DEFENSE ADVISORY COMMITTEE ON MILITARY PERSONNEL TESTING

September 20-21, 2018 Meeting

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

Minutes approved for public release.

December 27, 2018

Dr. Michael Rodriguez, Chair, DACMPT DATE

DEFENSE ADVISORY COMMITTEE
ON
MILITARY PERSONNEL TESTING

Minneapolis, MN
September 20-21, 2018

The meeting of the Defense Advisory Committee on Military Personnel Testing (DACMPT) was held at the Hyatt Place Downtown Minneapolis, MN on September 20-21, 2018. During the introductions, Dr. Paul Sackett noted that he had been a member of the committee 30 years prior.

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

1. Accession Policy Update (Tab D)

Mr. Christopher Arendt, Deputy Director, Accession Policy Directorate (AP), presented the briefing.

> Mr. Arendt began by summarizing the mission of the Accession Policy Directorate, which is to "develop, review, and analyze policies, resources, and plans for Services' enlisted recruiting and officer commissioning programs." He then presented an organizational chart detailing the structure and programs within Accession Policy. A table displayed recruiting outcomes for the Active Components as of August 2018. Regarding end-strength, the Army is somewhat below goal, with 97% of their 483,500 FY18 NDAA Authorized End Strength. The Army is also at 90% of their monthly mission attainment, 80% of their 6-month average contract mission, and 91% of their year-to-date mission attainment. An additional table showed figures for Reserve Component recruitment, also through August 2018. Once again, the Army is behind mission, with the Army Guard at 84% of monthly mission attainment and 78% of their year-to-date mission requirement. Corresponding figures for the Army Reserve are 68% and 73%. All other Services are meeting mission.

As Mr. Arendt briefed the recruiting metrics for 2018 (slides 4 and 5), a committee member asked him to clarify the figure provided for the Army's Delayed Entry Program (DEP) posture for FY 2019. Mr. Arendt explained that the Army had increased its recruiting mission in 2018, resulting in the depletion of persons in the DEP. He said the 5.0 number was the lowest it had been in some time and that AP was interested to see what the starting DEP figure would be on 1 October 2018.

As the briefing concluded, a committee member commented that the Army is using a larger percentage of their bonus dollars than the other Services. Mr. Arendt agreed and explained that the recruiting market is very tight, and that the Army has a different requirement than the other Services in terms of numbers.

2. Background – ASVAB Milestones and Project Matrix (Tab E)

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

Dr. Pommerich began the presentation with an overview of the projects to be covered in the briefing, including Armed Services Vocational Aptitude Battery (ASVAB) item pool development, the Career Exploration Program (CEP), ASVAB and Enlistment Testing Program (ETP) revision, the Internet-Based AFQT Predictor Test (APT), 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 – 14 from new items. New form implementation is projected for November 2019.
- Developing New CAT Item Pool for the CEP. The objective of this project is to build a CAT pool from 20B, 21 A&B, and 22 A&B for implementation of the *i*CAT in the CEP. The project is scheduled for completion in Fall 2018.
- Automated Generation of Word Knowledge (WK) Items. The objective of this effort is to develop procedures for automating WK item generation so WK item pools can be replaced on a more frequent basis. Anticipated completion date is September 2018.
- 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.
- 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.
    - 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 TAPAS; the WPA; and Army, Air Force, and Navy interest inventories. The project is ongoing.
- AFQT Predictor Test (APT). The objective of this project is to develop a short screening test that will accurately predict AFQT. The project was completed in Summer 2018.
- Air Force Compatibility Assessment (AFCA). The objective of this project is to program the AFCA for WinCAT administration. Project completion is slated for the Fall 2018.
- Defense Language Aptitude Battery (DLAB). The objective of this project is to transition to all computer-based testing and improve the predictive validity of the DLAB. The project completion is anticipated in December 2018.

- Expanding test availability through web-based delivery of special tests. The objective of this project is to transition the delivery of special tests from a Windows-based platform to a web-based platform. Project completion is scheduled for August 2021.
- Expanding test availability by moving to the cloud. The objective of this project is to examine the feasibility of moving test delivery to the cloud. Project completion is scheduled for August 2021.

As Dr. Pommerich briefed the ongoing development of new CAT-ASVAB item pools (slides 3-4), a committee member asked if this work included the development of items for the Word Knowledge (WK) test. Dr. Pommerich replied said that was a different effort being performed by Dr. Isaac Bejar under subcontract to the Human Resources Research Organization (HumRRO). Elaborating, she said HumRRO is currently identifying enemy items and evaluating the feasibility of automating that work.

When Dr. Pommerich said the new CAT item pool for the CEP was being developed based on retired paper-and-pencil (P&P) forms (slide 5), a committee member asked if that meant the items were not currently operational. Dr. Pommerich said that was a true statement.

As Dr. Pommerich addressed the technical bulletin requirements (slide 10), a committee member inquired about the delivery dates, size, and contents of the technical bulletins for Pool 10 of the CEP *i*CAT and the Armed Forces Qualification Test (AFQT) Predictor Test (APT). Dr. Pommerich said that the APT bulletin was almost complete, but that the Pool 10 bulletin might not be completed within the specified timeframe (i.e., Fall 2018). Dr. Segall confirmed that the bulletin would be delayed but stated that DPAC would expedite the priority due to the interest expressed by the committee. The committee member also asked if the bulletins would be available to the public, and Dr. Pommerich responded that they would be obtainable from the DPAC website. Dr. Pommerich warned the committee that the bulletins would be lengthy due to the inclusion of numerous tables.

As Dr. Pommerich concluded her briefing on the development of the Cyber Test (CT), a committee member asked if there was significant overlap in the AF's Joint Service CT and the Army's in-service version, or if the two tests were truly independent. Dr. Matt Trippe (HumRRO) replied that the Army's test was being developed to the same blueprint specifications that were used to develop the AF test. He explained that it might be helpful to think of the Army's in-service test as an alternate form of the AF's test. He added that item development for the Army's test has been completed and said the next step is to determine the relationship between scores and outcomes, which he said included attrition and training performance. Another committee member then asked for a more detailed description of the gaming approach, which was presented on (slide 19). Mr. Ken Schwartz (Air Force Enlistment Policy) explained that the proponents of the gaming approach believe that a different approach – one that incorporates mission-related performance versus knowledge – would be important in ascertaining candidate potential. He also mentioned that the gaming approach would be assessed by a different group of contractors. Dr. Velgach asked for the names of the contractors, and Mr. Schwartz said the approach was being developed by the Center for Applied Study of Language (CASL) and that the preliminary label for the assessment is the Cyber Aptitude and Talent Assessment (CATA) battery. He added that the literature review would help determine if the gaming approach should be used to enhance the existing tests or if it should be a separate measure.

Another committee member said s/he appreciated all the work being done in the CT arena, to include determining how the tests should be used. S/he also suggested that the constructs measured might differ based on the approach (i.e., knowledge vs performance) and said it would be interesting to see how the various tests might be used in combination or separately to meet the needs of the individual Services. Dr. Pommerich said DPAC would be covering the continued development and validation of the tests later in the current meeting as well as in future meetings. A committee member then asked if a test that was targeted for a specific Service could be integrated into the ASVAB platform. Dr. Segall replied that this question was more complicated than it seemed. He explained that the platform is used to administer special tests, even if only one Service uses them. He said the infrastructure is there and that placing the test on the platform allows a test to be shared (adopted by other Services), if desired. He said the AF's CT is a good example of this, but that it has been years since a test was formally added to the ASVAB, as opposed to just being administered on the platform. Dr. Pommerich then commented briefly on the complexity of adding a test to the ASVAB.

When Dr. Pommerich commented that the Mental Counters test (MCt) was considered more of a test of working memory than of nonverbal reasoning, a committee member commented that s/he had been looking at the 1918 Army beta and trying to figure out the types of tests that were administered at that time. Dr. Segall responded that he would like to see a copy of the 1918 beta if it was available. Dr. Pommerich added that DPAC has a copy, but that it was not complete. A committee member replied that the American Psychological Association (APA) has a site that provides previously copyrighted materials that are now in the public domain and that it includes some original forms of the Army beta. S/he clarified that the site has a cover page that says the APA "believes" the tests on the site are in the public domain. Dr. Pommerich commented that the problem with releasing a test in that manner is that, once it is released, it is out there forever.

As Dr. Pommerich briefed the addition of non-cognitive measures for selection and classification (slides 22-25), a committee member asked for more information about the Dark Tetrad facet items. Dr. Mark Rose replied that these covered narcissism, Machiavellianism, psychopathy, and sadism. He said that Machiavellianism is currently of special interest and that it was a popular topic at the recent Society for Industrial and Organizational Psychology (SIOP) conference. He added that some research suggests that "Dark" characteristics are not as bad as originally thought. Dr. Segall interjected, however, that the military is not selecting a lot of candidates with those characteristics. The committee member responded that the associated measurement dimensions could be interesting. LTC Rea (U.S. Army) then mentioned the GRIT measure developed by Dr. Angela Duckworth at the University of Pennsylvania and administered at West Point. He said that he also used the GRIT when he was a battalion commander to instill traits important for special operation forces (SOF). He said that, at West Point, the GRIT was the number one predictor of attrition, even more so than GPA. He then asked if the current TAPAS covered the facets addressed by the GRIT. Dr. Pommerich replied that she would defer to the Service representatives. Dr. Cristina Kirkendall responded first, saying that she would have to look more closely to see if the TAPAS scales used by the Army addressed the GRIT facets. Other Service representatives identified perseverance, resilience, and adaptation as hot topics. LTC Rea replied that he would like to discuss the matter further off-line, and the other Service representatives agreed. The discussion ended with a committee member mentioning that the acronym, AVID, the Army's interest inventory, is also the name of a common career and readiness program used in colleges. Dr. Pommerich said she had also seen the acronym used in that context.

On the topic of the APT (slide 26), a committee member said that the test showed promise, though the committee previously expressed concerns regarding its use by recruiters. The committee member said s/he would be interested in hearing an update in the future. Another committee member then recalled that part of the problem with the use of the APT had been the popularity of a competing measure, the Enlistment Screening Test (EST). Dr. Pommerich said that the Navy developed the EST approximately 30 years ago and that other Services had adopted it. She said that HumRRO had previously developed the Computerized Adaptive Screening Test (CAST) to improve on the EST, and that the APT was being developed as an updated, web-delivered version of CAST. She went on to say that the EST still appeared to be more popular than the APT among recruiters, which was causing an issue with APT implementation.

A committee member then noted that the EST is a P&P instrument and asked if it was scored by hand. Mr. Arendt replied that the EST is completed on a hand-scored bubble sheet. He said its popularity was due, in part, to its ease of administration. The committee member asked if DPAC knew if the EST was predictive. Mr. Arendt replied that it is not, but the driving factor behind its use is that recruiters perceive it to be more predictive than the CAST, which has led to a lack of buy-in on the CAST and APT. He clarified that the Navy Recruiting Command had looked at the data on the EST and APT and found that the tests were about equally predictive. He said that, after the recent scoring update, the validity of the APT should be reassessed. Mr. Arendt then said that he does not have any data on EST scores, because the EST is scored at recruiting stations. Dr. Segall then interjected that, if the APT is not more efficient than the EST, then DPAC has done something terribly wrong, simply because it is adaptive. A committee member asked whether the presentation of definitive data would be enough to convince recruiters that the APT is a more effective test. Another committee membered pointed out that the fact that the APT items are more current than the EST items should be enough to persuade recruiters of its value. The committee member then asked Dr. Pommerich about the length of the unproctored Pre-Screening Internet CAT-ASVAB (P*i*CAT). Dr. Pommerich replied that the P*i*CAT is the length of the full ASVAB, but that people first take the first five subtests and only take the technical tests if their initial scores are high enough. She added that the P*i*CAT AFQT scores are based on 55 items and that APT scores are based on 20 items. She said that DPAC could look at lengthening the APT to increase prediction accuracy, if necessary.

As Dr. Pommerich briefed the Air Force Compatibility Assessment (AFCA), Mr. Schwartz commented that the assessment is called the “compatibility assessment” because it identifies individuals who are likely to demonstrate inappropriate workplace behaviors. He said it is used to prevent the accession of outliers who come into a unit and then have to be removed. A committee member asked about the format of the items, and Dr. Rose replied that there are multiple item types, including Likert. The committee member then suggested the test is, perhaps, too transparent, to which Dr. Rose replied that test-takers still appear to answer the items accurately. Dr. Pommerich asked if the same statements were used repeatedly as a check, and Dr. Rose said that the items include similar statements within scales, but that the items are not the same. Dr. Pommerich then commented that the content and scoring provide checks and balances. The committee member asked Dr. Pommerich if DPAC was still eliminating WinCAT, to which Dr. Pommerich replied, “of course.”

At the end of the briefing, a committee member asked if DPAC would talk briefly about the RAND report on the reliability of the TAPAS. Dr. Pommerich replied that Dr. Tim McGonigle (HumRRO) would be talking about an expert panel that is being established to look at the issues raised by that report. She explained that the panel was being formed in response to RAND's independent review of the TAPAS, which found low test-retest reliability. Mr. Arendt added that RAND may be doing a follow-on study with an expanded data set, and that RAND would be working with ARI in that effort. He said he was not sure of the progress to date. Another committee member recalled that the Services had previously provided information on TAPAS test-retest reliability to the DACMPT. Dr. Pommerich replied that the information from the Services, as well as from RAND, can be brought to the attention of the expert panel, which she said would be an 18-month effort. CDR Hank Phillips (Navy) asked if the RAND report was available, and Dr. Pommerich replied that the report has not been officially released, even to DPAC.

