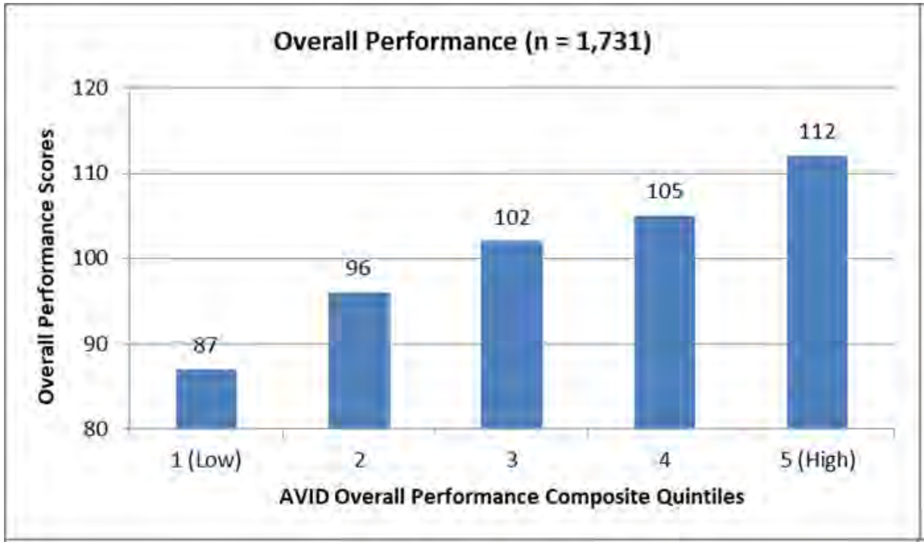
Results are based on models predicting overall performance in each MOS

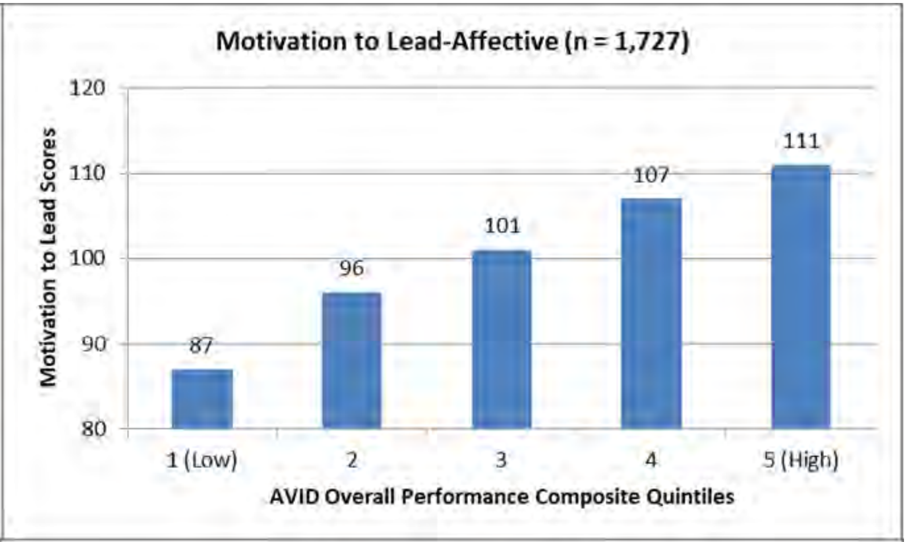
Values represent standardized regression weights for predicting each outcome. Sample sizes ranged from 262 (Combat Medic) to 449 (Transport Operator)

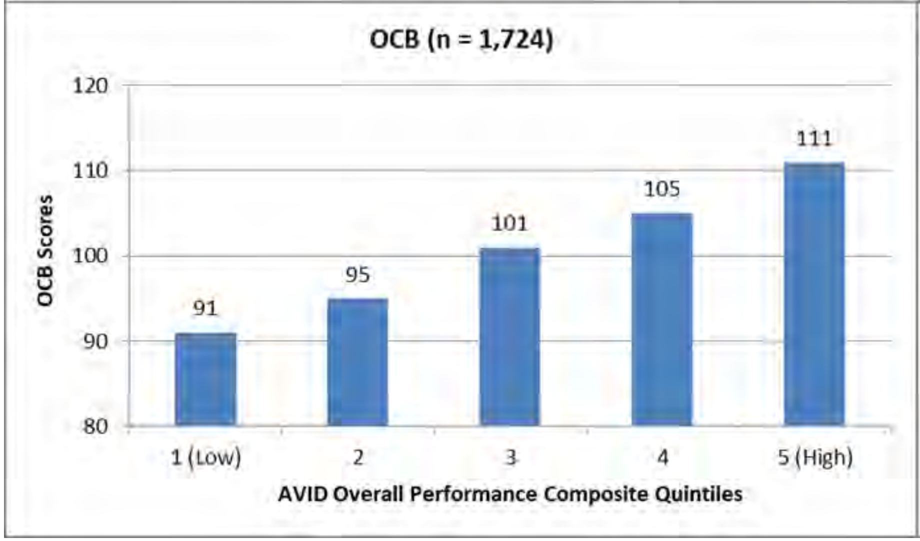
AVID Initial Validation Results: Sample 2

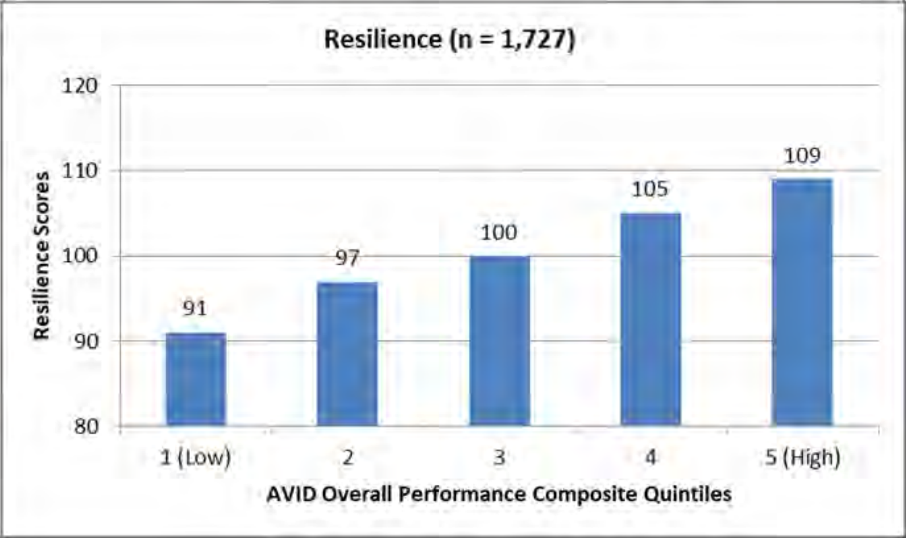
	MOS Fit	Army Fit	Affect. Commit	OCB	Res-ilience	Reenlist Intent	Mot. to Lead	APFT	Overall Perf.
Combat	.07	.07		.07		.07			**.08**
Construction		-.06	-.09			-.08	.05		**-.05**
Electronics									
Food Service				-.04	-.06			-.07	
Human Relations		.07	.07	.11	.07		.06		**.11**
Information Tech.									
Management	.06	.08	.11	.26	.12	.11	.42	.05	**.24**
Mathematics				.06	.05		.07	.06	
Mechanical	.07		.06			.06		-.06	
Medical Services	-.07					.05		.05	
Office Work		-.05		-.06	-.07			-.07	**-.05**
Outdoors	.05			.06					
Physical Activity	.05	.16	.09	.11	.27	.11	.09	.28	**.21**
Protection		.05	.10		.05		.06		**.06**
Teaching		.06		.10		.05			**.06**
Writing		-.09			-.05	-.05	-.05		**-.07**
Multiple R	.22	.31	.26	.42	.38	.23	.51	.32	**.47**
Adjusted R	.19	.29	.24	.41	.37	.21	.51	.31	**.46**

AVID Initial Validation Results: Sample 2


Overall Performance (n = 1,731)


Motivation to Lead-Affective (n = 1,727)


OCB (n = 1,724)


Resilience (n = 1,727)

AVID Initial Validation Results: Sample 2

	Infantry	Full Sample
Combat		.08
Construction	.17	-.05
Electronics		
Food Service		
Human Relations	.12	.11
Information Tech.	-.16	
Management	.18	.24
Mathematics		
Mechanical		
Medical Services		
Office Work		-.05
Outdoors		
Physical Activity	.26	.21
Protection		.06
Teaching		.06
Writing		-.07
Multiple R	.53	.47
Adjusted R	.47	.46

Values represent standardized regression weights. Sample sizes ranged from 215 to 1,731.

Results are based on models predicting overall performance in Infantry and in the full sample

AVID Initial Validation Results

- Examined the validity of interest fit for predicting overall performance in each MOS
  - Interest fit was operationalized using regression models with both AVID dimensions and MOS interest scores in the model.
  - Results indicated higher validities using this operationalization of fit.

		Military Police	**Combat Medic**	**Health-care**	**Transp. Op.**	**Mechanic**	**Full Sample 1**	**Infantry**	**Full Sample 2**
Without Interest Fit	Multiple R	.57	.50	51	.43	.49	.47	.53	.47
	Adjusted R	**.52**	**.44**	**.47**	**.38**	**.45**	**.46**	**.47**	**.46**
With Interest Fit	Multiple R	.70	.69	.64	.56	.63	.57	.64	.52
	Adjusted R	**.64**	**.62**	**.59**	**.51**	**.59**	**.56**	**.55**	**.51**

Sample sizes ranged from 215 (Infantry) to 1,731 (Full Sample 2).

Summary of Initial Validation

- The validities of the AVID were often larger than for the TAPAS when interest fit was calculated
- The differences across MOS suggest that the AVID may be useful for MOS classification
  - Correlations between MOS-specific composites of AVID scales were strong but indicated differences across MOS

	Military Police	Combat Medic	Health-care	Transport Operator	Mechanic
Military Police	1.00				
Combat Medic	.65	1.00			
Healthcare	.72	.80	1.00		
Transport Operator	.75	.78	.75	1.00	
Mechanic	.45	.59	.64	.70	1.00

- This provides a useful initial look at the AVID but more research is needed to examine validity in a broader range of MOS and to evaluate the adaptive version.

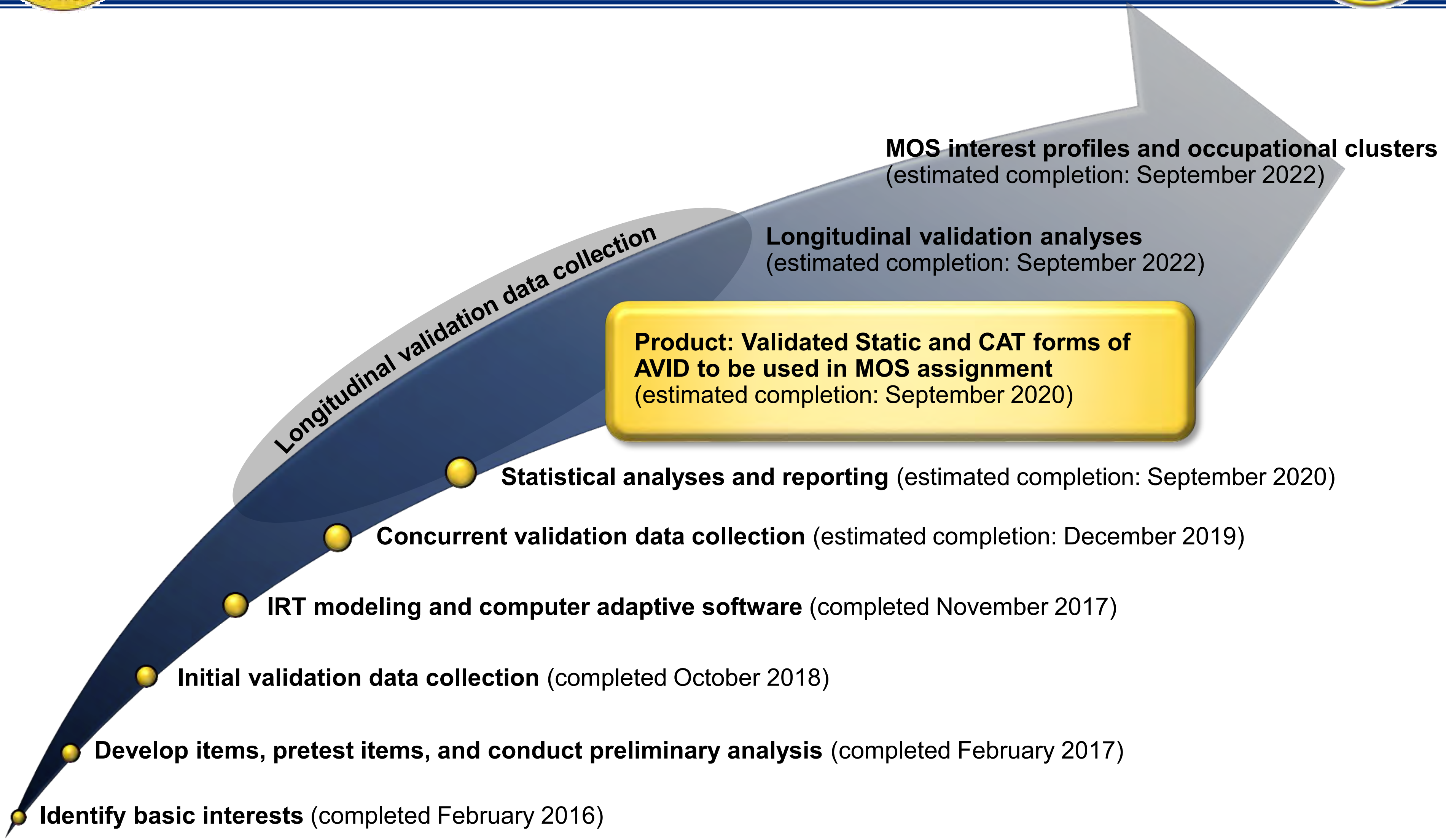
Next Steps

- Continued concurrent validation of AVID
  - Collect additional validity evidence for the static AVID form
    - Collect data on the four additional AVID dimensions that have not yet been examined (Science, Personal Service, Sales, Finance)
    - Target five different MOS:
      - Combat Engineer
      - Cavalry Scout
      - Human Resources Specialist
      - Automated Logistical Specialist
      - Unit Supply Specialist
- Conduct simulations to evaluate the best method of calculating the match between interest dimensions and job characteristics
- We will also examine the implications of “fit bandwidth” (i.e., some individuals will be interested in many jobs vs. others who are only interested in one job)

Next Steps – Potential tasks

- MOS Interest Profiles and Occupational Clusters
  - Data collected as part of the validity studies includes ratings of MOS characteristics; however, we only have ratings on a narrow range of MOS.
  - Ratings will be used to explore clusters of MOS with similar interest profiles
    - Can we identify job families with similar interests?
- Longitudinal validation
  - AVID data will be collected from early career Soldiers (e.g., at Reception Battalions)
    - These data will then be linked to end-of-training data
    - Outcomes will include 6-month attrition, the ALQ, and performance ratings

AVID Project Timeline



MOS interest profiles and occupational clusters (estimated completion: September 2022)

Longitudinal validation analyses (estimated completion: September 2022)

Longitudinal validation data collection

Product: Validated Static and CAT forms of AVID to be used in MOS assignment (estimated completion: September 2020)

Statistical analyses and reporting (estimated completion: September 2020)

Concurrent validation data collection (estimated completion: December 2019)

IRT modeling and computer adaptive software (completed November 2017)

Initial validation data collection (completed October 2018)

Develop items, pretest items, and conduct preliminary analysis (completed February 2017)

Identify basic interests (completed February 2016)

Summary

- AVID has the potential to be a valuable addition to ARI's non-cognitive measures and contribute to a whole-person assessment that more accurately predicts performance, behaviors, and attitudes
- Improved Personnel Assessment:
  - Enables greater flexibility to accommodate changes in force size, structure, mission demands, budget and availability of qualified applicants
  - Improves person-job match, performance, and retention
  - Saves money by reducing attrition

Whole-person assessment requires cognitive and non-cognitive measures

Tab O

OFFICE OF PEOPLE ANALYTICS

ASVAB Subtest Time Limit Analyses

Presented to the DAC

Furong Gao, HumRRO
Mary Pommerich, DPAC
Dan Segall, DPAC
March 29, 2019 | Carmel-by-the-Sea, CA

OVERVIEW

- **The purpose of these analyses is to evaluate current ASVAB testing time limits to ensure examinees have sufficient time to complete the tests**
- **Current time limits**
- **Time limit analyses**
  - Examinees' actual response time distributions
  - Sufficient time allocated?
- **Time adjustment**
  - Fit theoretical statistical distribution to the observed distributions
  - Examine the 95th, 98th, 99th quantiles of the fitted distribution
  - Compare with the current time limits
  - Propose adjustment

CURRENT TESTING TIME LIMITS

- **Set from analyses with data collected in the forms 5–9 equating study**
  - Time limits for MK, AI, SI, and AO increased from those used for forms 1–4
- **In place since the initial forms 5–8 implementation (2009) for tests without seeding and 2014 for tests with seeding**
- **Number of seed items is 15 per subtest, if administered**

	# of	**Time Limit (in minutes)**	
	Op. Items	**Without Seed Items**	**With Seed Items**
General Science (GS)	15	8	16
Arithmetic Reasoning (AR)	15	39	78
Word Knowledge (WK)	15	8	16
Paragraph Comprehension (PC)	10	22	55
Mathematics Knowledge (MK)	15	20	40
Electronics Information (EI)	15	8	16
Auto Information (AI)	10	7	18
Shop Information (SI)	10	6	15
Mechanical Comprehension (MC)	15	20	40
Assembling Objects (AO)	15	16	32
Total	***145***	***154***	

SEEDING CONFIGURATION

- **Examinees are randomly assigned to one of five groups**
  - For example, group 01 examinees take 15 GS seed items and 15 AR seed items
- **Seed items are randomly dispersed among the operational items in the subtest**

Examinee Group	Take Seed Items in
01	GS, AR
02	WK, PC
03	WK, MK, EI, AI
04	WK, MK, AO
05	SI, MC, AO