LTC Rea (Army) asked for an update on the use of calculators on the ASVAB. Dr. Pommerich said that DPAC had provided an information paper on that subject, which she said explained why using calculators had been discouraged. Dr. Arendt said he would be glad to provide that paper again to the Army and commented that the rationale had been consistent over time. LTC Rea said he just wanted to know if there had been updates, and Dr. Arendt said that the Office of People Analytics (OPA) continues to evaluate the matter. Dr. Pommerich explained further that it is a time consuming and costly process to conduct the research and standardization that would be required by the introduction of calculators. She said that adding calculators would change the items and constructs measured, require the development of new items, and require an entire line of research to validate the updated tests against success in training. She added that the resulting test would be viewed as a new test to load on the platform and said that DPAC had outlined all the steps that would be required, which would include a new norming study costing around 20 million dollars.

Mr. Arendt reinforced Dr. Pommerich's assertions, citing requirements for norming on all existing and new pools; he said the work would represent a seismic shift in the tests. He also said the Manpower Accession Policy Working Group (MAPWG), when briefed on the matter, had been in agreement. Dr. Segall then said that interest in using calculators had originally surfaced in the recruiter community, which knew that other standardized tests had begun to allow the use of calculators. He added that, if the military began to allow their use, scores would rise but the number of "qualified" applicants would not increase, due to equating. He said the report Dr. Pommerich had cited was the result of their investigation and that DPAC had concluded that integrating the use of calculators would require the same amount of work as adding an entirely new test to the battery. He went on to say that there are more promising tests that are on the docket. Dr. Velgach clarified that, even if the use of calculators were to lead to higher scores, the requisite norming would prevent an increase the number of qualified candidates. Dr. Segall agreed. Dr. Salyer then identified an additional complication in relation to the CEP: allowing calculators would be problematic in that the types of calculators used across high schools would have to be standardized. She said this would destroy the CEP program. Mr. Arendt then summarized the discussion by saying that, though it sounds simple to allow calculators to be used, it is very complicated and would not qualify more applicants. Dr. Pommerich said that the military would be 50 million dollars poorer for having undergone the effort, and that the only benefit would be an increase in the face validity of the ASVAB.

3. Next Generation ASVAB and ETP Update – (Tab F)

Dr. Mary Pommerich, DPAC, presented the briefing.

Dr. Pommerich began by providing an overview of her presentation, the goal of which is to give background and update on the history, status, and plans for the next generation of ASVAB and special tests administered in the Enlistment Testing Program (ETP). An expert panel was convened in 2005-2006 to consider the status of the ETP and make recommendations for improvements and enhancements. The panel (a) reviewed the ASVAB content, methodology, and use; (b) discussed problems associated with the current battery; (c) reviewed new types of cognitive and non-cognitive skills not measured by ASVAB that might prove valid for selection and classification, and; (c) developed recommendations for change to the battery. Dr. Pommerich then presented a slide showing 17 panel recommendations with associated ranks assigned by the Military Accession Policy Working Group (MAPWG). Dr. Pommerich then turned to new tests of interests.

- Extensive research has been conducted by the Services on the usefulness of the Tailored Adaptive Personality Assessment System (TAPAS) as a screening instrument for military applicants. However, some concerns have been raised about the seemingly low stability of test-retest scores over time after an independent review conducted by RAND. Given the outcomes of that review, an expert panel has been assembled to review TAPAS.
- Extensive research has also been conducted by the Services on the usefulness of the Cyber Test (formerly known as the Information Technology and Literacy Test) as a screening instrument. Concerns have been raised about the vulnerability of the test to compromise with the 29-item, fixed form currently in use, as well as the fact that it may be too difficult for the applicant population given the lack of moderately difficult items. The feasibility of a CAT version of the test will be revisited after further item development has taken place.
- Mental Counters (MCt) is a working memory test currently administered to Navy applicants on the CAT-ASVAB platform. The test has very promising characteristics, including high reliability and short testing times. However, there is a persistent floor effect, with approximately 4-9% of examinees receiving a score of zero each year. Options for eliminating the floor effect are being considered.
- The Abstract Reasoning Test (ART) is a test of nonverbal reasoning that is currently being administered to Defense Language Institute (DLI) applicants as part of the Defense Language Aptitude Battery (DLAB2). A planned research study of nonverbal reasoning tests has been deferred indefinitely. Analyses of ART and MCt data could yield more information about the desirability of investigating ART further. (MCt is also administered as part of DLAB2.)

Dr. Pommerich then turned to the question of why no modifications to the ASVAB have been undertaken. In June 2011, the DAC encouraged DPAC to determine the "uses that each Service requires the ASVAB to meet, in order to establish a philosophy of the test." Since that time, the MAPWG has had numerous discussions about potential modifications to the battery, but these have been stymied by several key issues. First, there are unresolved questions about what the philosophy of the ASVAB should be. There are also concerns that there are insufficient resources to accommodate a revised ASVAB that takes more time than the current battery. Finally, there are the logistical difficulties associated with making changes that would impact existing composites and systems set up to operate on those composites. DPAC is now hopeful that application of an arguments-based approach to validation of the ASVAB will help answer the question of what the philosophy of the ASVAB should be.

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.

Dr. Pommerich then turned to a discussion of next step in these efforts. These include:

- Continuing efforts to evaluate and resolve issues/concerns pertaining to the new tests of interest (TAPAS, Cyber Test, Mental Counters, Abstract Reasoning).
- Continuing efforts to evaluate tests currently in the ASVAB.
- Completing the effort to apply argument-based approach to validation of the ASVAB.
- Developing a shared vision among stakeholders that defines the purpose and general makeup of the next generation ASVAB.
  - Revisit the question of the philosophy of the ASVAB as needed, following establishment of a validity framework.
- Establishing a systematic process to follow for evaluating potential changes and making decisions regarding tests in the ASVAB.
  - Recommended by the DAC following the last revision to the battery.
  - DPAC presented a proposed process for potential changes to the ASVAB in 2014.
- Reviewing and updating the psychometric checklist, as needed, for the purpose of evaluating tests to be administered as part of the ASVAB.
  - Current checklist was developed for making decisions about adding tests to the ASVAB platform, not the battery.
- Having Services/proponents complete the updated psychometric checklist for new tests of interest, documenting all new information since the last checklist was completed.
- Revisiting logistical questions with stakeholders, including the feasibility of lengthening the ASVAB and the feasibility of dropping existing tests.
- Having stakeholders summarize the impact of potential modifications to the battery and identify resources to support a revised battery.
- Compiling all information, then identifying and discussing potential changes to the contents of the ASVAB and tests administered in the ETP.
  - Given the complexities associated with making changes to the battery, DPAC believes it is best to consider all new and existing tests at once, rather than on a case-by-case basis.

As Dr. Pommerich reviewed the status on each of the expert panel's recommendations, Mr. Brad Tiegs (MEPCOM) asked about implementing content controls. Dr. Pommerich said that they had incorporated pseudo controls, such as the tracking of enemy items and not allowing more than one item of a certain type, but she said that these did not technically fall under the definition of content controls.

On the topic of the MCt (slide 6), a committee member asked if DPAC still believed that the floor effect resulted primarily from confusion over the instructions. Dr. Pommerich said that likely accounted for some of the cases and explained that Dr. Ping Yin (HumRRO) would brief some options for better explaining the task to test-takers. Another committee member asked if similar issues were faced with the Abstract Reasoning Test (ART). Dr. Pommerich said DPAC would be evaluating the ART and that the committee might have seen the prototype in a previous briefing. Another committee member then asked for an explanation of why there have not been any recent additions to the ASVAB. Dr. Pommerich said DPAC had experienced stumbling

blocks in that area, with questions about the overall philosophy of the ASVAB being the primary cause of the delay. She said that the Services were not able to agree on what tests to add and remove, and that the discussion became contentious. She said that a current effort, led by Dr. Art Thacker (HumRRO), is looking at the path forward for the ASVAB philosophy through the lens of Kane's work in validity arguments.

Later in the briefing, as Dr. Pommerich addressed "next steps," the committee member pointed to DPAC's 2014 presentation of a proposed process for modifying the ASVAB and commented that the more systematic the process, the better it will be. Another committee member commented that it is helpful to know where DPAC stands, what options are available, and how to make those decisions. S/he said that the validity framework will help, but that it also takes a vision of the future. The first committee member then mentioned the utility of a needs assessment, and Mr. Arendt replied that such an assessment will be incorporated into the program. He said that many factions within the Department of Defense (DoD) are pushing for testing space and that the test-time constraint is the main enemy of any effort to move forward. Another committee member asked if there was a single entity that had ownership over the testing space. Mr. Arendt identified AP as the owner, but also emphasized the advisory role played by the DACMPT. The committee member replied that attempting to make such decisions by consensus would be problematic, because the Services realize that the tests selected for administration affect people's lives in a profound way. S/he added that strong leadership would be required to say, effectively, "get in line or get out."

In response, Dr. Segall clarified that the current process of dropping and adding tests, which he described as an evolutionary process, should probably be replaced by an approach that makes all the required changes concurrently, and that this be done periodically. The committee member reiterated that any decisions would need to be made considering expectations for future changes. Dr. Pommerich said that, fortunately, the MAPWG, AP, and DACMPT can provide information about what the Services need, but she explained that the Military Entrance Processing Command (MEPCOM) also has constraints on the administrative side. She thanked the committee member for making the point and said it was going to be a difficult task to meet everyone's needs and expectations. She went on to note that an emphasis on cyber testing was currently of great interest, but that not long-ago interest in a TAPAS-like test was very high; she mentioned this to emphasize that cyclical nature of events. She also said DPAC was fortunate to be able to administer new tests on the platform without adding them to the ASVAB battery. She then noted that pressure will exist to administer tests that make recruiting easier.

A committee member asked if there might be a way to administering tests other than putting them on the ASVAB platform. Dr. Pommerich replied that it sounded like the committee member was referring to unproctored tests administered from some alternate platform. Mr. Arendt responded that the solution is P*i*CAT. He said that the PiCAT currently accounts for about 40% of ASVAB administrations but needs to be increased to around 70%.

Returning to the difficulty of adding and dropping tests, Dr. Segall said that it is relatively easy to add tests, at least in comparison to dropping tests, which he said they have not been able to do. A committee member then commented that the Services like to have options, and that a test should be available if it is shown to be valid for their Service. S/he then said that it should be possible to

count battery time separately from time required for Service-specific tests. Mr. Arendt responded that each Service has a composite score for the ASVAB tests, and that dropping a test off the ASVAB would cause a change in the composite score for that Service. Another committee member asked if a test could be moved to a special section of the platform. Dr. Segall said that they have implemented that approach, with Coding Speed (now a Special Test) being an example, and that it could be done it again. Mr. Arendt clarified that every Navy applicant takes the Coding Speed test, and that it takes additional time on the back end of testing (i.e., after ASVAB administration). Another committee member suggested that the approach used with the Coding Speed test is probably the best approach moving forward. Mr. Arendt agreed.

CDR Phillips said he could envision an expansion of examination time once the PiCAT is at or close to full utilization. He asked, however, if applicants would be sitting for seven or eight hours to take numerous special tests. Dr. Pommerich said that MEPCOM would veto a scenario that required that amount of continuous testing time. Mr. Arendt said it would depend on the utilization of web delivery and that if the tests were unproctored, it could be an option. He added, however, that applicants might conclude that taking "the next test" is not worth it, especially if they have just completed an extensive exam.

A committee member raised the point that a lack of change in exam scores indicates that human ability has not changed in the last seven or eight years. Dr. Pommerich agreed and said that only four changes have been made to the ASVAB since 1968. Dr. Segall clarified that scores on the ASVAB AFQT, which he called a measure of *g*, have not changed much over time. He said that the *g* components of the battery are probably working the best. The committee member responded that, if criterion data on each of the ASVAB tests were available, there would be some basis on which to make judgments about which tests could be removed. Dr. Pommerich agreed, saying that DPAC has those data. She also indicated that proneness to compromise and other data could be useful. Dr. Velgach then asked if there was a consolidated database of criterion data, that is, performance-based outcomes. Dr. Pommerich replied that Dr. Janet Held (a former DPAC researcher) had been working to establish such a database, but that the effort fizzled. Dr. Segall said that there has been progress within the individual Services in this area. He said that the Army has a system and some of the other Services may have something as well. Dr. Kirkendall responded by clarifying that the Army has a database of some criterion data, but she said that it does not include everything; she explained that creating a more comprehensive database is a very complex endeavor. Mr. Arendt added that this problem is the hardest nut to crack. Another committee member then noted that s/he thought there had been an effort to consolidate criterion data across Services some years ago.

A committee member commented that the overall effort to introduce changes to the Enlistment Testing Program (ETP) constitutes a level of institutional change that s/he has never seen before, and that it will not be easy to achieve. S/he said that the previous panel wanted to see the desired changes take effect in the lifetime of its members; s/he then noted that it has now been 12 years since that panel made its recommendations. The committee member then admitted that DPAC has, indeed, developed tests that were suggested by that panel, and Dr. Segall said that the CT was an example, but that it is not in the ASVAB battery. The committee chair closed the discussion, saying the conversation could continue during the next briefing, an update on the ASVAB validity framework.

4. <u>Validity Framework Update</u> (Tab G)

Dr. Art Thacker, HumRRO, presented the briefing.

Dr. Thacker began by providing an overview of the presentation, which includes a summary of the validity argument approach to validation, a discussion of how this can be applied to validation of the AFQT and ASVAB, drafts of the theory of action (TOA) and claims structure for the AFQT and ASVAB, some specific validity evidence, and next steps and challenges.

The validity of an assessment depends on the purpose of that assessment, the inferences made from scores, and the use of those scores. Argument-based validation tests the underlying claims that must be true to support the inferences made. An assessment score may be valid for multiple purposes, in which case it is rare that the assessment is equally valid for all of them. Evidence is collected for a validity argument to support claims in a chain of reasoning, where any claim in the chain found to be weak may undermine subsequent claims. If multiple inferences are drawn from a single assessment score, each inference may have its own unique validity argument.