TIME LIMIT ANALYSES

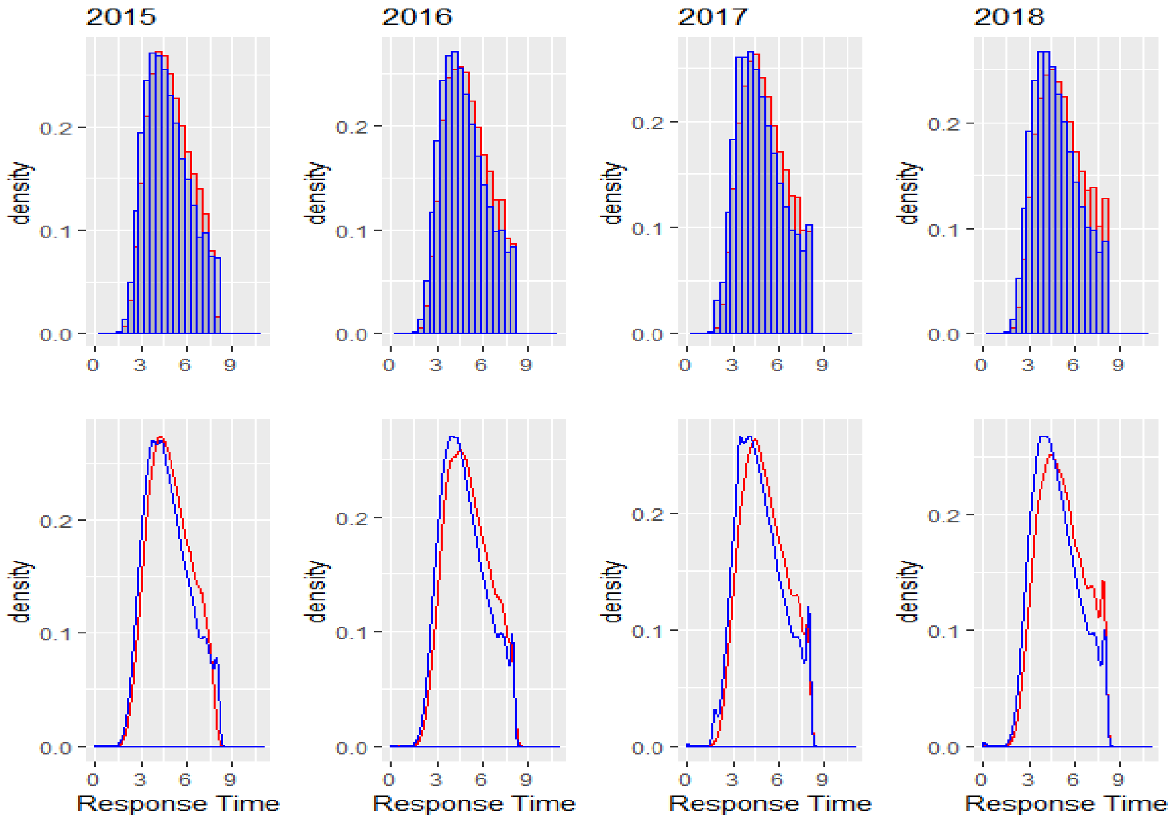
- **Questions**
  - Is sufficient time allocated for each subtest? Will adjustments be needed? If so, how to adjust?
- **Approach**
  - Examine empirical response time distribution
  - Fit a theoretical statistical distribution/model
  - To ensure that examinees have sufficient time to finish the test, look at the various quantiles of the statistical distribution and evaluate with regard to the current time limit.

TIME LIMIT ANALYSES, CONT’D

- **Data**
  - 2015–2018 (for 2015, the new seeding design was implemented in WinCAT first)
  - WinCAT and iCAT
- **The analyses were conducted for each subtest by examining examinees’ response time distribution on the test and by**
  - seeding status: with or without taking the seed items
  - year of testing and across years
    - 2015–2018 data were analyzed
  - test administration platforms and across platforms
    - iCAT
    - WinCAT
  - Combining platforms across years
    - iCAT and WinCAT data 2015–2018
- **The analyses were focused on looking at trends and variation across years and platforms**

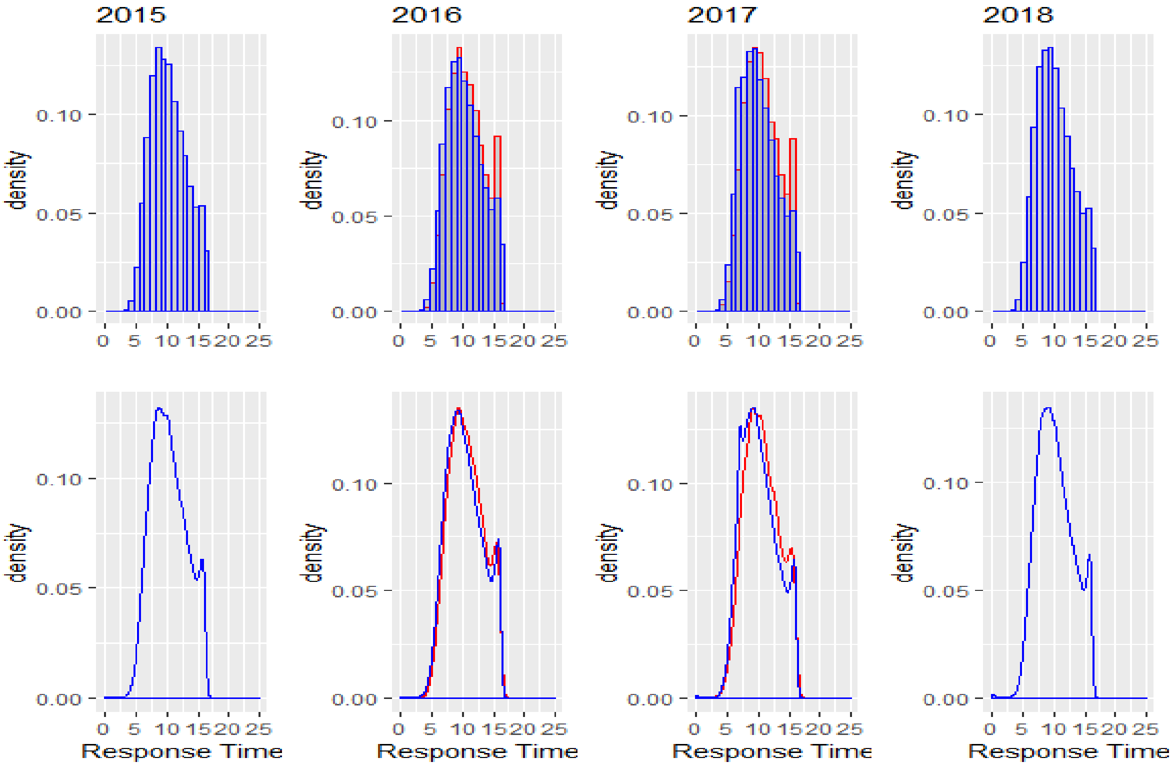
EXAMINEES' RESPONSE TIME DISTRIBUTIONS FOR GS

- **GS without seed items, by year and platform**
  - Blue: WinCAT
  - Red: iCAT
- **Similar across years**
- **iCAT examinees tend to take slightly longer than those who take the test on WinCAT**



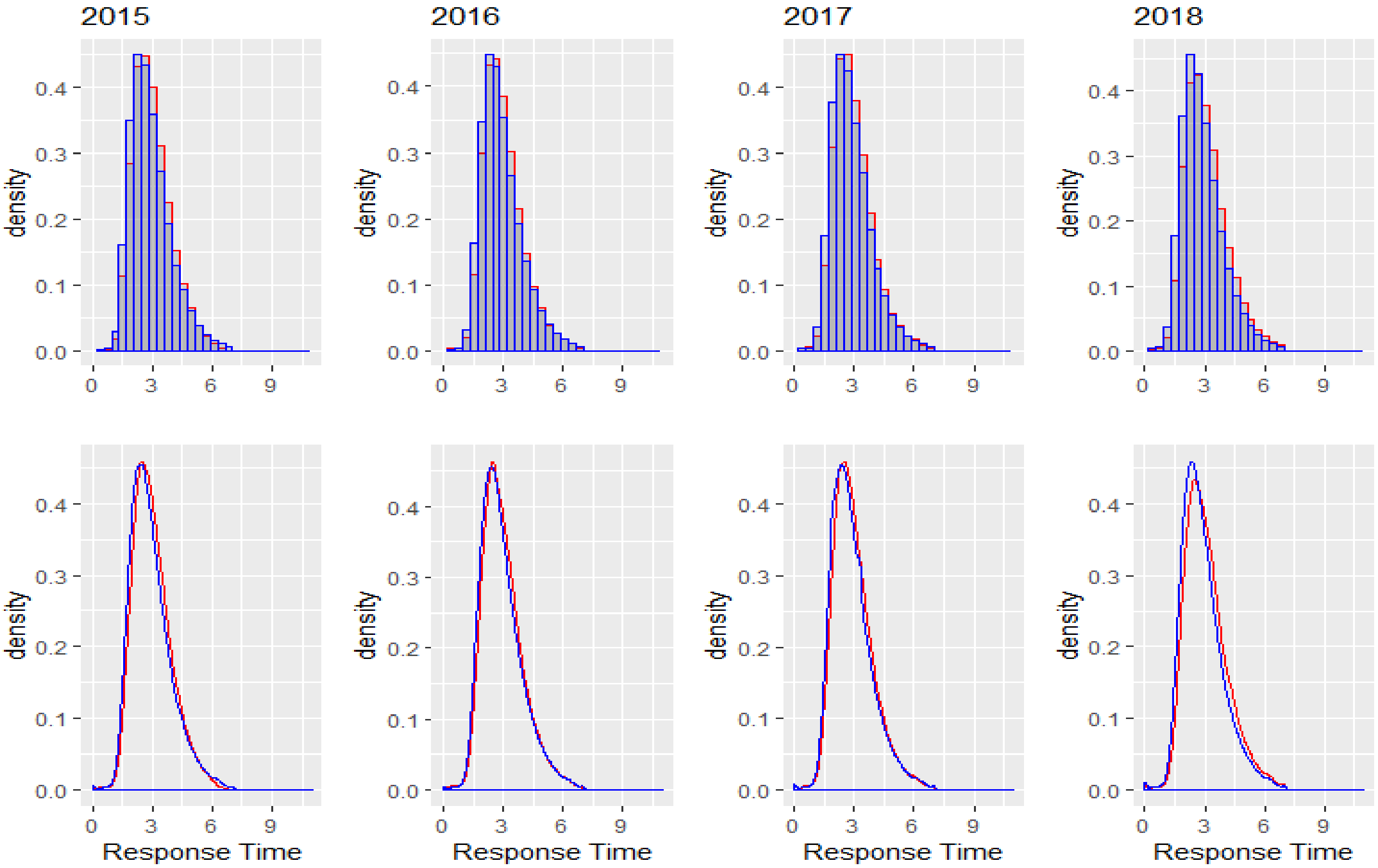
EXAMINEES' RESPONSE TIME DISTRIBUTIONS FOR GS

- **GS with seed items, by year and platform**
  - No seed items in 2015 iCAT
  - iCAT 2018 used special seeding (WK items only), so no iCAT data for GS
- **Similar across years**