The process starts by identifying the most important inferences to be made from an assessment. In the case of the ASVAB, these include admission to the military, placement into training programs or advanced educational opportunities, and prioritization of recruiting efforts. The ASVAB primarily relies on an informal reasoned approach, where evidence is not typically tied to organized claims or assumptions. A TOA is required to frame the interpretive and validity arguments. Addressing the validity of the ASVAB for selecting individuals into specific training programs would require its own body of evidence and is beyond the scope of this effort.

Dr. Thacker continued by providing an illustration of the validity argument. The TOA identifies all the things the test and test scores are expected to be used for, and the expected advantages of using the test for those purposes. The interpretive argument provides a description of the inferences that the test scores support. The validity argument consists of evidence providing justification for the inferences and the interpretive argument. The theory of action for the AFQT could take the form of assuming that the tests that make up the AFQT measure *G*, and because *G* is predictive of a broad range of future performance, the AFQT will broadly predict success in military occupations. We can then develop claims that must be supported for each step in the TOA to be true (i.e., AFQT measures *G*, *G* is broadly predictive of performance, candidates sorted based on AFQT scores are sorted according to likelihood of success in military occupations).

Dr. Thacker continued by presenting two draft TOAs for the ASVAB. The first of these starts with the claim that the ASVAB subtests measure many KSAs, and job analysis identifies the KSAs required to perform specific jobs. Experts match the job/training KSAs to ASVAB content, evidence demonstrates that ASVAB subtests predict job/training performance and who will succeed in training and on the job. This model relies on clear linkages between KSAs required for military training and jobs and KSAs measured by the ASVAB. The second TOA begins with the claim that ASVAB subtests measure content students were taught in high school, and therefore the ASVAB is a strong measure of achievement in high school subjects. High school achievement is then demonstrated to predict military job/training performance. This model relies on prior education success being predictive of future success in the military. However, this is not the way that the Services conceptualize the ASVAB, so this model is not being pursued.

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

- AFQT Measures G
    1. A candidate's score on the AFQT is an estimate of that candidates true G.
    2. The predictive nature of G is continuous for nearly the full scale for the AFQT (i.e. a higher score always results in a better prediction of outcome, irrespective of the area of the scale the score falls in).
    3. The AFQT categories represent important differentiators among candidates.
    4. AFQT scores have high overall reliability, especially near the cut points for the categories.

5. AFQT scores have high classification accuracy.
6. AFQT scores are largely free from construct irrelevant variance.

- *G* is Broadly Predictive of Performance
  7. Other *G* measures are used to predict performance broadly in non-military contexts, and these measures correlate positively and strongly with AFQT.
- Candidates Categorized Based on AFQT Are Sorted According to Likelihood of Success in Military Occupations
  8. AFQT scores correlate positively and strongly with success in military careers.
  9. AFQT scores are unbiased with regard to race/ethnicity, gender, etc.

The TOA may also include some uses of the AFQT scores. These may not fall under the heading of inference but are vital to the success of the assessment. Examples might include the following.

- Utility and Implementation Factors
  10. Users of the AFQT scores can interpret and understand score reports.
  11. Users of the AFQT scores are sufficiently trained to help candidates understand their options based on the AFQT performance.
  12. Factors outside of AFQT scores that contribute to the decision to enlist a candidate enhance predictions based on AFQT alone.

Dr. Thacker continued by presenting an excerpt from a draft validity argument for the AFQT, which included an assumption, the claims based on that assumption, and evidence supporting those claims. He continued by identifying the next steps to be followed in developing the validity argument, which include:

1. Revising the TOAs to better reflect the logic model underlying ASVAB (in general) and AFQT (specifically) (Iterative).
2. Defining/revising the assumptions associated with the revised TOAs.
3. Developing /revising the specific claims that support the assumptions.
4. Identifying the required evidence necessary to support validity claims.
5. Referencing evidence for specific validity claims from the literature and from ASVAB documentation (e.g. technical manuals).
6. Identifying evidence gaps or weaknesses and commission analyses/studies to address them.
7. Maintaining and updating the validity argument as necessary.

Dr. Thacker concluded by highlighting some of the challenges that will be faced in developing a validity argument for the ASVAB, including the lack of models for comparable assessment systems, the long history of the ASVAB and the fact that it has multiple users and uses, and discerning which ASVAB literature is relevant for the validity argument, particularly when not all the literature is unbiased.

As Dr. Thacker described the validity argument, as illustrated on slide 5, a committee member asked if a theory of action (TOA) could include a "use" that does not involve an interpretive argument. In response, Dr. Thacker replied that a use of that type would probably be limited to a more distal purpose, or a secondary effect, such as causing a change that is not directly or entirely dependent on test scores. Along that line of thinking, a committee member later suggested that military readiness is a less direct effect of military testing. The committee member clarified that this "readiness" was not preparedness for training, but the force readiness required for mission success. Mr. Arendt clarified that force readiness is obtained largely through training outcomes, which are improved through selection and classification, making testing a step in the direction of readiness. Dr. Segall replied, however, that there are specific metrics for quantifying force readiness, such as unit strength. Mr. Arendt clarified that readiness at the unit level is achieved when trained assets arrive, and that a decrement exists when an untrained person fills a position. The committee member then suggested that the TOA could include a more distal outcome, such as force readiness.

Another committee member asked if the specialty tests, as well as the AFQT (i.e., a measure of *g*) would fall under the proposed TOA. Dr. Thacker said that all the tests, as well as composite scores, fall under the model. Mr. Arendt clarified that the AFQT is used as the primary indicator of quality and that composite scores are used for classification. He said the composites include AFQT scores for many occupations.

At this point, Dr. Thacker explained that the major advantage of adopting the recommended type of model is that it supports the measurement of the knowledge, skills, and abilities (KSAs) recruits need to possess, but that the disadvantage is that those KSAs change over time, requiring continual examination of job requirements and training. He noted that this can be a challenge. A committee member resurfaced the matter of the next generation ASVAB, raising the need to focus on success in current as well as future jobs and KSAs.

As Dr. Thacker began to brief the AFQT claims shown on slides 9 and 10, Dr. Segall said that an AFQT-like battery could legitimately include typical measures of *g* as well as more job-specific measures. He noted that, historically, DPAC has taken this approach, citing the inclusion of math and verbal tests. He said that their approach has been to look at current textbooks to ensure the tests are tapping into what high school students are being exposed to. Mr. Tiegs responded, suggesting that not just any measure of *g* would be effective, but that the measure should be related to the types of actions and knowledge that are required in job training. He said that he can think of ways of measuring *g* that miss the KSAs that are relevant for job training. He also said that the early legal cases for job selection resulted in the requirement for a connection between measures of *g* and job requirements. In response, a committee member pointed out that some people say *g* predicts everything and that any measure that is more job-specific would correlate highly with other measures of *g*. Another committee member then noted the possibility of increased adverse impact if a very general measure of *g* were used. The first committee member agreed but said that selecting for a vast universe of jobs almost requires some generalized type of measure. S/he also said that the debate about the use of measures of *g* rages on and that legal cases are sometimes decided in light of political considerations.

A committee member asked Dr. Thacker if the claims shown on slides 8 and 9 were going to be investigated. When Dr. Thacker replied that his team is still searching the literature, another committee member asked if "commissioned studies," technical reports not published in journals, were candidates. Dr. Thacker replied that they were.

The discussion concluded with a brief discussion, clarifying that TOA 2 (slide 8) was not on the table. Dr. Thacker reported that TOA 2 was not well received at the recent MAPWG meeting. Dr. Segall, however, said that he would still like to keep TOA 2 in the background to potentially revisit in the future.

5. <u>Mental Counters</u> (Tab H)

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 a small practice effect; and (d) is an excellent candidate for automatic item generation. 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 fairly good distributions except for a floor effect where nearly 9% of test takers 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. She explained that there was hope that the floor effect would be reduced over time, but instead it has gotten worse, leading to consideration of how to eliminate it. Option 1 is to change the test by ordering items from easiest to hardest and add easier items to the test. The rationale is that the current administration is unbalanced, which could be alleviated by adding easier items and increasing the time between counter adjustments. The positives to this option are that it could eliminate the floor effect and balance the administration design. The negatives are that it would change the test, could introduce a possible ceiling effect, and could make it more difficult to compare performance on the test from previous administrations without an equating study. Option 2 is to make the instructions clearer, add an animated demo of the task, and require that examinees answer at least one practice item correctly in order to start the test. This would very likely reduce the floor effect, given that minor clarifications of the instructions between versions 2.0 and 3.0 reduced it by half. It would also not require an equating study. The down side to this option is that it would introduce practical constraints involving updating instructions, creating the demonstration and practice items, and developing new sequencing, as well as minor adjustments during the operational test.

Dr. Yin went on to explain that currently there are at least 50 screens for instruction, demonstration, and practice. Recommendations for improvement include reducing the number of screens, providing a video-like visual demonstration to clarify the instructions, and directing the test administrator (TA) to ensure that candidates understand the instructions before beginning the test. A pilot test of the new instructions could be conducted with a few volunteers who are instructed to "think aloud" at they go through them. Two groups could be run, with one receiving the current instructions and the other the revised. This should provide useful information before changes are made.

Currently, practice items are presented in two groups. The first includes two easy to moderate items, with five additional items presented if the first two are answered incorrectly. Similarly, the second group includes two harder practice items, followed by four more if the first two are answered incorrectly. Therefore, the total number of practice items ranges from 4 (if answered correctly) to 9 (if answered incorrectly). The test starts even if the test taker answers all items incorrectly. To reduce the floor effect, the number of screens for practice items should be reduced, as should the difficulty level of the 2nd group (from hard to moderately difficult). The delay for the easiest practice items should be increased, with more detailed instructions from the TA, if needed. The examinee should be required to answer at least one practice item correctly before starting the operational test. Simply providing more, and harder, practice items is not enough to ensure a clear understanding, without which the floor effect won't be reduced. A practical concern is what to do if the examinee fails to answer any practice items correctly after repeated attempts. One solution would be to allow the examinee to continue with the test but assign a code to his/her record for later identification.

In the new practice test item sequence, items will still be shown in two groups, with the first two being easy items and the second two moderately difficult. The second group would include 5 easy items with response time adjusted to 830 MS, followed by 5 moderately difficult items with time adjusted to 500 MS. If the examinee answers either one or both easy items correctly, they will be given the second set and start the test afterwards regardless of whether they answer them correctly. If the examinee answers both items in the first set incorrectly, they will review the instructions and demonstration and start the practice again. If they fail the easy items again, the TA will be signaled to help, the examinee will review the instructions and demonstration again, and restart the first set of practice items with the TA's assistance.

Dr. Yin continued by addressing minor updates that will be made during operational testing. A pause will be added between screens to allow for additional time between answering one item and starting the next.

> There is currently a break between items 16 and 17, with a message appearing telling examinees that they are halfway through the test. Given that the average completion time for MCt is 4 minutes, this break is unnecessary and will be eliminated. Equating may be considered if there is any concern over the impact of these minor updates. Additional changes include implementing a short post-test feedback questionnaire to solicit additional guidance on future changes. This will seek input on the instruction, demonstration, practice, the test itself, TAs, motivational level, fatigue, and other factors. Statistical analyses will also be conducted to differentiate examinees who are engaged in rapid-guessing from those seeking solutions based on response time. Finally, if the floor effect is not reduced after implementing these changes, consideration will be given to implementing Option 1, described earlier.

As Dr. Yin described the task required by the MCt, a committee member clarified that boxes, and not numbers, appear. Dr. Yin agreed and said the boxes appear multiple times in different places and that the test-taker must track the number of times boxes appear at various places. The committee member then asked if there were ever more than two boxes shown at a time. Dr. Yin said that up to four boxes are shown at a time. Mr. Arendt reminded DPAC to make sure they explained this clearly in the instructions.

As Dr. Yin explained the findings related to the floor effect (slides 7 and 8), a committee member asked if DPAC had looked at the association between scores of zero on the MCt test and low scores on other tests. Dr. Pommerich said they have looked at the relationship of MCt scores and AFQT scores, and Dr. Manley added that, though there are zero scorers who score high on the AFQT, there is an overall correlation. Dr. Yin said that there are two options to address the floor effect, with one being to make the test easier by reducing the difficulty of some of the items. Another committee member then asked several questions about test administration, to include: do the items have varying amounts of delay between box presentation, how are responses recorded, how short is the allowed response time? Dr. Yin replied that responses are entered via the keypad. Dr. Pommerich added that the delay between the presentation of boxes varies and that response time is unlimited. The first committee member then sought and received confirmation that users respond after the boxes have stopped appearing. Dr. Velgach asked if it is clear to the user when the stimulus (i.e., flashing boxes) is complete, and Dr. Yin said yes. She also clarified that boxes appear in a random fashion, to which Dr. Velgach asserted that the order and locations of box flashes represent a component of the difficulty of the items. Dr. Yin replied that they did not want to create patterns that could be remembered easily.

LTC Rea asked whether a demonstration was provided as part of the instructions, suggesting that this was the best way for people to learn how to do the items. Dr. Yin said there was a demonstration. LTC Rea also asked if there were practice items in addition to the demonstration and whether DPAC thought the users understood the practice items. Dr. Yin said there are practice items, and Dr. Manley said that a user can proceed with the test even if they get the practice items wrong. LTC Rea asked if users could skip the practice items, and Dr. Yin said they do not have to answer them correctly. A committee member asked if the test provided explanations of correct answers. Dr. Pommerich replied that they are only told the correct answer, but no explanation is provided. The committee member asked why, and Dr. Yin said it was because they would have to show the item again in slow motion, and that the respondent would have to remember his/her incorrect answer to make sense of it. Another committee member suggested that the examples could be used like a teaching model. Dr. Yin replied by reiterating that they are attempting to update the instructions to provide more clarity. The committee member then commented that the test presents a complicated task.