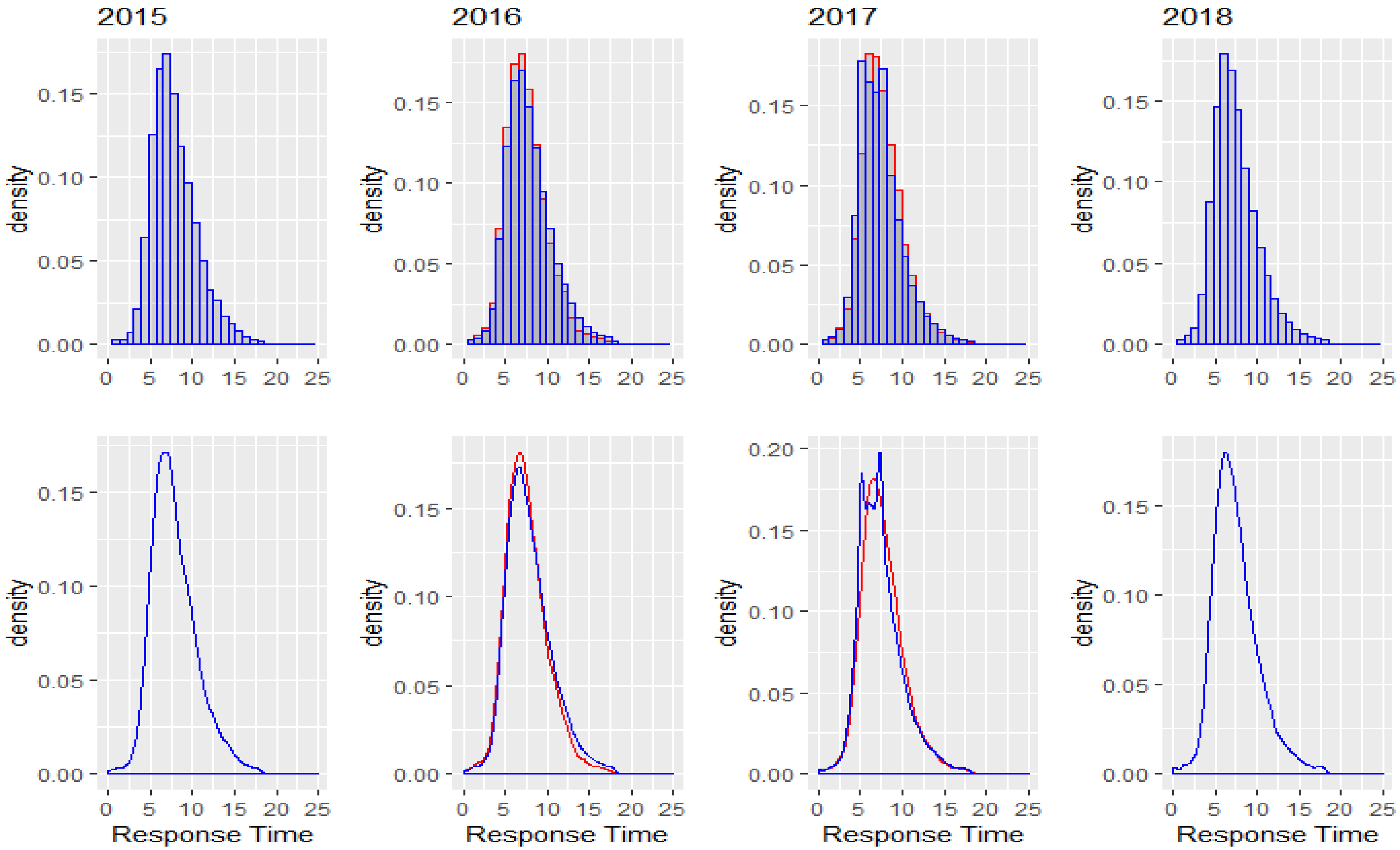
EXAMINEES' RESPONSE TIME DISTRIBUTIONS FOR AI

- **AI without seed items**



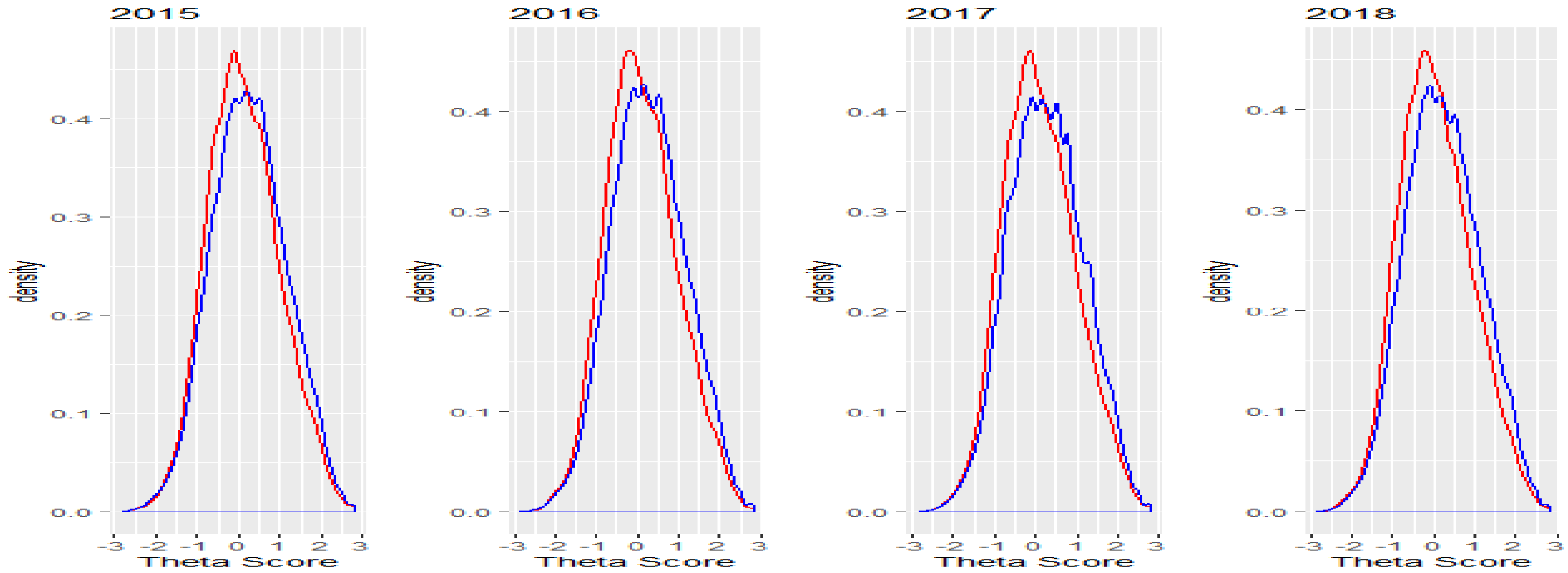
EXAMINEES' RESPONSE TIME DISTRIBUTIONS FOR AI

- AI with seed items



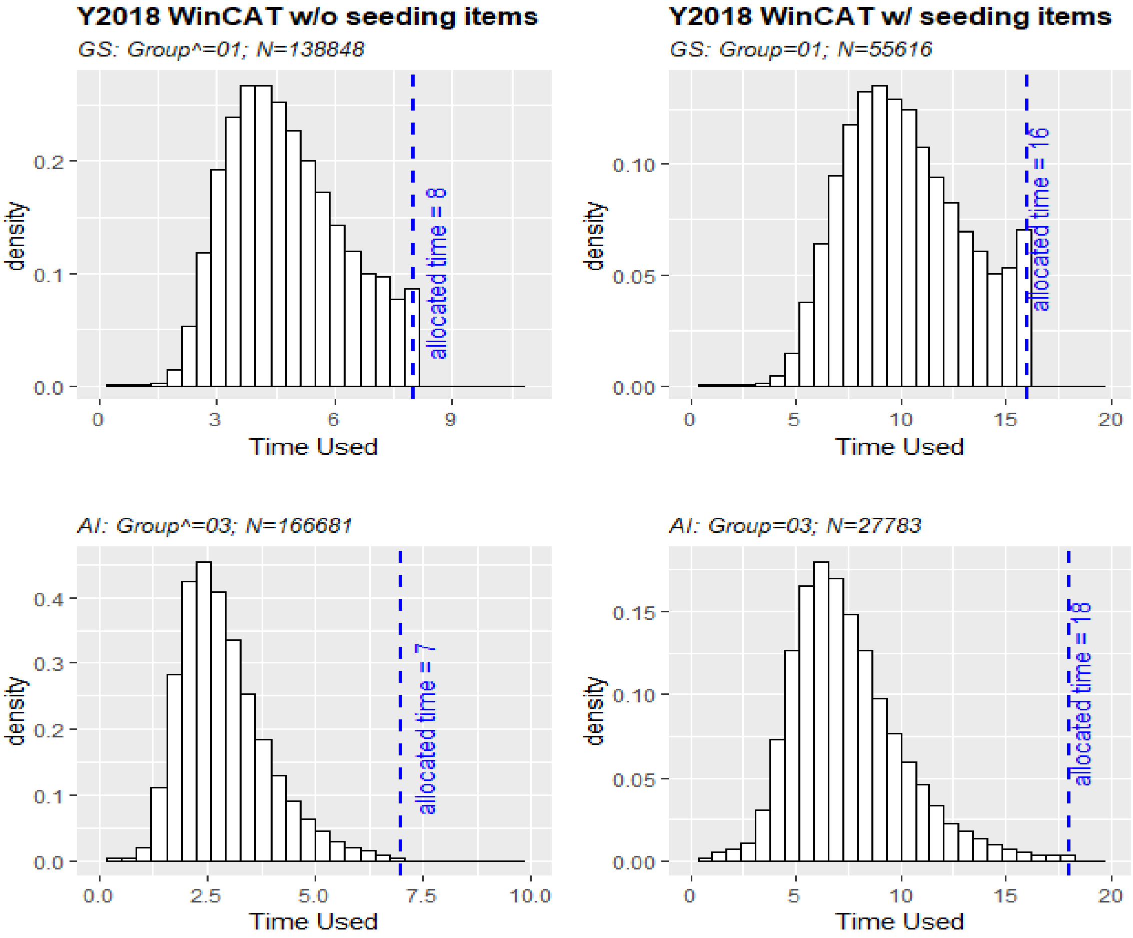
WHY DO iCAT EXAMINEES TAKE LONGER?

- Do we need to set different time limits for iCAT?
- Comparison of score distribution for GS
  - iCAT (red) vs. WinCAT (blue)
- Examinees who completed the tests
- iCAT examinees seem to have slightly lower abilities
- The longer time needed with iCAT is likely due to the population difference; therefore, no need to set different time limits across WinCAT and iCAT



EXAMINEES' RESPONSE TIME DISTRIBUTIONS

- **Examination of the empirical examinees' response time distributions in each subtest indicates that**
  - Some tests might be slightly speeded (e.g., GS)
  - Some tests might require less time than currently allocated (e.g., AI)



Y2018 WinCAT w/o seeding items

Y2018 WinCAT w/ seeding items

GS: Group^=01; N=138848

GS: Group=01; N=55616

AI: Group^=03; N=166681

AI: Group=03; N=27783

OBSERVED POSSIBLE SPEEDEDNESS AND IMPACT ON SCORES

- **Possible speededness observed in GS, AR, MK; more severe in MK**
- **If a subtest is not completed by an examinee, a "penalized" theta score is estimated for the examinee**
  - Non-completed items were scored as though they were answered at random

Proportion of examinees who didn't finish the tests

Year	Platform	GS	AR	WK	PC	MK	EI	AI	SI	MC	AO
2015	WinCAT	0.02	0.02	0.00	0.01	0.03	0.01	0.00	0.00	0.00	0.01
	iCAT	0.04	0.02	0.01	0.01	0.06	0.02	0.00	0.01	0.00	0.01
	Combined	**0.03**	**0.02**	**0.01**	**0.01**	**0.04**	**0.02**	**0.00**	**0.00**	**0.00**	**0.01**
2016	WinCAT	0.02	0.02	0.00	0.01	0.03	0.01	0.00	0.00	0.00	0.01
	iCAT	0.03	0.02	0.01	0.01	0.04	0.01	0.00	0.00	0.00	0.03
	Combined	**0.02**	**0.02**	**0.00**	**0.01**	**0.03**	**0.01**	**0.00**	**0.00**	**0.00**	**0.01**
2017	WinCAT	0.03	0.02	0.00	0.01	0.03	0.01	0.00	0.00	0.00	0.01
	iCAT	0.04	0.02	0.01	0.01	0.04	0.01	0.00	0.00	0.00	0.02
	Combined	**0.03**	**0.02**	**0.00**	**0.01**	**0.03**	**0.01**	**0.00**	**0.00**	**0.00**	**0.01**
2018	WinCAT	0.02	0.02	0.00	0.01	0.04	0.01	0.00	0.00	0.00	0.01
	iCAT	0.04	0.02	0.01	0.01	0.05	0.02	0.00	0.01	0.00	0.01
	Combined	**0.03**	**0.02**	**0.00**	**0.01**	**0.04**	**0.01**	**0.00**	**0.00**	**0.00**	**0.01**

OBSERVED POSSIBLE SPEEDEDNESS AND IMPACT ON SCORES (CONT'D)

Proportion of examinees who didn't finish the subtests: WinCAT data

Year	Seeding Status	GS	AR	WK	PC	MK	EI	AI	SI	MC	AO
2015	Without	0.02	0.02	0.00	0.01	0.03	0.01	0.00	0.00	0.00	0.01
	With	0.02	0.01	0.00	0.01	0.03	0.01	0.00	0.00	0.00	0.01
2016	Without	0.02	0.02	0.00	0.01	0.03	0.01	0.00	0.00	0.00	0.01
	With	0.02	0.01	0.00	0.01	0.03	0.01	0.00	0.00	0.00	0.01
2017	Without	0.03	0.02	0.00	0.01	0.03	0.01	0.00	0.00	0.00	0.01
	With	0.02	0.02	0.00	0.01	0.04	0.01	0.00	0.00	0.00	0.01
2018	Without	0.02	0.02	0.00	0.01	0.04	0.01	0.00	0.00	0.00	0.01
	With	0.02	0.02	0.00	0.01	0.04	0.01	0.00	0.00	0.00	0.01

OBSERVED POSSIBLE SPEEDEDNESS AND IMPACT ON SCORES (CONT'D)

- **Impact on AFQT scores**
  - Among the 932,746 examinees (across WinCAT and iCAT, from 2015-2018), 36,294 (~3.9%) of them received a lower AFQT score due to incompletion (but finished at least 2/3 of the items in each test) of at least one of the 4 subtests (AR, MK, WK, PC).

Distribution of the AFQT Score Differences with and without Penalty due to Incompletion

	1	2	3	4	5	6	7	8	9	10+
N	22786	9049	2892	996	346	128	58	23	5	11
Proportion	0.02	0.01	0	0	0	0	0	0	0	0

OFFICE OF PEOPLE ANALYTICS

OBSERVED POSSIBLE SPEEDEDNESS AND IMPACT ON SCORES (CONT'D)

- **Impact on classification: AFQT cut score 31:**
  - 1041 (2.9% of the 36,294 who received a lower AFQT scores) is 0.1% of the total 932,746.

		Without Penalty		
		< 31	**>= 31**	**rowTotal**
With Penalty	**< 31**	8824 (24.3%)	1041 (2.9%)	9865
	>= 31	0 (0%)	26429 (72.8%)	26429
	columnTotal	8824	27470	***36294***

- **Impact on classification: AFQT cut score 50:**
  - 879 = 0.09% of the total 932,746

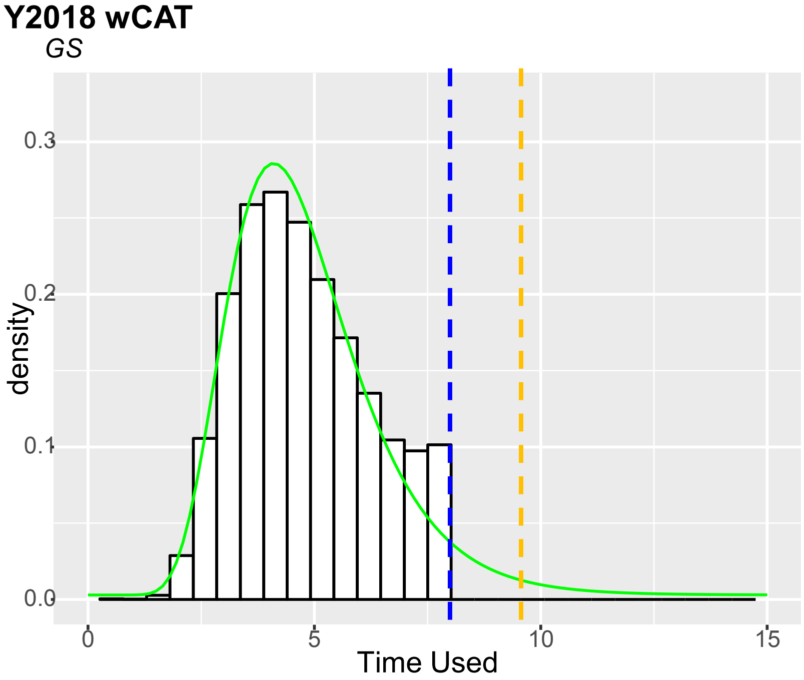
		Without Penalty		
		< 50	**>= 50**	**rowTotal**
With Penalty	**< 50**	20252 (55.8%)	879 (2.4%)	21131
	>= 50	0 (0%)	15163 (41.8%)	15163
	columnTotal	20252	16042	***36294***

ESTIMATING THE APPROPRIATE TIME LIMIT

- **Fit a theoretical statistical distribution to the empirical test response time distribution**
  - Response time (item or test) generally follows a log-normal distribution
  - If the test is speeded, then the empirical distribution will be right-censored at the time limit
- **Examine the 95th, 98th, and 99th quantiles of the fitted distribution to evaluate current and potential adjusted time limits**