As Dr. Yin briefed Option 1 (adding five easier items; slide 10), a committee member asked if the new items would replace current items or lengthen the test. Dr. Yin said it would lengthen the test. The committee member replied that s/he thought that this approach was attacking the wrong problem, and that the instructions needed to be improved. Dr. Yin agreed and pointed to Option 2.

On Option 2 (slide 12), a committee member described the "cons" as practical constraints rather than constraints in the same sense that the "pros" were pros. Dr. Pommerich explained that the cons were termed "cons" due to their associated costs. The committee member then suggested that the use of a cognitive lab could help DPAC better define the performance problem. Another committee member agreed, suggesting that many people may not even know they are performing the task incorrectly. A third committee member pointed to the fact that, although scores had improved with slight improvements in the instructions, they had gone back down again. Dr. Pommerich said that finding may be related to sample size, and another committee member agreed. Dr. Pommerich then said DPAC has considered conducting a "think aloud" study, but that it would be difficult to do due to logistics. A committee member said that they would not need very many cases. Dr. Pommerich said they had talked about using nine cases, so that no approvals would be required. The committee member replied that nine cases would probably be sufficient to explain what is going on.

When Dr. Yin described the importance that test-takers think their scores count (slide 13), a committee member replied that the scores do not matter. Dr. Yin replied that DPAC did not want users to know that in order to maintain motivation. Another committee member said that s/he has seen similar distributions for other tests, and that the effect had been due to low motivation. A third committee member suggested that it is likely that members of the population taking the MCt would be willing to put forth some effort. The first committee member agreed. Dr. Pommerich then interjected that there are some who do the task correctly, and so a lack of motivation does not apply across the board. Mr. Tiegs asked if there was a similar floor effect on the Coding Speed test. Dr. Pommerich said she was not aware of one, however, Dr. Segall replied that the distributions for Coding Speed are bimodal, suggesting a mixture of two groups: One scoring around chance and another scoring above chance and, apparently, following directions.

Continuing the discussion on factors contributing to the floor effect, a committee member said that mental exhaustion probably does not explain the zero scores but may explain low scores. A second committee member agreed and said that allowing additional time between items would likely not solve the problem. Dr. Yin responded that DPAC is focusing on clarifying the instructions rather than on the other solutions listed on the slide.

After Dr. Yin presented the MCt task demonstration, she explained that the demonstration used 50 different screens. A committee member asked if 50 screens were used in the actual test, and Dr. Segall said, no, that there would not be time for that. He said that they used so many screens in the demonstration presented to the DACMPT to allow a more critical look at the content. A committee member asked if test-takers could pause the demonstration. Dr. Pommerich said no, but that they could repeat it. Dr. Yin commented that color had been added for the DACMPT's version of the demonstration but was not present in the real demonstration. Another committee member asked how many milliseconds elapsed between screens in the demonstration, and Dr.

Manley said they had slowed it down to two seconds. The committee member then asked if test-takers could make notations during the test, to which Dr. Manley replied that they had eliminated that possibility in Version 3 of the test to reduce the ceiling effect.

A committee member said s/he was worried about test administrator (TA) interference. Dr. Manley said the proctoring procedures currently require that TAs intervene only if multiple practice items are answered incorrectly. Dr. Segall said that the ratio between TAs and test-takers is typically about 1:15. He also said that start times are staggered to ensure TAs would be available.

Toward the end of the briefing, as Dr. Yin briefed the pilot test described on slide 17, a committee member said the pilot test was like a cognitive lab, and Dr. Yin agreed. The committee member asked if a test-taker could view the demonstration more than once and if the test-taker manually cued the real test to begin. Another committee member asked how the test was initiated. Dr. Pommerich replied that there is a screen that says the practice items have been completed and the test-taker can hit a "back" button to see the practice items again. She said that the TA sees a red screen and must reset the test if the test-taker gets all the practice items wrong. A committee member asked if DPAC thought there might be value in using the 50-screen version of the demonstration operationally. Dr. Yin responded that the existing screens provide the same information as the 50-screen demonstration. Another committee member responded, however, that if the 50-screen version helped everyone a pilot to understand the task, then that would answer the research question. Dr. Yin said DPAC also wants to know whether the instructions are clear. Another committee member said that was a different research question. A third committee member asked, why bother with the old instructions if you know they are not working? Dr. Pommerich replied that they do not want to make unnecessary software changes. The committee member asked if making software changes was that large of an investment. Dr. Pommerich replied that it is difficult to make changes. A committee member concluded the discussion by asserting that DPAC was really proposing two experiments.

As Dr. Yin explained the new practice item sequence (slide 21), a committee member asked why the practice items included difficult items at all. Dr. Yin replied that the intent was to provide a practice experience that reflected the actual test as closely as possible. Another committee member asked why the instructions could not just warn test-takers that some items would be more difficult than the practice items. At this point, Dr. Manley pointed out that the most difficult items provided in the practice section are only of medium difficulty. Another committee member responded that only one or more of the easier items must be answered correctly. Another committee member asked if everyone received the same practice items, and Dr. Yin said they do. She said that if they fail the items a second time, the TA will receive a flag and the TA will work through the items with the test-taker. She said the test-taker will then take the items again; she said, however, that this cannot go on indefinitely, and that three times is the current limit. She said that, in the pilot test, DPAC wants to record these conditions to allow selected test-takers to be screened out. A committee member then asked if DPAC was going to allow test-takers to seek TA support if they fail to understand the instructions after the demonstration, but before they take the practice items. Dr. Manley said they were trying to minimize use of TAs. Dr. Pommerich clarified that there is a "help" button that will flag the TA, and that this is the case with all ASVAB tests. Dr. Segall said the button is available on every screen.

The discussion of the pilot test concluded with a committee member questioning whether practice item scores can be linked to test scores. Dr. Yin said that, hopefully, they can do that. Dr. Velgach asked if test-takers received the same practice items again if they missed them, which could lead to a correct response on second try without understanding the true response. Dr. Yin said they received the same items. Dr. Manley commented, however, that DPAC is not sure whether they will use the same items. A committee member emphasized that they should not repeat the same practice items.

6. CAT-ASVAB Form 10 Equating Study (Tab I)

Dr. Matt Trippe, HumRRO, presented the briefing.

Dr. Trippe began the presentation by explaining that CAT-ASVAB Form 10 was developed from old P&P forms for use in the CEP. Form 10 item parameters have been calibrated and scaled (through linear transformation) to be on the same scale as operational CAT-ASVAB forms 5–9. As an extra precaution, form 10 theta scores will be equated to theta scores on CAT-ASVAB form 4 (a reference form used for the purpose of equating analyses). Equating ensures that form 10 scores have the same meaning or can be treated interchangeably with operational form scores. Rigorous equating procedures were developed by DPAC to equate forms 5–9, which was the most recent equating. This was used as a template for equating Form 10, as well as a template for new Forms 11-14. Linear equating methods were used to derive constants to transform IRT-based theta scores on form 10 to scale of the reference form 4. This was done at the subtest level. Linear equating constants match the mean and variance of each subtest distribution, which works well to the extent that subtest distributions have similar shapes. Therefore, it was necessary to evaluate the comparability of composite distributions to ensure subtest equating resulted in sufficient precision.

Equating was performed in three phases of operational administration of Form 10 to military applicants, with each phase including progressively larger sample sizes. The intent of the phased design was to maximize the accuracy of the reported operational scores. This constituted a random groups design. Each applicant was assigned to a single Form with 1/6 assignment probability to either the reference Form 4, an operational Form (5, 6, 8, or 9), or Form 10. Provisional transformation constants were updated after phase 1 and phase 2 sample sizes were achieved. The differences in qualification composite cumulative distribution functions (CDFs) between reference Form 4 and Form 10 were evaluated. Provisional transformation constants were not replaced after phase 1 and phase 2 target sample sizes achieved as originally planned. Replacement of composite transformation constants is a non-trivial update to CAT-ASVAB that requires changes to the software. This is incompatible with data collection schedule. However, evidence suggests provisional constants were sufficiently accurate for reported operational scores. Transformation constants were updated after phase 3 target sample size was achieved on July 23, 2018. Operational Form 10 transformation constants will be replaced based on phase 3 results. Dr. Trippe then presented graphs showing the difference in qualification rates between Forms 10 and 4 based on phase 3 data and a table showing the data collection results (i.e., forms involved and numbers of applicants taking each.

Dr. Trippe then turned to phase III analyses conducted. To answer the question of whether assignment procedures produce equivalent groups with respect to key demographic variables, distributions were compared across assignment to Forms 4 and 10. As expected, assignment of groups to different forms was randomly equivalent. Dr. Trippe then turned to the question of whether linear transformation is adequate and whether subtest distributions have similar shapes. There was evidence of systematic differences in the shapes of subtest distributions, however qualification decisions are based on composite scores which are likely to be more normal in shape. Dr. Trippe then displayed graphics demonstrating these points. Composite scores can have different variances if the forms display different patterns of subtest correlations. Such differences were evaluated using the Kolmogorov-Smirnov test and the Cumulative Distribution Function (CDF) for reference group minus CDF for the new form group. Dr. Trippe then showed graphs displaying the composite distribution equivalence.

> The next question centers on whether subgroups (i.e., females, Blacks, Hispanics) perform at the same level across forms. One-way ANOVAs were performed with groups defined by form revealed two statistically significant differences in female analyses, however effect sizes were small. No other significant differences were observed. To identify how equated scores on Form 10 compare to operation forms, means differences were compared in Form 4 to Forms 5, 6, 8, and 9. Statistically significant mean differences representing small effect sizes were found in several tests, but there were no differences in the remaining tests or the AFQT. To identify how closely the provisional equating transformations matched the final transformation, all applicants who took Form 10 were rescored using the final transformation constants and the rescored values were compared to those used operationally based on provisional constants. Total errors were calculated as the sum of equating errors and measurement efforts, and total error was compared with standard errors of measurement. Only small differences were detected.

As Dr. Trippe briefed the equating design and procedure (slides 4-6), a committee member asked if Phase 2 of the data collection included the two weeks of Phase 1, or if it was a separate 3-week period. Dr. Trippe said that Phase 2 was conducted over three weeks following the conduct of Phase 1.

On the topic of equating results (slide 7), a committee member said s/he understood that the items in Form 4 had been calibrated previously and that equating constants were based on those parameters. S/he then noted that Form 4 had been administered in the current effort to obtain the distributional properties of the scores. S/he asked if the items on Form 4 had been recalibrated. Dr. Trippe explained that Form 4 defined the operational scale, and that the adjustment was made to the Form 10 scores. Another committee member asked if equating was accomplished by transforming thetas and not item parameters. Dr. Trippe replied that, prior to Phase 1, rescaling/linear transformation had been performed on Form 10. Dr. Pommerich clarified that Form 10 had been calibrated on data obtained at pretesting (i.e., an earlier stage). The first committee member asked how the Form 10 items were previously administered, and Dr. Trippe said that the test had been administered by P&P. Dr. Segall commented that some programs, after putting the item parameters on the same scale, may skip the step of equating new form scores to the operational form scale. Dr. Trippe added that this is an extra precise way to follow up on what is typically done.

Continuing the discussion of equating results (slide 8), a committee member asked for clarification on the scales used for the three tests. Dr. Trippe said that the composite score scale values were shown on the horizontal axes and the differences between Forms 4 and 10, in terms of qualification rates (scoring at that level or above), was shown on the vertical axes. He also clarified that the red line showed the difference for provisional scores and the blue line showed the difference based on updated scores, or scores that had been adjusted by equating. The committee member asked about the qualification rate and whether the graphs were indicative of accession decisions. Another committee member explained that the charts identified the percentage of people that qualified at various scores. Dr. Segall confirmed this interpretation as being correct. Dr. Trippe explained that the graphs illustrated that the differences were "practically tolerable," indicating that the study could proceed with the use of the provisional constructs. He continued, saying the differences did not rise to a level that would justify "blowing up the study," and that post-hoc analysis drove that home; that is, there was an expected level of equating error in relation to measurement error. A committee member then asked about how the updated transformation constants were obtained. Dr. Trippe said his team had carried out the study as designed and updated the constants after each phase. He added,

however, that the equating constants had not been replaced in the WinCAT system as planned, which meant that people who participated in the study and were administered Form 10 received scores based on the provisional transformation constants and not on the updated constants. To clarify, the committee member asked if this work had been performed with existing data, and Dr. Trippe said that it had.

On slide 10, a committee member asked if the provisional results were used at the time of Sample 3. Dr. Trippe said, yes, and he explained that they compared and rescored one sample with the updated results to see what it would have been otherwise. He went on to say that all scores on Form 10 were ultimately based on the provisional results. Another committee member asked if continuing work should be based on provisional scores. Dr. Segall responded that some people compare the new form with a reference form, so he would want to know how the composite distributions compare, and that would require a case-by-case review. He said that sometimes it makes sense to do this, but that sometimes it does not. Dr. Trippe concurred.

Dr. Trippe described how some of the subtest distributions (slide 13) were "less tidy" than others, but he said that the composite distributions (slide 14) looked "pretty good."[1] At that point, a committee member asked if the updated equating method could be used when Form 10 is operational, and Dr. Trippe said, yes, the final form will be based on Phase 3 and that form is what will become operational.

On composite distribution equivalence (slides 16-17), a committee member reiterated that the intent is to determine whether Form 10 can be used in the CEP; that is, it is providing scores that are comparable to what test-takers would have received on the operational form. S/he stressed that the most important component of the battery is the AFQT, because qualification is based on AFQT scores. S/he said that it is comforting to see good alignment with the operational form in that area. Dr. Trippe responded that the AFQT had one of the largest levels of agreement.

Upon viewing the estimates of provisional transformation accuracy (slides 21-22), a committee member asked how equating error was estimated. Dr. Trippe answered, in part, that it included an unknown component, after which he cited the equations on slide 26. The committee member asked if the provisional scores contained equating error, and Dr. Trippe said that they did. Another committee member then observed that the slides revealed the difference between the original and equated scores. Dr. Segall clarified that the difference was in the amount of unwanted variance introduced by using the provisional scores operationally. The committee member said s/he appreciated the clarification, because s/he would not have wanted to use those scores. Dr. Trippe described that part of equating error as being the extent of damage caused by using the updated scores. The committee member suggested that a better label for the estimate might be provisional equating error.

[1] After the meeting, Dr. Trippe clarified that equating is applied to the individual subtests (e.g., WK), but no decisions are made based on individual sub-tests; rather, decisions are made at the composite level (i.e., combinations of subtests).

7. Development of New Cyber Test Items and Pools (Tab J)

Dr. Matt Trippe, HumRRO, presented the briefing.

Dr. Trippe began the presentation by explaining that the development of the Joint Service Cyber Test (CT), formerly known as the Information and Communications Technology Literacy (ICTL) test, began over 10 years ago. The CT is modeled after ASVAB "information" tests (e.g., Electronics, Auto, Shop Information) and is administered on the ASVAB platform as a linear (i.e., non-adaptive) test. The CT has demonstrated evidence as a valid predictor of training success in several Air Force, Navy, and Army technology-related occupations. Like any selection test, the CT requires periodic maintenance to review and refresh the item pool. The Air Force developed 251 new Cyber Test items in 2015. These items, along with the prior pool of items (n = 167), are intended to be transitioned to a computerized adaptive testing (CAT) platform. The item pool includes many items that provide information on relatively high-ability applicants. The objective of the work described here was to review, calibrate, and equate the items to the current operational scale, and replace the static forms with two CAT forms/pools. This involved the development of 200 new items targeted toward the middle and low end of the ability distribution.

The first step was to conduct a review of experimental item quality of the 243 experimental items administered at the MEPS. Items were screened in similar manner to ASVAB experimental items, beginning with empirical evidence such as classic test statistics (p-value, item-total, option-total r) and IRT evidence (three-parameter logistic model; 3PL). SME content reviews were guided by item statistics. SMEs evaluated items with potential issues identified by distractor analyses or other empirical guidance. The review was based on a large applicant sample, with a total n of 84,988, and an average of 3.386 responses per item. About 48% (117) of the items were retained, which is a relatively low survival rate. Item difficulty was the primary factor in experimental item loss, given that many content areas are inherently complex or technical.

The 117 new items were placed on the operational scale established in 2011 via Stocking & Lord (1983) procedure. All viable CT items (n=284) were included in form assembly using Automated Test Assembly (ATA) as described in van der Linden (2005). A binary/integer programming approach was taken in which form specifications are set up as quantities to be minimized (e.g., TCC between forms) or maximized (e.g., score information) against an objective function. Additional constraints such as item enemies, content specifications, and keyed response, are incorporated into the model. The goal was to create two parallel CAT forms.

In addition to equating and form assembly, this work focused on developing 200 new items targeted toward the middle and low end of the ability distribution scale. SMEs were provided feedback on the empirical difficulty of items they wrote in 2015. Items of "easy" and "moderate" difficulty are challenging to write in this content area because information is concentrated at the high end of the ability distribution. Prior to developing each new set of experimental items, a blueprint "validation" or review was conducted to determine test item content. This comprises 40-50 knowledge, skill, and ability (KSA) statements organized into four broad content areas. The individual KSA statements serve as stimulus for development of new items (e.g. "Knowledge of network addressing concepts," "Knowledge of operating system internals").

For the blueprint validation, SMEs from the Services reviewed the CT blueprint for relevance or potential obsolescence. They were provided with KSA statements in the current CT blueprint (n = 49) and KSA statements from National Initiative for Cybersecurity Education (NICE) framework (n = 61). In standardized rating exercise, SMEs rated (a) should this KSA be acquired prior to enlistment? (b) how important is this KSA for successful performance in entry-level training for enlisted cyber occupations? (c) given ongoing technological change, how stable do you think this KSA will be over time? and (d) whether KSA statements were missing. This resulted in minor/marginal updates to knowledge, skill, ability (KSA) statements included and content area weighting. Dr. Trippe then presented a table showing the total number of KSAs, and the number retained, dropped, or added to the previous blueprint for the years 2008, 2011, 2015, and 2018. Another table showed the blueprint category weights for those same years. The weights for

2018 were 30% networking and telecommunications, 30% computer operations, 25% security and compliance, and 15% software programming and web design.

Dr. Trippe concluded the presentation by providing an update on the project status and schedule. The technical review is complete and equating is in progress. Form assembly should be completed by October 2018. In all, 200 new items have been written, and editorial and technical reviews are complete. The work will be documented in a technical report which should be completed by December 2018.

As Dr. Trippe described the agenda (slide 2), a committee member asked if he could address data dimensionality. Dr. Trippe replied that Dr. Gao would be talking about that in the next briefing. Another committee member, commenting on the addition of medium and low difficulty items, asked if scores in the medium and low range were useful. Dr. Trippe said that he did not have a slide on the use of scores, but he said that the operational cut score is 60, equating to a 1 on the theta metric. He said the committee member's point was relevant in that not everyone is expected to have a lot of knowledge in the area and that the military is mostly interested in candidates who score on the high end. Dr. Pommerich recalled a previous DACMPT concern that there was a shortage of information in the moderate ability range. The committee member then asked if the addition of the new items would make the test function better in a CAT environment, or if the information was really necessary. Dr. Segall joined the discussion, asking Dr. Trippe if the Services use the Cyber Test as a standalone cut. Dr. Trippe said the AF uses the test as a compensatory measure; that is, if a person misses the initial cut, but has a relatively high Cyber Test score, that may be enough to qualify them a candidate for certain jobs. Dr. Velgach added that, for the Navy, the test was included in the composite but was also used as a hurdle. CDR Phillips agreed and said that it was difficult to obtain a good distribution of scores around the cut score due to the difficulty of writing moderately difficult and easy items. Dr. Trippe agreed, reiterating that the distribution was not comparable to those of other tests.

A committee member asked whether the 243 experimental items described on slide 5 had been developed to fill the gap at the low end of the difficulty range. S/he also asked if the items included new types of items or focused on new content. Dr. Trippe said that the items referenced were not fundamentally different than the original item set in any of those ways. He said, however, that the new set of 200 items was to include a larger number of easy and moderately difficult items.

As Dr. Trippe briefed on equating and form assembly (slide 6), a committee member asked if the two forms were really "CAT pools," and Dr. Trippe said they were, and that two pools were required to support retesting. He said that each pool would include 142 items, and Dr. Segall added that DPAC had developed smaller CAT pools that compared favorably to conventional tests. The committee member asked about the degree to which the tests are adaptive, and Dr. Segall replied that most tests have some exposure control and minor content balancing. The committee member asked if there were content specifications, and Dr. Gao said that she would address that topic in her briefing. Dr. Segall added, however, that he thinks unidimensionality can be assumed, making content coverage a non-issue. Another committee member asked if there would be two forms and whether those two forms would overlap. Dr. Trippe said he expected to recommend the creation of two, non-overlapping forms. The first committee member then said that the key will be to relax the content constraints. Dr. Trippe replied that the two forms will be balanced in relation to the blueprint. The committee member asked how many items would be administered, and Dr. Trippe said 29. Dr. Segall then said that the goal would be

to reduce the number to 15 on the adaptive test as they replace the conventional test. The committee member then asked if Dr. Trippe was only interested in the single CT score theta scores. Dr. Trippe concurred and added that, though there are four content areas, earlier analyses had revealed exceptionally high correlations among subscores (i.e., in the range of desirable reliability estimates). To clarify further, Dr. Segall said that there is a relationship between difficulty and content area; that is, there are very few if any "easy" items in the software programming content area, and so enforcing content balancing essentially undermines the purpose of the CAT, which is to administer items targeted toward the ability level of the examinee.

On new item development (slide 7), a committee member commented that the ability of test-takers to answer items correctly depends on their experience in the targeted content areas. S/he said that competency is *not* likely something that is learned over many years (e.g., math), but is based primarily on whether a person has had some experience in the area. Dr. Trippe agreed, citing the example of setting up a home wireless network. Mr. Arendt said the military is using the test to identify individuals who are more likely to assimilate training more easily. Dr. Trippe said that an Army study found a relationship between Cyber Test scores and related technical area MOS satisfaction, which he described as a job-fit assessment.

As Dr. Trippe discussed the blueprint validation review (slide 8), Mr. Arendt asked if anyone was looking at high school curricula to see what is being taught. He said if coding is being taught, for example, items on coding should be included. A committee member agreed, whereupon Mr. Arendt said that if a person has a propensity or interest in an area, then they would likely have had some involvement with it during or before high school. Another committee member said that NAEP had a similar assessment. Dr. Trippe responded that his recollection of that assessment was that it measured a person's ability to use technology (e.g., the Internet) to do research on a topic, rather than a person's understanding of computer hardware and software. The committee member replied, however, that it may provide material if DPAC is looking for easier items to include on the test. Dr. Trippe concurred, but he added that they were interested in content in the areas covered by the National Initiative for Cybersecurity Education (NICE) or DoD Cyber Workforce Framework (DCWF) frameworks. Dr. Velgach clarified that the DCWF is what the military is using to cover the competencies across all cyber positions.

A committee member then suggested that successful performance during entry-level training is important, but that it is not clear whether any knowledge of course content is required before training or if it can all be learned during training. Dr. Trippe said that the most common argument against the test is: "we don't care if they know anything about it; we just want smart people." He said, however, that everyone wants smart people, but that the CT can be used as a compensatory factor for people with moderate *g* scores. He said that hobbyists tend to work out well, and that they are, in fact, sometimes the best candidates. He said they may not have any formal training, but they know how to do the work. The committee member responded, however, that their real skill might simply be problem-solving (e.g., the use of trial and error).

Another committee member then asked if the plan was to make the CT available on the platform so that a Service could have all its applicants take the test, or if the idea was to allow Services to point selected applicants to the test. Mr. Arendt said one Service had taken the first approach. Dr.

Trippe added that the Army uses the test for applicants who want to get into certain MOS. Mr. Arendt cited the use of DLAB test, which was to identify cryptolinguists by testing everyone who came to Language Training. LTC Rea commented that the Army (for officers) followed the model used in the medical and judge advocate general areas, noting that the Army has specific requirements for cyber commissioning. Mr. Arendt said that a person's background and civilian certifications are important. LTC Rea agreed and said that four-year degrees and certain areas of applied experience are relevant.

Mr. Schwartz commented on the fact that the SMEs liked the idea of puzzle items. Dr. Trippe said that the puzzle items performed sufficiently, but were not more informative than the AO items, which were already capturing that space. Dr. Segall explained that the addition of *g*-like items had been considered in early versions of CT, but that *g* was already being measured by other tests, and so it was decided that the CT should include only cyber-related items. Dr. Trippe said it is hard to beat the ASVAB in measuring *g*.

On scoring, LTC Rea said that the cut score for the combination of the General Technical (GT) and Skilled Technical (ST) was 112, and now there was a CT cut score on top of that. He asked about the cut score for the CT, and Dr. Trippe replied that the AF uses a cut score of 65.

Toward the end of the briefing, a committee member asked if the new forms would conform to the balance of blueprint category weights shown on slide 10. Dr. Trippe said that the team's Automated Test Assembly (ATA) program would result in a close match to the blueprint weights; that is, the plan is that the item pool will be representative of the blueprint (4 areas), but items received during the test session would be content balanced, due to unidimensionality. Another committee member asked if the new items would be completed soon. Dr. Trippe said they had been developed and were ready for seeding.

8. Sparse Data Dimensionality Assessment with Application to the Cyber Test (Tab K)

Dr. Furong Gao, HumRRO, presented the briefing.

Dr. Gao began by explaining that the IRT model fit affects the accuracy of item parameter estimation, test scores, and classification of the test takers. The current algorithm for item selection in the CAT-ASVAB assumes unidimensionality for a given test without content constraint. Concerns were raised by the DAC about potential content or item difficulty shift in continually developed new items and the potential impact on the CAT item-rendering algorithm and content constraints. Strictly unidimensional tests are theoretical in nature and do not exist in practice. Tests that are carefully constructed to measure only a single dimension (construct) often show one or more minor dimensions. If there is unidimensionality, a unidimensional model will adequately represent the test data. However even tests that are essentially unidimensional usually display a bi-factor structure; the intended dimension/construct with items from different content domains, with items in each content domain measuring a secondary (minor) dimension.

Dr. Gao continued by explaining the approach taken to assess dimensionality with complete data. An item score matrix is formed including the number of examinnes, number of items, and the item score for each examinee. Dimensionality assessment can be covariance or IRT based. However, for seeded items, each examinee gets a small set of items from the experimental pool, resulting in a sparse data matrix. Missing data patterns can be missing at random (MAR, where the "missingness of an item score is not related to the missing value, but related to some of the observed data), or missing completely at random (MCAR).