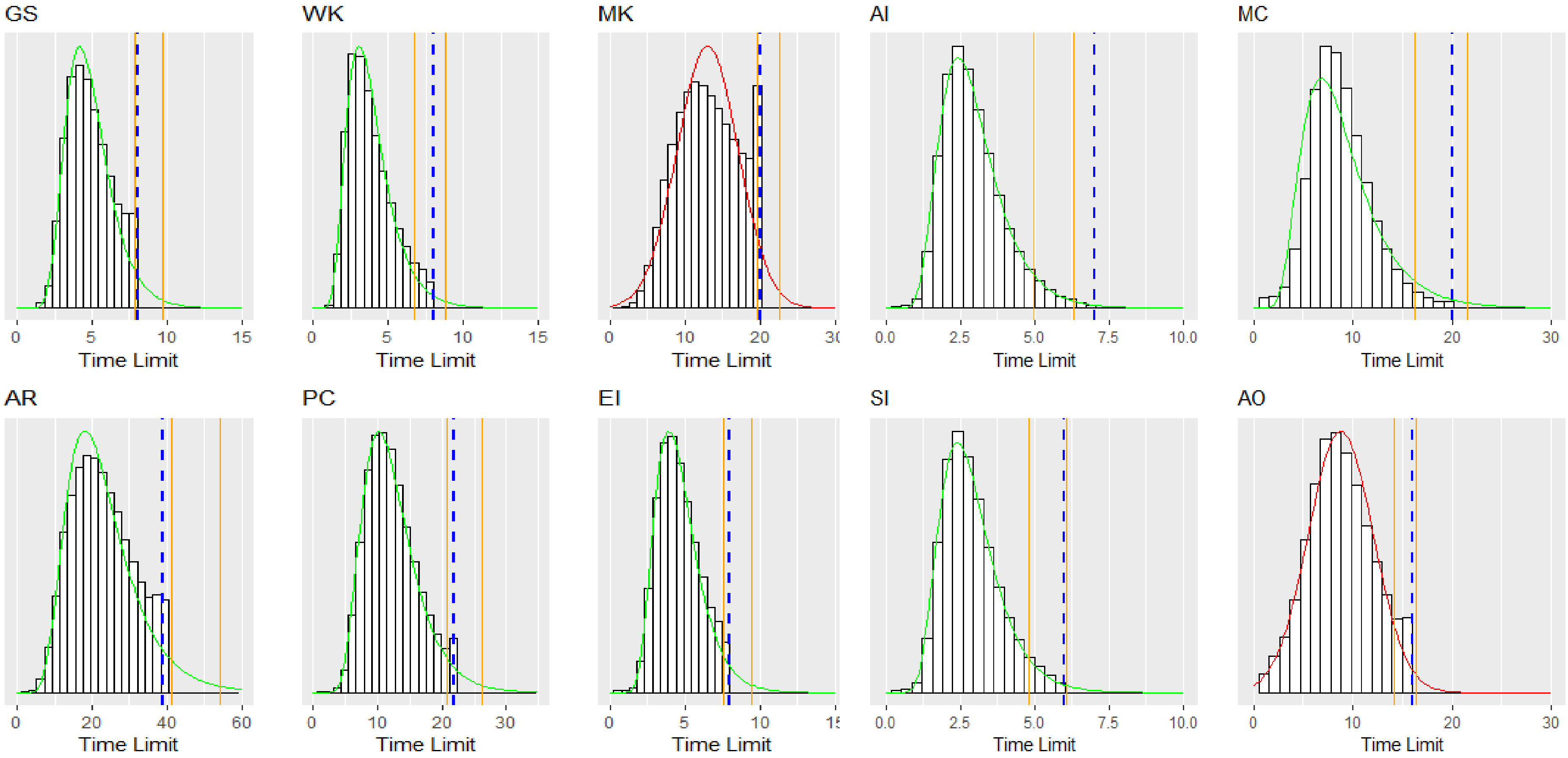
FITTING A LOG-NORMAL DISTRIBUTION TO CENSORED DATA—THE METHOD OF MAXIMUM LIKELIHOOD

- A positive random variable T, **if log(T) ~ N(μ, σ^2)**, the T is said to have a log-normal distribution **T ~ log-normal(μ, σ^2)**
- Using MLE to find the **(μ, σ)** values to fit a log-normal distribution to a censored data set with specified censoring point
  - R-package: maxLik (Henningsen & Toomet, 2011)
  - E.g: 2018 GS without seed items, examinees' response time distribution, censored at the allocated time limit: 8
    - Fitted log-normal distribution:
    - μ = 1.52, σ = 0.33
    - P(T > 8) = 0.042 (about 4.2% of the examinees would be expected to take longer than 8 minutes)
    - The 99th percentile = 9.8; a time limit =10 minutes would allow at least 99% of the examinees to complete the test within the limit.

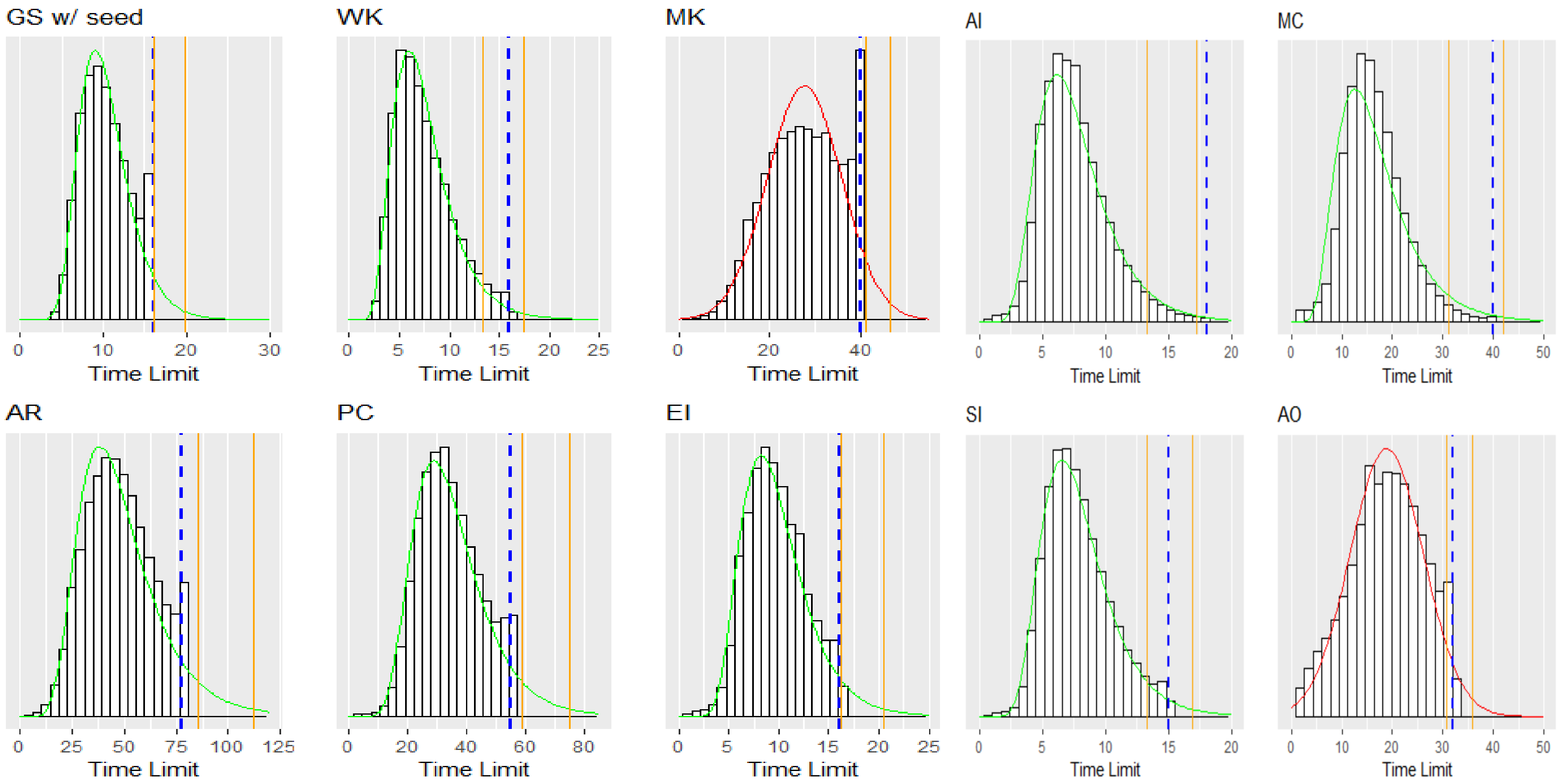


Henningsen, A. & Toomet, O. (2011). maxLik: A package for maximum likelihood estimation in R. *Computational Statistics* 26(3), 443-458.

FITTED DISTRIBUTION AND Q95, Q99—ALL DATA, WITHOUT SEED ITEMS



FITTED DISTRIBUTION AND Q95, Q99—ALL DATA, WITH SEED ITEMS



DESCRIPTIVE STATISTICS FOR FITTED AND OBSERVED DISTRIBUTIONS

- **Means, SDs of the fitted and observed response time distributions are mostly very close or identical**

	Without Seed Items					With Seed Items				
	Fitted Distribution			Observed		Fitted Distribution			Observed	
Test	Mode	Mean	SD	Mean	SD	Mode	Mean	SD	Mean	SD
GS	4.2	4.9	1.6	4.9	1.5	9.0	10.3	3.2	10.3	2.9
AR	18.0	23.0	9.7	22.8	8.1	38.3	48.5	19.9	47.9	15.9
WK	3.0	3.8	1.6	3.8	1.5	6.0	7.5	3.1	7.5	2.9
PC	10.1	12.2	4.5	12.2	4.0	28.7	34.8	12.8	34.4	10.4
MK	13.0	13.0	4.1	13.0	4.1	27.6	27.6	8.2	27.7	8.2
EI	3.9	4.6	1.6	4.6	1.4	8.2	9.8	3.5	9.7	2.9
AI	2.4	2.9	1.1	2.9	1.0	6.1	7.6	3.0	7.6	2.7
SI	2.4	2.9	1.0	2.9	1.0	6.6	7.9	2.9	7.9	2.6
MC	6.8	8.9	3.9	8.7	3.1	12.6	16.8	7.7	16.5	6.0
AO	8.8	8.8	3.3	8.8	3.3	18.9	18.9	7.3	18.9	7.3