In a situation where there are sparse data, the covariance-based approaches will not work. Other possible approaches include full information maximum likelihood (FIML) estimation, where only observed data are used with no direct imputation of missing values. This will produce unbiased estimates under MAR and MCAR. A factor analysis approach can also be taken using an R package for structural equation modeling. IRT model-based item factor analysis software is also available in which a Markov change Monte Carlo estamation method is employed. Assumptions include that the test is designed to be unidimensional and measure a single construct but with broad content coverage tha may introduce minor additional unintended dimensions to the test data. Items are rendered in such a way that the "missingness" in the response data is MAR or MCAR. Both the CAT-ASVAB and the current seeded item design produce MAR data. Confirmatory analyses determine if the data will fit both a one-factor and a bi-factor model. In the bi-factor model there is one general factor on which all items load (*G)* and secondary factors, one for each of the content sub domains. All factors are assumed to be independent of one anothter. When the *G* factor loadings of the two models are compared small and neglible differences are expected. Specific factors do not distort the meaning of the general factor that is measured by all items on the test. Explained common variances (ECV) provide an indicator of essential unidimensionality, and are calculated using the factor loading values of the *G* factor and the secondary factors of the bi-factor model. Dr. Gao continued by displaying the ECV formula. The larger the ECV, the stronger the unidimensionality,

Turning to the Cyber Test (CT), Dr. Gao displayed the four broad content areas: Computer Operations, Networks and Telecommunications, Security and Compliance, and Software Programming and Web Development. The bi-factor for the CT would include one general factor (G), and one secondary factor for each of the content areas, with all the factors being independent of one another. The dimensionality assessment was used with CT Form 1 data, which includes 29 items and a total of 65,289 test takers. The IFACT program was used with 2,000 burn-in cycles and an additional 2,.000 cycles for posterior summarization. Dr. Gao then showed charts summarizing the results, which yielded a ECV of .865 (.866 when adjusted for standard error of measurement). Similar analyses were run on the same data with the addition of 117 seeded items. The results showed an ECV of .909 (.922 when adjusted). Results of analyses of Form 2 data (n = 68,928) yielded an ECV value of .893 (.895 when adjusted). Analyses of data combining Forms 1 and 2 and 117 seeded items, with case counts on seeded items ranging from 6,493 to 7,678, yielded an EDV of .911 (.916 when adjusted).

Dr. Gao concluded by reiterating that well-constructed unidimensional tests often display a bi-factor structure with one dominant general factor and minor secondary factors that are negligible. These tests are essentially unidimensional, and the response data can be adequately modeled/explained by unidimensional IRT models. The CT analysis results demonstrate this to be the case. Further analyses will be carried out on CAT-ASVAB data, both simulated and operational. DPAC will continue to monitor for potential content or item difficulty shift.

As Dr. Gao briefed the assumptions related to the dimensionality assessment approach (slide 8), a committee member said that unidimensionality is an assumption for using confirmatory factor analysis. With the test being constructed to be unidimensional, the confirmatory approach either confirms or rejects the unidimensional assumption. Dr. Segall replied that, though there can be sub-dimensions in these tests, the general dimension is always predominant. The committee member then said that s/he understood that the assessment would take the bi-factor model into consideration as well. Another committee member then asked if the missing data should be completely random, and the first committee member said that it would. Dr. Segall, further pointed out that an exploratory analysis would be helpful in determining the number of group factors if the test blueprint was not available.

On the topic of confirmatory analyses (slide 8), a committee member asked how DPAC would determine the number of group factors. Dr. Gao replied that, for the CT, the number of group factors would be determined by the number of blueprint content areas.

As Dr. Gao described the explained common variances (slides 9 and 10), she confirmed a committee member's assertion that all the items on the test would be accounted for. Another committee member then asked if the equation was really for the bi-factor model and said that it seemed unusual that using the "adjusted explained-common variance" formula each item's factor loadings (squared) were adjusted by subtracting their estimated errors (squared). Dr. Segall replied that adjustment accounted for error in the estimated loadings, which would tend to bias the statistic and make the test appear less unidimensional. The committee member replied that this was for the optimal adjustment. Dr. Segall responded that, for small samples, you would otherwise get spuriously high or low loadings and the direction would not be known. He added that the squared values are high (regardless of sign), so it would not appear to be unidimensional. He concluded by saying that the method selected was an attempt to eliminate the chance of that happening. The committee member replied that this appeared to be a direct reduction. He added, however, that because the point is that they are being summed, it did not present an issue. Dr. Segall replied that, in this case, it did not make much difference because the sample sizes are so large.

As Dr. Gao described how sparseness came into play upon considering the seeded items, a committee member noted that the test was fixed and not adaptive. Another committee member then asked if the seeded items were included in the 29 items mentioned on slide 13. Dr. Gao said they were not included in that total and that the seeded items were administered, but not in the present scenario. The committee member asked if the seeded items, when included, would be mapped to the four subscales, and Dr. Gao said that they would.

A committee member noted that, in the bi-factor graph on slide 14, the items that loaded higher on the *g* factor seemed to load negatively on the second factor. S/he said that there was a downward trend, which s/he called curious. Another committee member replied that the analysis might be parsing noise, because the correlation between the two *g* factors was .99. S/he said that any remaining variance must go somewhere, so it is probably not an issue.

As Dr. Gao briefed the scenario that included the seeded items (slide 15), a committee member observed that the distribution of items across factors looked similar to the distributions in the scenario that did not include the seeded items. S/he asked if it was important to maintain the same distribution of content with regards to seeded and operational items. Dr. Gao said that this was something they would need to consider when they constructed the CAT pools. The committee member then said that s/he agreed that this was not an issue for the current analysis but that it was important/critical to pay attention to when constructing CAT pools. The committee member then asked if the second scenario reflected the data from the same subjects and if the seeded item data were just not included in the first scenario. Dr. Gao said that was the case.

When reviewing the results from scenario 2 for Form 1 (slide 16), a short discussion clarified that the seeded items had previously been reviewed and piloted, and that the item statistics supported including the items in the study. A committee member then observed that the same variations seen in scenario 1 were also present in scenario 2. Dr. Gao replied that the test still appeared to be unidimensional. A discussion followed on whether the analysis results, in which the four sub-factors were determined in accordance with the four content areas, were truly indicative of a unidimensional outcome. During the discussion, Dr. Segall noted that the analysis used a multi-dimensional IRT model and, thus, could be interpreted like a 3PL model. He also

responded to a committee member's comment that there is even greater dispersion than was initially apparent, noting that he suspects that there is a range of items that are highly discriminating. Items with low/small loading values are indicative of low discrimination values. He added that as the items become less difficult or extremely difficult, they are likely less discriminating. A committee member suggested that the difficult items were those primarily responsible for discrimination, but Dr. Segall replied that there is still some variance in difficulty even within the difficult items. The committee member asked Dr. Gao if she had calculated fit indices, and Dr. Gao said that she had and that she had also looked at the predictive *p* values. She said that the *p* values matched the fit indices very well, but that she had not looked at other fit indices yet. She also pointed out that DPAC is conducting simulation studies to assist in that area. The discussion closed with Dr. Segall commenting that the IFACT program, which uses a bi-factor approach, will result in a good fit.

As Dr. Gao briefed the Form 2 data IFACT results (slide 17), a committee member asked for and received confirmation that the ECV is the proportion of variance accounted for by the general factor.

At the end of the briefing, a committee member suggested that DPAC should be careful about saying the test is unidimensional based on an analysis that relied on defining factors in accordance with confirmatory factor analysis as opposed to factors that could be generated from an exploratory factor analysis. The committee member also said that a negative relationship between item difficulty and item discrimination may account for what was happening with this test. Dr. Segall pointed out that it depends on the test (e.g., the ASVAB vs SAT); he said that Dr. Susan Embretson had once commented on the fact that the ASVAB contained highly discriminating Arithmetic Reasoning items of moderate difficulty, whereas the SAT found these item types to display high discrimination only over low ability regions. So, items from particular subtests or domains might have high discrimination only over limited ability regions, and these regions of high-discrimination might depend on the examinee population.

9. TAPAS Expert Panel Update (Tab L)

Dr. Tim McGonigle, HumRRO, presented the briefing.

Dr. McGonigle began by explaining that the Tailored Adaptive Personality Assessment System (TAPAS) was originally developed by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) and Drasgow Consulting Group (DCG) to measure up to 27 facets of the Big Five personality dimensions, TAPAS uses multidimensional pairwise preference (MDPP) items which generally present two statements from different personality dimensions that are matched on the strength of the dimension and on the socially-desirable nature of the response options. It is structured to make faking more difficult because the "correct" answer is difficult to identify. Items are generated on-the-fly by selecting from pools of pre-calibrated personality statements that measure construct dimensions relevant to performance in the military. Approximately one million statement combinations are possible. TAPAS is scored using multi-unidimensional pairwise preference IRT (ideal point) model. The Army, Navy, Air Force, and Marines have all collected TAPAS data on applicants, showing evidence of incremental validity beyond ASVAB for training and military success criteria (e.g., attrition).

Some stakeholders have raised technical concerns about TAPAS, especially due to low test-retest reliability. RAND recently completed an independent evaluation of the reliability and validity of TAPAS. They analyzed data from candidates who completed TAPAS between March 2010 and April 2015 and

subsequently completed at least six months of service. RAND found small, significant incremental validity over education credential in predicting attrition. They also found low test-retest reliability in some conditions (r_{xx} = 0.07 (TAPAS 9/10/11, Army recruits who failed first test), but not as low under other conditions (r_{xx} = 0.59 (TAPAS 5/7/8, Air Force all recruits). Consequently, DPAC requested the establishment of a Technical Expert Panel (TEP) to independently review the body of TAPAS research and make recommendations regarding the readiness of TAPAS for operational use. The panel will review related research conducted by the Services, both on TAPAS and on other instruments (e.g., interest inventories), and make recommendations for future research and development. They will also comment on the readiness of TAPAS for operational use.

Through discussion, it was decided that the TEP should include five experts whose research and practice have involved personality measurement and the use of non-cognitive measures for selection, bringing both fresh perspectives and familiarity with TAPAS research. Five criteria were applied for recruiting TEP members. The overall panel should include members with (a) a familiarity with TAPAS research and development; (b) an independent perspective on personality measurement; and (c) knowledge of psychometrics and IRT, particularly ideal point IRT modeling with forced-choice pair-wise comparisons. In addition, the TEP should be diverse in regard to race/gender and come from both academic and practitioner backgrounds. DPAC and HumRRO independently identified potential candidates (N = 35) who were grouped into four categories: (1) National-level testing experts, (2) Psychometrics experts, (3) Personality theory experts, and (4) Operational testing experts. Candidates were ranked within the four categories and recruited from the top down. All top choices agreed to serve. Dr. McGonigle then displayed a table showing panel members, their titles and affiliations, and their areas of expertise. He also provided short biographies of each member demonstrating their qualifications.

Dr. McGonigle then explained that HumRRO will provide research support to the TEP, which will focus mostly on synthesizing existing research. The Services have been asked to provide a comprehensive set of existing research studies. Potential research topics include (a) a meta-analysis of TAPAS validity across Services; (b) additional analysis if test-retest reliability of TAPAS; and (c) the effects of coaching, regression to the mean, random responding, and motivation on reliability. The first TEP meeting will be held October 22 in Atlanta, GA. Points of Contact will be identified from each Service to provide existing research. Both RAND and DCG will be asked to be involved. At this time, four meetings of one day each are anticipated, with a final report delivered by October 2019.

As Dr. McGonigle described the TAPAS (slides 3-4), a committee member asked for clarification on the meaning of "failing the first test." LTC Rea responded by saying that individuals who fall below the 10th percentile in relation to a cumulative type score may be considered to have failed the test. The committee member said that, when evaluating test-retest reliability using a sample that does not meet the minimum score the first time they completed TAPAS, severe restriction of range may surface, which might attenuate the test-retest reliabilities. Dr. McGonigle replied that individual studies have done different things, such as use different TAPAS forms, different samples, and different testing conditions, and have found different levels of reliability. A committee member then said that some of these studies used old TAPAS forms and that the test has always been adaptive. In response, Mr. Tiegs pointed out that there are static forms, but that they have not been used since the test went operational. Dr. Salyer replied that the forms used in DLAB are static, but Dr. Segall explained that all the forms are based on the adaptive version. Dr. McGonigle summarized by saying that the test had been designed, from the beginning, for adaptive implementation.

A committee member continued the conversation, saying that the correlations shown for the Army recruits who failed the first test were not useful because the sample was range restricted to candidates who failed the test. The committee member also said that s/he also did not know how much the TAPAS had evolved over the years in terms of the number of items or other factors

that might influence reliability. Dr. Segall replied that there have been changes in the facets measured. Dr. Velgach then said that the Navy research group had done some work in this area and suggested that Dr. McGonigle reach out to that group.

A committee member asked if the TAPAS had been originally designed for use in a military context, and Dr. McGonigle said that it had. Mr. Arendt then said that he did not know if it was used more widely, to which another committee member responded that Drasgow Consulting Group (DCG; the developer of the test) uses the framework/methodology in other venues, but that the item content for military TAPAS is different. Dr. Manley explained that DCG owns the rights to develop other versions of the test. Dr. Segall further explained that the items used in different versions of the test can be very similar to the items used in the military's version and the Small Business Innovation Research (SBIR) contract, under which the test was developed, promotes transitioning methods from military to industry use. CDR Phillips said that the Navy had its own item pool, which has been in use for eight years.

10. <u>Adverse Impact</u> (Tab M)

Dr. Greg Manley, DPAC presented the briefing.

Dr. Manley began by explaining that impact can occur when groups that are not matched on ability perform differentially on an item or test. Adverse impact occurs when a group is disadvantaged by those performance differences. Bias occurs when an item or test unfairly favors one group over another. The occurrence of bias is problematic because it can negatively affect test validity. However, the occurrence of (adverse) impact does not necessarily mean that a test is biased. Adverse impact is examined by comparing the performance of a reference group in relation to a focal group, with the focal group being potentially disadvantaged. In the case of the ASVAB program, the groups are (a) males/females, (b) non-Hispanic Whites/Hispanic Whites, (c) non-Hispanic Whites/Non-Hispanic Blacks, and (d) Non-Hispanic Whites/Non-Hispanic Asians. Pairs a-c are the same groups that are used in evaluating differential item functioning (DIF), while group d is added because non-Hispanic Asians now represent more than 2% of the applicant population. Adverse impact has been assessed in 2005, 2009, 2011, 2013, and 2015. The data presented here concern applicants testing in fiscal year 2017 (October 1, 2016 through September 30, 2017).

The four-fifths rule is often used to determine the occurrence of adverse impact: "A selection rate for any race, sex, or ethnic group which is less than four-fifths (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact." [Section 60-3, Uniform Guidelines on Employee Selection Procedures (1978); 43 FR 38295 (August 25, 1978)]. The ratio comparing the selection rates is called the impact ratio, with the selection rate for the focus group divided by the selection rate for the reference group. Dr. Manley then presented a formula used to determine the statistical significance of the impact ratio and the confidence intervals around the impact ratio.

The four-fifths rule and accompanying statistics are applied to the ASVAB by comparing qualification rates across the focal and reference groups of interest with regard to: (a) examinees who qualify for entry into the military (i.e., those scoring in AFQT category IIIB or higher, AFQT $\geq$ 31), and (b) examinees who qualify for enlistment incentives (i.e., those scoring in AFQT category IIIA or higher, AFQT $\geq$ 50). Adverse impact is measured using initial test scores only (i.e., scores from retests or confirmation tests are excluded from the analyses). Effect sizes (i.e., standardized mean differences) provide another method of evaluating impact across individual ASVAB tests, where no direct selection occurs. The effect size is calculated by subtracting the mean score of the focal group from the mean score of the reference group and dividing by the pooled standard deviation across the two groups. Effect sizes can be plotted and classified with respect to Cohen's (1988) standards of evaluation, with a small effect size starting at .20, moderate at .50, and large at .80.

Dr. Manley then displayed a chart showing the sample sizes for the various comparison groups for the FY 2017 analyses. A series of charts displayed the results of the analyses conducted. Based on these outcomes, Dr. Manley concluded that the magnitude on the ASVAB has remained fairly constant across fiscal years, but still varies in size from negligible to large across tests and groups. A comparison of the impact across different testing programs gives some indication of whether the observed FY 2017 magnitudes are reasonable. Sufficient information for estimating effect sizes is available online for two other large-scale testing programs—the SAT and the National Assessment of Educational Progress (NAEP). Dr. Manley then presented charts showing comparisons of the effect sizes from these programs and the ASVAB in math, verbal, and science tests for the various groupings. He concluded that, for the AFQT tests (and General Science), the direction and magnitude of overall impact is largely consistent with that observed on comparable SAT and NAEP tests, which suggests that the impact on ASVAB tests may reflect legitimate differences in the studied groups. He noted, however, that comparisons across programs may be somewhat restricted due to differences in such factors as group definitions, testing populations, and test content.

Adverse impact does not reflect bias if validity research shows that the test is equally valid for relevant groups. Historically, a regression-based approach has been advocated to evaluate the existence of bias. Lack of bias is indicated when the regression line relating the test score and the criterion is the same for each group. Previous research on the ASVAB technical tests showed similar prediction lines across (1) males and females, and (2) Blacks and Whites, suggesting no bias for the tests and groups studied. DMDC recommended in 2010 that an updated validity study be conducted for relevant tests and groups, but a lack of access to criterion data across Services (except the Air Force) presents an impediment to updating the study. More recent thinking in the realm of bias detection is that regression-based approaches may not accurately reflect bias. Reducing adverse impact will be a high priority when considering revisions to the content of the ASVAB and AFQT.

As Dr. Manley briefed the effect sizes for the ASVAB scores (slides 14-18), a committee member asked if all recruits take the tests listed on the slide. Dr. Manley said that they do, because they are all included in the ASVAB battery. Referencing the effect sizes for Non-Hispanic Whites versus Non-Hispanic Blacks, a committee member noted that the variation between whites and blacks – about a standard deviation – was similar to what is seen with the National Assessment of Educational Progress (NAEP) examinations. Dr. Manley commented that the difference is large in this comparison, but not atypical.

In response to seeing the effect size comparisons between Non-Hispanic Whites and Hispanic Whites for non-AFQT tests (slide 22), a committee member asked if the effect sizes had been corrected for range restriction. He then asked if the Assembling Objects (AO) test was less reliable than the other tests, resulting in a lower impact. Dr. Pommerich said that AO is not less reliable, but that there is no language component to the test.

As Dr. Manley explained the effect sizes for Non-Hispanic Whites versus Non-Hispanic Asians on AFQT scores (slide 25), a committee member asked if DPAC had TOEFL (Test of English as a Foreign Language) scores for the Asian sample. Dr. Pommerich said, no, because it is not a military test. Mr. Arendt added that the number of non-national Asians is small, perhaps only five to eight individuals per year.

Mr. Tiegs asked Dr. Manley if he had generated a graph like the gender representation chart (slide 34) for other demographics, such as race. Dr. Manley said that they only produced the chart for the male-female comparison, and it was done only to further examine the interesting results showing a small effect size in favor of males on the WK and VE tests (slide 32). He said the chart was designed to show that females are a much smaller percentage of the military

population than the NAEP and SAT populations, and that they, thus, might not be representative of the females in those populations.

When Dr. Manley had finished briefing the effect size comparisons, a committee member again raised the issue of inconsistency in effect sizes between some of the ASVAB tests and the national testing programs (e.g., NAEP and SAT). Referring to the small effect size in favor of *males* on the Paragraph Comprehension (PC), WK, and Verbal Expression (VE) tests versus the small effect size in favor of *females* on the NAEP, s/he suggested that the military may be recruiting a less competent group of females than exists in the NAEP population. Dr. Manley responded that the group may include many who think that college is not for them and decide to join the military.

The committee member then identified two papers that have criticized Aguinis and colleagues: one by Mattern & Patterson (2013)[2] and one by Berry & Zhao (2015)[3] [formerly] at Texas A&M, who used a mathematical basis for the argument. Dr. Manley said that Aguinis previously advocated the use of Moderated Multiple Regression (MMR) analysis to detect test bias (Cleary Model), but later published papers stating that the MMR approach may not have adequate power in most cases to detect slope bias, as it is an interaction term that requires more power. Thus, absence of a significant interaction term in the MMR does not necessarily imply a lack of slope bias. The committee member said that Mattern & Patterson (2013) conducted their research with "college admissions data" and that Berry & Zhao (2015) used a mathematical approach. He said the second (mathematical approach) deserves more attention. Dr. Manley responded that, in DPAC's analyses, test bias has never been found, but adverse impact has. He said that if test scores do show adverse impact in selection, they need to be related to the criterion and that the criterion should be representative of the performance to be predicted. He also said that no other tests with equal validity but less impact could be used. He concluded by saying that DPAC will watch closely for the occurrence of adverse impact in the next generation ASVAB.

When Dr. Manley described the difficulty of obtaining criterion data – that the available data are not standardized across Services – a committee member underscored the need for a standardized criterion data collection, which s/he said has been a topic of prior DACMPT meetings. Mr. Arendt responded that the issue across Services is, in part, the difference in how they measure outcomes. He said, however, that the existing criterion data are useful when looking at the ASVAB data by Service. Dr. Kirkendall reiterated that ARI is working on a project to identify measures that are collected across Services. The committee member then raised the possibility that criterion data, if biased, would constitute another barrier to conducting prediction studies. Dr. Manley agreed.

[2] Mattern, K., & Patterson, B. F. (2013). Test of slope and intercept bias in college admissions: A response to Aguinis, Culpepper, and Pierce (2010). Journal of Applied Psychology, 98, 134-147.
[3] Berry, C. M., & Zhao, P. (2015). Addressing criticisms of existing predictive bias research: Cognitive ability test scores still overpredict African Americans' job performance. Journal of Applied Psychology, 100, 162-179.

11. Device Evaluation (Tab N)

Dr. Tia Fechter, DPAC, presented the briefing.

Dr. Fechter began by explaining that the goal of the device evaluation study is to facilitate delivery device expansion of the ASVAB iCAT and PiCAT by evaluating examinee performance differences among electronic devices (e.g., tablets, smart phones). This 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" user interface that incorporates a "Responsive Design" approach, which automatically formats the test display to alternative devices. Dr. Fechter then cited a literature review (Buckland, Becker, & Wiley, 2018) that summarized studies addressing mode impact on device usability, item difficulty and score differences. Studies comparing effects of modern electronic devices are sparse. Most focus on device usability and are found in the unpublished literature. Considerations that may impact performance include (a) screen size (minimum of 9.5 inches), (b) participants' device fluency, (c) item types/features, (d) whether item content is visible at one time, (e) device capabilities (e.g., touch screen), and (f) the fact that test completion times are higher when using mobile devices. Simple text-based items tend to not perform differently across devices. However, consensus on what impacts performance has not been reached.

Dr. Fechter then turned to the potential courses of action when it comes to implementing the ASVAB on different platforms. The first would be to proceed with implementation with no additional research studies. This would significantly reduce evaluation cost efforts and cut the time to operationalize expansion by one year. However, it degrades confidence in test score interpretations, raises concerns about score comparability for career counseling and enlistment, and could hurt the perception of the ASVAB testing program if score and measurement invariances are not upheld. Further, it may give the false impression that research was already carried out, and evidence to the contrary may weaken the ASVAB program. Finally, it could degrade the quality of the testing experience.

An additional option would be to proceed with implementing operational device expansion for the ASVAB testing platform for CEP grade 10, where scores cannot be used for enlistment. The resulting data could be analyzed to determine the impact of device expansion on score comparability for enlistment and classification purposes in a post-hoc manner. This would provide greater flexibility for the CEP to include students, could reduce the need to conduct an evaluation study within the MEPS, and may reduce evaluation costs. However, it would not be a controlled experimental design, and CEP grade 10 student outcomes are unlikely to generalize to the applicant population. Further, the outcomes of the evaluation may show that the CEP scores obtained lack measurement invariance and, like option 1, may give the false impression that research was carried out. In addition, it is possible that attempting to test on various devices could led to numerous failed test attempts.

The third possible course of action is to conduct an evaluation of performance differences observed across various devices for select ASVAB subtests before implementing any operational device expansion plans. This allows for a controlled experimental design and for the evaluation to be carried out with the most representative sample, namely applicants. In addition, this option allows for the evaluation of device familiarity as well as measurement invariance issues across devices and operating systems and provides the opportunity to obtain feedback on the responsive interface design. The major downside to this option is it increases the time to operational implementation.

DPAC recommends proceeding with the device evaluation and exploring the findings before operational implementation of alternative electronic devices in the enlistment testing and career exploration programs. Concurrent with the device evaluation, efforts to adapt the ASVAB testing platform and interface to be compatible with various web browsers can be undertaken. Dr. Fechter continued by indicating that the DAC role in this process will be to evaluate the recommendation to proceed with the device evaluation study and provide input on the appropriateness of the current design. In addition, the DAC can assist in mitigating technical challenges associated with the study and provide additional ideas for shortening the duration of the evaluation efforts.

The questions to be answered through the study are as follows:

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

Dr. Fechter then turned to the sampling design for the study. The plan calls for applicants to be tested at 10, low-volume Military Entrance Processing Stations (MEPS). Dr. Fechter presented a table showing eight examinee groups differing by the form and subtest from which items are taken. In all cases, the testing time will be approximately 30 minutes, with the number of items delivered varying by subtest (e.g., 12 AR, 24 MC). The total number of desired subjects in each group is 1,750, yielding an overall sample of 14,000. Dr. Fechter acknowledged that there are a number of challenges to be overcome, including (a) gaining access to participants, (b) issues of representativeness given the limited number of MEPS, (c) whether results will generalize to the CEP and subtests not included in the study (i.e., AI, EI, SI), (d) questions regarding the motivation of test takers who are in the midst of the application procedure, and (e) time constraints.

Seven devices will be selected for the study based on Graphic User Interface usability evaluations. These include one Windows-based PC (control condition), two laptops, 2 tablets, and 2 smart phones. Each participant will take two test forms, one on each of two randomly assigned devices. The forms are parallel and consist of ASVAB items from a selected number of subtests. The application will allow for two different logins for each participant, one per device, so each participant serves as his/her own control. Each participant will be randomly assigned to a condition code 01-16 which allows for counterbalanced administration of forms. Each will also have two access codes that contain information about their group assignment, device used, and form assignment.

Eight pairs of forms were developed for each ASVAB subtest to be parallel to one another. They contain the same number of items that a paper-and-pencil (P&P) form would have and they adhere to the table of specifications for the P&P forms. Items with special features, such as extended text length that may result in difference on item difficulty when administered on devices will be oversampled. Eight pairs of forms were developed that contain some items from a selection of ASVAB subtest pairs. Items were selected based on the proportion of items from each subtest that contain items with special features. These include some text-only items as controls. Dr. Fechter then showed a table that displayed the special features and the relevant subtests. This was followed by a series of charts that displayed the test characteristic curves for various subtest item pairs.

The research plan calls for delivering Forms 11 and 12 first, because they contain WK and AO items. The hypothesis is that device delivery mode should have no impact on WK performance. If such differences are found, there would be little incentive to move forward with the study. On the other hand, an additional hypothesis is that performance on AO will be affected by device delivery mode given the graphic-based nature of the test and results from prior research. If no differences are found on AO performance related to the device used, it provides some confidence in moving forward with operationalizing the device expansion.