FITTED DISTRIBUTION, Q95, Q98, AND Q99

- **From the fitted response time distributions on all the data combined**

	without seed items				with seed items			
	Current	Q95	Q98	Q99	Current	Q95	Q98	Q99
GS	8	7.8	8.9	9.7	16	16.2	18.3	19.8
AR	39	41.2	48.6	54.3	78	85.9	101.0	112.5
WK	8	6.7	7.9	8.8	16	13.3	15.6	17.4
PC	22	20.6	23.9	26.3	55	58.7	67.9	74.8
MK	20	19.7	21.4	22.5	40	41.1	44.5	46.7
EI	8	7.6	8.7	9.5	16	16.2	18.7	20.5
AI	7	5.0	5.7	6.3	18	13.3	15.5	17.2
SI	6	4.8	5.5	6.1	15	13.3	15.3	16.9
MC	20	16.2	19.2	21.6	40	31.2	37.3	42.0
AO	16	14.2	15.5	16.4	32	30.8	33.8	35.8

RECOMMENDATIONS

- **Set time limits so that at least 99% of examinees are able to complete the test in the time given**
  - Note that examinees do not need to take all the time allocated; once a subtest has been completed, an examinee may continue to the next subtest
- **This will yield completion rates similar to those observed when time limits were initially studied and established prior to implementation of new forms 5–9**
  - This means no equating is needed!
- **Continue to monitor the completion rates and total battery time once the new recommended time limits are implemented**

NEW TIME LIMIT RECOMMENDATIONS (in min.)

	without seed items					with seed items				
	Current	New	Change	Mean	Exp.Mean	Current	New	Change	Mean	Exp.Mean
GS	8	10	2	4.9	4.9	16	20	4	10.3	10.3
AR	39	55	16	22.8	23.0	78	113	35	47.9	48.5
WK	8	9	1	3.8	3.8	16	18	2	7.5	7.5
PC	22	27	5	12.2	12.2	55	75	20	34.4	34.8
MK	20	23	3	13.0	13.1	40	47	7	27.7	28.0
EI	8	10	2	4.6	4.6	16	21	5	9.7	9.8
AI	7	7	0	2.9	2.9	18	18	0	7.6	7.6
SI	6	6	0	2.9	2.9	15	17	2	7.9	7.9
MC	20	22	2	8.7	8.9	40	42	2	16.5	16.8
AO	16	17	1	8.8	9.0	32	36	4	18.9	19.6

- **For AR, the fitted response time distribution (with seed items) has a mean of 48.5 compared to its observed mean of 47.9, indicating that on average, the examinees' response time would increase by less than 1 minute after the time adjustment; for PC, on average, there would be no change.**

* Rounded up

OPA

OFFICE OF PEOPLE ANALYTICS

Backup Slides

FITTED DISTRIBUTION PARAMETERS—WITHOUT SEEDING

similar fitted distributions across year & platform

		μ					σ				
		2015	**2016**	**2017**	**2018**	**All Years**	**2015**	**2016**	**2017**	**2018**	**All Years**
GS	**WinCAT**	1.51	1.52	1.52	1.52	**1.52**	0.32	0.32	0.33	0.33	**0.32**
	iCAT	1.55	1.58	1.58	1.59	**1.57**	0.29	0.29	0.30	0.30	**0.30**
	All	**1.54**	**1.53**	**1.53**	**1.54**	**1.53**	**0.31**	**0.31**	**0.32**	**0.32**	**0.32**
AR	**WinCAT**	3.04	3.05	3.04	3.08	**3.05**	0.41	0.40	0.41	0.42	**0.41**
	iCAT	3.03	3.06	3.06	3.08	**3.05**	0.38	0.38	0.39	0.40	**0.39**
	All	**3.04**	**3.05**	**3.05**	**3.08**	**3.05**	**0.40**	**0.40**	**0.41**	**0.41**	**0.40**
WK	**WinCAT**	1.24	1.24	1.22	1.21	**1.23**	0.39	0.39	0.40	0.40	**0.39**
	iCAT	1.34	1.33	1.32	1.38	**1.35**	0.36	0.37	0.36	0.37	**0.37**
	All	**1.31**	**1.25**	**1.24**	**1.27**	**1.26**	**0.38**	**0.39**	**0.39**	**0.40**	**0.39**
PC	**WinCAT**	2.44	2.45	2.43	2.44	**2.44**	0.36	0.35	0.36	0.37	**0.36**
	iCAT	2.45	2.46	2.44	2.47	**2.45**	0.32	0.34	0.35	0.35	**0.34**
	All	**2.45**	**2.45**	**2.43**	**2.45**	**2.44**	**0.34**	**0.35**	**0.36**	**0.37**	**0.36**
MK	**WinCAT**	2.49	2.50	2.49	2.50	**2.49**	0.37	0.36	0.38	0.39	**0.38**
	iCAT	2.54	2.53	2.53	2.57	**2.54**	0.33	0.35	0.35	0.35	**0.34**
	All	**2.52**	**2.50**	**2.50**	**2.52**	**2.51**	**0.35**	**0.36**	**0.37**	**0.38**	**0.37**
MK: Normal	**WinCAT**	12.83	12.91	12.80	12.97	**12.88**	4.08	4.12	4.15	4.19	**4.14**
	iCAT	13.23	13.21	13.24	13.82	**13.40**	3.72	3.98	3.96	4.05	**3.91**
	All	**13.05**	**12.94**	**12.89**	**13.20**	**13.02**	**3.89**	**4.10**	**4.11**	**4.17**	**4.09**

FITTED DISTRIBUTION PARAMETERS—WITHOUT SEEDING

		μ					σ				
		2015	2016	2017	2018	All Years	2015	2016	2017	2018	All Years
EI	**WinCAT**	1.47	1.48	1.46	1.47	**1.47**	0.33	0.33	0.34	0.34	**0.34**
	iCAT	1.52	1.50	1.49	1.53	**1.50**	0.30	0.32	0.31	0.32	**0.31**
	All	**1.49**	**1.48**	**1.47**	**1.48**	**1.48**	**0.32**	**0.33**	**0.33**	**0.34**	**0.33**
AI	**WinCAT**	1.01	1.01	0.99	1.00	**1.00**	0.36	0.36	0.36	0.37	**0.36**
	iCAT	1.05	1.04	1.03	1.06	**1.05**	0.33	0.35	0.34	0.35	**0.34**
	All	**1.03**	**1.01**	**1.00**	**1.01**	**1.01**	**0.34**	**0.36**	**0.36**	**0.37**	**0.36**
SI	**WinCAT**	1.00	1.00	0.97	0.97	**0.98**	0.35	0.35	0.35	0.36	**0.36**
	iCAT	1.05	1.02	1.02	1.07	**1.04**	0.32	0.33	0.33	0.34	**0.33**
	All	**1.02**	**1.00**	**0.98**	**0.99**	**1.00**	**0.34**	**0.35**	**0.35**	**0.36**	**0.35**
MC	**WinCAT**	2.10	2.10	2.07	2.09	**2.09**	0.41	0.41	0.42	0.44	**0.42**
	iCAT	2.12	2.07	2.08	2.14	**2.11**	0.38	0.42	0.42	0.41	**0.40**
	All	**2.11**	**2.10**	**2.08**	**2.10**	**2.09**	**0.41**	**0.41**	**0.42**	**0.44**	**0.42**
AO	**WinCAT**	2.10	2.09	2.06	2.05	**2.07**	0.47	0.48	0.50	0.51	**0.49**
	iCAT	2.11	2.09	2.09	2.13	**2.10**	0.43	0.46	0.46	0.46	**0.45**
	All	**2.10**	**2.09**	**2.06**	**2.07**	**2.08**	**0.45**	**0.48**	**0.49**	**0.50**	**0.48**
AO:	**WinCAT**	8.88	8.86	8.63	8.62	**8.72**	3.29	3.32	3.33	3.34	**3.32**
Normal	**iCAT**	8.85	8.80	8.83	9.13	**8.87**	3.00	3.17	3.21	3.27	**3.13**
	All	**8.86**	**8.86**	**8.67**	**8.71**	**8.76**	**3.13**	**3.30**	**3.30**	**3.31**	**3.28**

Tab P

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

TAPAS Evaluation Project (TEP)

Status Update

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

March 29, 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 the TEP
- Summary and Highlights of First Two Meetings
- Upcoming Activities

Overview and Goals of the TEP

- Some stakeholders have raised technical concerns about TAPAS, especially low test-retest reliability
  - RAND recently completed an independent evaluation of the reliability and validity of TAPAS, finding:
    - Small, significant incremental validity over education credential in predicting attrition
    - Evidence of low test-retest correlation in some conditions
- DPAC requested a TAPAS Evaluation Project (TEP) to independently review the body of TAPAS research and make recommendations regarding the readiness of TAPAS for operational use. The evaluators will:
  - Review related research conducted by the Services, both on TAPAS and on other instruments (e.g., interest inventories)
  - Comment on the readiness of TAPAS for operational use
  - Make recommendations for future research and development
- Evaluators have expertise in psychometrics, personality theory and measurement, and operational testing programs
  - Dr. James Robert (Chair), Georgia Tech
  - Dr. Paul Sackett, University of Minnesota
  - Dr. Mark Reckase, Michigan State University
  - Dr. Winfred Arthur, Texas A&M University
  - Dr. April Zenisky, University of Massachusetts

Key Points

- PoP: Oct. 2018 – Oct. 2019
- Four TEP meetings
  - Attended by DPAC and Service representatives
  - Presentations from TAPAS developers, RAND, and Service representatives
- End result is a report on TAPAS readiness for operational use

Summary of First Meeting (October 22, 2018)

- Attendees from DPAC, ARI, USAF, USN, RAND, and Drasgow Consulting Group (DCG)

- Evaluators received draft of RAND report prior to meeting:
  - ***An Evaluation of the Tailored Adaptive Personality Assessment System: Is It Valid for Predicting Attrition from Military Service? Is It Reliable?***

- Agenda focused on TAPAS research and development
  - Drs. Stephen Stark, Fritz Drasgow, Sasha Chernyshenko, Chris Nye (DCG), Tonia Heffner, and Leonard White (ARI): ***Development of the Tailored Adaptive Personality Assessment System (TAPAS) and Ongoing Psychometric Research***
  - Dr. Chris Nye (DCG) and Dr. Leonard White (ARI): ***Validity Evidence for the Tailored Adaptive Personality Assessment System***
  - Mr. John Trent (AFPC/DSYX): ***Stability and Validity of TAPAS Under Operational and Experimental Conditions***

- Provided detailed minutes to evaluators

Highlights from Presentations

- Dr. Stark presented on the initial research and development of TAPAS:
  - Reviewed research leading to development of TAPAS
  - Described the personality and measurement theory underlying TAPAS
  - Summarized history of TAPAS development and use
  - Described research to improve reliability: use of marginal reliability index, recalibration of item pool, smart-CAT algorithm, use of triplet items
- Drs. Nye and White discussed ongoing TAPAS validation and research:
  - Described validity of TAPAS composites for predicting Will-Do (0.31*), Can-Do (0.25*), and Adaptation (0.12**) criteria
    - Incremental validity for Will-Do (over AFQT) and Adaptation (over AFQT and Educational Tiers)
    - Different predictors of 36-month and misconduct attrition
  - Evaluated validity for MOS-specific (Infantry, Military Police, Combat Medic, Transportation, and Mechanic) TAPAS composites— differences primarily found for Adaptation composites
  - Compared Soldiers' predicted performance in their current MOS to their predicted performance in other MOSs—40% to 47% of individuals were predicted to perform at least .5 SDs better in a different MOS
  - Examined use of TAPAS for in-service testing, such as to identify high-potential individuals for special duty assignments that are only available to experienced Soldiers (Special Forces, Recruiters, Drill Sergeants, Instructors)—multiple Rs = .18 - .48
- Mr. Trent presented research comparing TAPAS scores across three conditions: (1) operational, pre-accession, (2) retest administration post-accession under honest conditions, (3) retest administration post-accession under directed faking conditions:
  - Generally small to moderate effect sizes between conditions
  - The most predictive scales (i.e., Achievement, Cooperation, Even-Tempered, Non-Delinquency, Self-Control, and Selflessness) were significant across all conditions in predicting ethical decision-making performance.
  - Compared TAPAS FFM reliability to other FFM measures from two meta-analyses—TAPAS reliability somewhat lower than meta-analytic reliabilities, but potentially over different retest intervals
  - Physical Conditioning, Non-delinquency, Dominance, and Adjustment most strongly correlated with training and job outcomes across select Air Force careers
  - SMEs rated Adjustment, Achievement, Self-Control, and Even-Tempered as most important across three Air Force careers
- Evaluators also:
  - Reviewed and approved TEP charter
  - Elected Dr. Roberts as Chair
  - Identified topics for second meeting

*adjusted multiple R; ** Multiple R

Summary of Second Meeting (January 29, 2019)

- Attendees from OUSDPR, DPAC, ARI, USAF, USMC, RAND, DCG
- Evaluators requested articles, chapters, and technical reports recommended by the Services
  - ARI and USAF provided documents (see Appendix)
- Ms. Stephanie Miller and Mr. Matt Boehmer provided opening remarks
- Agenda focused on evaluation and operational use of TAPAS
  - Drs. Lawrence Hanser, Chaitra Hardison, & Denis Agniel (RAND): ***Evaluating the Usefulness of TAPAS: Reliability and Validity Results***
  - Dr. Tracy Kantrowitz (PDRI): ***TAPAS Research Review: Validity, Reliability, Demographic, and Faking Subgroup Differences***
  - Dr. Leonard White (ARI), Dr. Chris Nye (DCG), and Mr. Jeremiah McMillan (ARI): ***The Tailored Adaptive Personality Assessment System (TAPAS): Reliability and Validity***
  - Mr. John Trent (AFPC/DSYX): ***Use of the TAPAS in the U.S. Air Force***
  - MAJ Rachel Gonzales (USMC): ***USMC Use of TAPAS***
- Included individual Q&A sessions with RAND and ARI
- Provided detailed minutes to evaluators

Highlights from Presentations

- **Dr. Hardison presented on RAND's evaluation of TAPAS**
  - Found test-retest correlations ranged from .19 – .59
  - Reported low but significant validity for some TAPAS scores in predicting attrition, but little practical effect on reducing attrition
  - Discussed simulation showing observed levels of correlation may be due to large proportion misrepresenting or responding randomly
  - Encouraged both operational and lab-based research strategies to improve TAPAS
- **Dr. Kantrowitz discussed an ongoing TAPAS literature review**
  - Described project aimed at reviewing previous research findings and analyzing data to make recommendations regarding which aspects from the NCAPS and SDI would be worth incorporating into the TAPAS system
  - Coded 30 articles that had psychometric information on TAPAS
  - Presented pattern of predictive relationships between TAPAS facets performance and attrition measures; subgroup differences
- **Dr. White presented research in response to questions from the first TEP meeting**
  - Simulated data to examine impact of range restriction on test-retest correlations and retest score gains
    - When T1 score range is restricted, test-retest reliability is reduced and scores regress to the mean from T1 to T2
  - Used TAPAS to predict Can-Do, Will-Do, Adaptation, and Good Conduct criterion composites
    - Results varied across MOS, but were most predictive of Will-Do criteria
- **Mr. Trent described USAF's experience with TAPAS**
  - Reviewed current operational use of TAPAS in the Air Force (for some classification; not for selection)
  - Described testing policies and procedures, including retesting policy
  - Discussed positive (validity, faking resistance, flexibility) and negative (limitations on administration time) experiences with TAPAS
  - Described potential future of TAPAS in the Air Force, including adding classification models, modifying TAPAS format and content), and potentially using TAPAS in selection
- **MAJ Gonzales described USMC's experience with TAPAS**
  - Described current testing policies and use—administered to all recruits; scores automatically waived
  - Discussed the operational challenges USMC faces, such as small staff

Upcoming Activities

- Third Meeting (May 15, 2019)
  - Location TBD
  - Draft agenda
    - Operational policy and use presentations from Army and Navy
    - Additional discussion of reliability (RAND)
    - Summary of requested dimensionality analyses (DCG)
    - Working time for evaluators
- Fourth Meeting (Summer, 2019)
  - Working time for evaluators
- Final Report

TAPAS Articles, Chapters, and Technical Reports

ARI and USAF provided the following documents to the evaluators:

- *Constructing Fake-Resistant Personality Tests Using Item Response Theory* (Stark, Chernyshenko, & Drasgow, 2011)
- *Tier One Performance Screen Initial Operational Test and Evaluation: 2015–2016 Biennial Report* (Knapp & Kirkendall, 2018)
- *Adaptive Testing With Multidimensional Pairwise Preference Items: Improving the Efficiency of Personality and Other Noncognitive Assessments* (Stark, Chernyshenko, Drasgow, & White, 2012)
- *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to Support Army Selection and Classification Decisions* (Drasgow, Stark, Chernyshenko, Nye, & Hulin, 2012)
- *Moderators of the Tailored Adaptive Personality Assessment System Validity* (Stark, Chernyshenko, Nye, Drasgow, & White, 2017)
- *Assessing the Tailored Adaptive Personality Assessment System (TAPAS) as a MOS Qualification Instrument* (Nye, Drasgow, Chernyshenko, Stark, Kubisiak, White, & Jose, 2012)
- *An IRT Approach to Constructing and Scoring Pairwise Preference Items Involving Stimuli on Different Dimensions: The Multi-Unidimensional Pairwise-Preference Model* (Stark, Chernyshenko, & Drasgow, 2005)
- *Validation of the Noncommissioned Officer Special Assignment Battery* (Horgen, Nye, White, LaPort, Hoffman, Drasgow, Chernyshenko, Stark, & Conway, 2013)
- *Constructing Personality Scales Under the Assumptions of an Ideal Point Response Process: Toward Increasing the Flexibility of Personality Measures* (Cherynshenko, Stark, Drasgow, & Roberts, 2007)
- *Toward a New Attrition Screening Paradigm: Latest Army Advances* (White, Rumsey, Mullins, Nye, & LaPort, 2014)
- *From ABLE to TAPAS: A New Generation of Personality Tests to Support Military Selection and Classification Decisions* (Stark, Chernyshenko, Drasgow, Nye, White, Heffner, & Farmer, 2014)
- *Assessing the Tailored Adaptive Personality Assessment System for Army Special Operations Forces Personnel* (Nye, Beal, Drasgow, Dressel, White, & Stark, 2014)
- *Personality Assessment Questionnaire as a Pre-Accession Screen for Risk of Mental Disorders and Early Attrition in U. S. Army Recruits* (Niebuhr, Gubata, Oetting, Weber, Feng, & Cowan, 2013)
- *Examining Personality for the Selection and Classification of Soldiers: Validity and Differential Validity Across Jobs* (Nye, White, Drasgow, Prasad, Chernyshenko, & Stark, in press)
- *Tailored Adaptive Personality Assessment System (TAPAS) as an Indicator for Counterproductive Work Behavior: Comparing Validity in Applicant, Honest, and Directed Faking Conditions* (Trent, Barron, Rose, & Carretta)

Questions?

Tab Q

OPA
OFFICE OF PEOPLE ANALYTICS

Future Topics

Daniel O. Segall
Briefing presented at a meeting of the Defense Advisory Committee on Military Personnel Testing, 28-29 March 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 Panel
- 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
  - Expert Panel Recommendations
  - 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 Study

OFFICE OF PEOPLE ANALYTICS