Feedback will be gathered from subjects electronically after use of each device. Dr. Fechter showed a question that assesses the test takers motivation when taking the test. That was followed by another question that will be administered after the second device asking for subjects' perceptions of whether their performance was affected by the device used. Other questions to be asked include whether it was easier to use one device over the other, how comfortable the subject is using a tablet to take tests, and their familiarity with various devices.

Dr. Fechter then turned to next steps to be taken in the evaluation and their associated considerations. Work has been initiated on developing training materials for test administrators (TAs). A procedure is being developed for collecting appropriate information for ASVAB score matching. A factor that needs to be

taken into account is that the results for newer devices may not generalize for those subjects who use older devices that are more commonly available to applicants and students from lower socioeconomic groups. Another issue is that pulling one or two participants at a time may not be an efficient use of TA time.

Dr. Fechter then described the planned analyses of the data collected through the study. Item level comparisons will be carried out, including item difficulty, item information, the area between the item characteristic curves, response time, and differential item functioning (DIF) for applicant subgroups. Score level comparisons will include the differences within participant and between devices, factor analysis to determine the measurement invariance between devices, ASVAB subtest score correlations, device familiarity, and motivation. Feature-level analyses will involve statistical models to estimate the impact of systematic difficulty differences due to item features for groups of items. Challenges include ensuring participants accurately report their level of motivation and estimating score differences using the minimal items administered. The latter relies on the accuracy of the assumption that items without special features will perform similarly across all devices and that participants are equally motivated during the test as they were when taking the ASVAB to obtain a score of record. Finally, DIF analyses may not be feasible depending on the sample sizes.

As Dr. Fechter described existing research findings (slide 4), a committee member asked how items with extensive content (e.g., paragraph comprehension items with long paragraphs in addition to test questions) would be handled on mobile devices with smaller screen sizes. Dr. Fechter said that the research design includes looking at those types of items. Another committee member asked whether the existing literature addresses the impact of font size. Dr. Fechter responded that it probably has some impact, but that the literature review did not specifically address that question. She also said that mobile devices typically provide a zoom capability and so the team was trying not to be overly restrictive in that area.

A committee member agreed with Dr. Fechter that DPAC should not proceed with course of action (COA) 1 (slide 5), which was to begin operational device expansion for ASVAB testing without additional research.

When Dr. Fechter reported that some schools only want to administer the CEP if it can be administered on tablets, Mr. Arendt said that the issue with that approach is that the current test is designed to be administered on a device with a mouse, and that it does not support the use of touch screens. He said this eliminated the use of Apple products. He continued, however, saying that there is a desire to pilot the test to compare scores obtained with the use of alternate devices. Dr. Segall replied that COA 3 recommends conducting a study to see what alternatives are possible. He said that one option is to administer the CEP on alternate devices, while not counting those scores as operational. However, he also said that recruiters viewed this approach as a "no-go." Mr. Arendt inquired about the possibility of implementing the approach with students in grade 10, but Dr. Segall said that was not an option.

A committee member asked how the test interface would work on tablets and smart phones. Dr. Segall said that one option in that area is to write the test on applications that are self-contained for each device, but he explained that such an approach would be a significant undertaking because there are numerous types of devices, including Apple, Android, and others. Dr. Segall then said that DPAC is considering using Internet browsers, allowing the test to be delivered via the Internet. He said, however, that the issues with that approach are the same as those they currently face when administering the *i*CAT, primarily bandwidth. He said that there are currently bandwidth issues with the P*i*CAT as well, but that he is hoping that moving to the

Cloud will resolve or reduce those issues. The committee member then asked about the possibility of using an app. Dr. Segall said that some agencies currently offer apps, but that pursuing that path would require an entirely different level of effort. Mr. Arendt replied that DPAC is looking at streamlining the mode of delivery so that content can be consumed at a more rapid pace, making connection speed a non-issue. Dr. Pommerich explained that this approach involved implementing a thick client architecture. Mr. Arendt cited Amazon as an example. Dr. Velgach then asked if the selection of a browser would be a concern. Dr. Segall replied that the specific browser used should not make a difference in test delivery. Dr. Salyer replied, however, that based on what she has seen, text will appear in different fonts on different browsers. She cited Safari as an example of a browser that sometimes presents text in a "weird" font.

As Dr. Fechter described COA 3 (slide 7), a committee member expressed concern with conducting such an evaluation in the context of an operational testing program, citing potential adverse effects on candidate performance. Dr. Fechter replied that the proposed evaluation would not impact ASVAB scores, because the assessment would be implemented with just a few items at a time on separate devices, and that the evaluation itself would not be part of operational testing. Dr. Velgach asked if DPAC would be testing the use of smart phones. Dr. Fechter said that they would. The committee member replied that smart phones were the platforms that s/he was most worried about. Dr. Fechter said DPAC had the same concern and their approach would be to determine a cutoff for the types of devices that could be used effectively.

As Dr. Fechter briefed the evaluation design sampling plan (slide 12), Mr. Arendt asked if DPAC would be testing devices with different operating systems and browsers. Dr. Fechter said that they would, and that device selection would be based on a user interface study. CDR Phillips commented that human factors considerations would make that selection competitive, but he recommended dropping Windows Phone.

A committee member then noted that there are two aspects to the study: examinee performance and item functioning. S/he then asked if the items included for each subtest were representative of all the item types on those subtests. Dr. Fechter referred to slide 16, saying that forms would be parallel and constructed according to the P&P form specifications. She said that item selection would oversample items that have special features, including graphics. She then pointed to a list of special features to be sampled (slide 17) and emphasized that these included items with features that would require the "redisplay" of graphics due to the size of phone screens. The committee member asked if Dr. Fechter believed that presentation on different devices would change the difficulty level of some items. Dr. Fechter said that had been one of Dr. Embretson's concerns, and it will be studied as part of the research. She explained that recalibration might be device dependent. The committee member then asked if reconfigured items would be presented consistently across the various devices. Dr. Fechter replied that reconfiguring items would present a challenge in that they may be configured differently across devices; she then said that other types of items could be presented the same way across devices. Mr. Arendt then reminded discussants that the study would provide answers to many of these questions.

The discussion then moved to item calibration, as a committee member asked if DPAC would be estimating performance by subtest – the answer was yes – and calibrating items or persons. Dr. Fechter said they would be conducting simple analyses of variance (ANOVAs) comparing *p*

values and comparing equated scores across forms for subtests. When the committee member noted that the study would not include as many items as are included on the ASVAB subtests, Dr. Fechter replied that DPAC would like to use the subjects' ASVAB scores of record to impute responses for items that are not on the experimental form. The committee member clarified that the ASVAB scores could be the starting values for estimating thetas on the experimental form. Another committee member said s/he liked the approach of using real ASVAB test-takers for that reason.

Dr. Fechter continued by saying that DPAC would like to correlate performance in the two conditions, with each condition being defined by the device used. A committee member said that some subtests have fewer items, so it would be difficult to have a powerful test. Another committee member reported being worried that results might be skewed due to the inclusion of more difficult items, which would represent a confounding factor. S/he said that if test-takers only take more difficult items, their scores would be different than if they took a representative sample of items across the range of difficulty. When another committee member said that the items used in the experimental conditions should cover a range of the difficulty distribution, Dr. Fechter said that they are planning to do that. Dr. Pommerich clarified that they also plan to include "normal" items, that is, those for which there should not be a large impact based on reconfiguration. Dr. Fechter further clarified that if there is a meaningful effect by device on tests of most interest (e.g., WK and AO), then they should not move forward with the use of alternate devices. Dr. Segall said, referring to slide 25, that they would examine the tests that are most likely to identify issues before studying items that are expected to show less impact. Mr. Arendt agreed with taking that approach. Dr. Fechter pointed out that the study design presents a second opportunity to examine effects on AO items. Mr. Arendt said that PC items would also be examined twice, and Dr. Fechter replied that PC items are unique because of the scrolling requirement. She said they wanted to have two opportunities to look at that type of item.

A committee member asked how well the study would generalize to future devices. Dr. Fechter replied that they were planning to use the newest devices available to reduce the extent to which study results would be quickly outdated. Dr. Segall added that they were aware this was an issue. The committee member then cited a recent study that had demonstrated substantial differences in scores by device type on a writing test; s/he also noted that familiarity with new devices may also have been an issue. S/he went on to say that, in general, there has been some decline in scores on the writing components of some national tests but said that it might have been due to changes in the population. S/he said that the prompts in these tests are very rich and require substantial space; s/he said that, when added to the writing space, the total space required is significant. Dr. Fechter asked if the keypads for these tests were on the display, or if desktop keyboards were provided. The committee member replied that the tests used keyboards and not display keypads. Another committee member pointed out that there is no writing requirement on the ASVAB, explaining that all the items are multiple choice with the same number of options.

Dr. Fechter described how, on a tablet, a touch response could be used to select a response, which would allow the response identifier (i.e., A, B, C, and D) to be removed, allowing more response content to be seen at the same time. A committee member suggested this design might affect the response process, and Dr. Segall added that laptop administrations currently require users to click on the response identifier as opposed to anywhere on the response.

A committee member asked if a device constant could be developed to account for score differences across devices; s/he also asked about the procedure for integrating future evolutions in devices. When Dr. Fechter responded that they had not answered those questions yet, the committee member said the task of accommodating new devices could be overwhelming. To this point, Dr. Segall replied that some studies have found less of an effect for the device than for familiarity of the device. He said that, if this is the case, then AP could make recommendations that schools allow students to use the devices with which they are most familiar. He added that it would be a huge burden if they had to re-equate, however. Mr. Arendt said that NAEP has taken that approach and asked if there was someone with a technical background that DPAC could consult to better inform the study. A committee member replied that Partnership for Assessment of Readiness for College and Careers (PARCC) had done work in that area, and Dr. Fechter said that her team had reviewed that research. Another committee member suggested an ACT-related report that had been written by Laurie Davis in early 2018 that may not have been included in the literature review. Dr. Fechter said her team's literature review had been completed in January 2018, and so it was not likely included. The committee member said the results had been published later in 2018 and that he would coordinate with a contact to help relay the information to DPAC. After a brief exchange on the need to accommodate the timeline of DPAC's current study, Dr. Segall said that they had been facing this issue for decades, starting with the transition to computers.

As Dr. Fechter presented the questions to be used to collect post-test feedback, a committee member asked if it would be difficult for participants to respond to the following item: "I am comfortable using a tablet to take tests?" Dr. Fechter replied that it might be theoretical if they had not used a tablet in the past and if it was not one of their assigned conditions in the study. Another committee member said s/he thought the questions would be asked based in accordance with the devices each participant used in the study. Dr. Fechter said that, generally, they will be, but that the specific question cited is a more general type of item that will be given to everyone. She said that if she had to answer the question, she would disagree.

On the item dealing with device familiarity (slide 30), a committee member asked if the item could be modified to make it more specific; that is, by using more specific language than "on a regular basis." Dr. Fechter asked if the committee had any suggestions, and the committee member responded that s/he was not sure what timeframe would provide the best information. After surveying the devices listed, Mr. Arendt commented that he would like to see "Desktop PC" be the device that was the least familiar. A committee member said that it would be useful to know if people use multiple devices, and Dr. Fechter agreed. A committee member then asked if it would be useful to get information on many types of devices or just those that can be used in the testing setting. S/he cited devices that can be connected to a television that allow connections to the Internet and that can be used to write papers via voice recognition. Mr. Arendt recommended adding a box for free input to identify additional devices. LTC Rea responded by asking if the ultimate objective is to allow test-takers to use their own devices. He cited, as an example, his unit's use of cell phones to complete the GRIT assessment and said they ran into few difficulties. He also noted, however, the need to disable notifications that would be distractions to a test-taker. Dr. Fechter replied that distractions can be an issue. She explained that DPAC sees higher test completion times with the P*i*CAT than the proctored ASVAB. LTC Rea asked if the P*i*CAT is used to collect operational scores, and Mr. Arendt said that it is. A

committee member mentioned that s/he has seen sophisticated testing programs that can lock down the device to eliminate interruptions.

To conclude the discussion, a committee member asked how the committee can help going forward. S/he listed several possible ways they could contribute, including suggesting ways to reduce the scope of the evaluation and providing input on decision-making methods and criteria. Another committee member asserted that there is more information available than what had been included in the literature review. S/he suggested that Laurie Davis's work might provide a rationale for eliminating cell phones from consideration. S/he said that research suggested a very low probability of being able to display the required content in a usable manner. S/he also said, however, that s/he liked the overall design of the study, especially the examination of the most potentially problematic tests (i.e., PC and AO) first. Another committee member suggested that DPAC create a plan for dealing with truly major changes that come along. Mr. Arendt called that type of change "revolutionary" as opposed to "evolutionary." Another committee member again expressed concern over the reduced length of the experimental tests, as well as the oversampling of selected item types, and suggested that the current study focus on determining the types of devices that demonstrate potential and then looking more closely at score comparability and equating in a separate study.

12. <u>WK Automated Item Generation</u> (Tab O)

Dr. Isaac Bejar, ETS, presented the briefing.

Dr. Bejar began the briefing by explaining that the goal of the project is to automate the production of 4-option ASVAB Word Knowledge (WK) items of both the definitional and contextual type. The expectation is that the generated items will have fairly well-estimated difficulty levels and be above a certain level of discrimination. The development process involves conceptualizing the approach, designing the system, conducting field tests of generated items, and performing CAT simulations. Dr. Bejar then presented a list of the tasks involved, indicating that all had been completed except for system packaging.

Automated Item Generation (AIG) enhances both efficiency and validity. It allows for many items to be produced with enough of an understanding of their difficulty levels to reduce the need for pre-testing. Validity is enhanced because the construct representation is grounded in relevant science, with difficulty a function of word familiarity and depth of word familiarity. In addition, having many more items to work with improves test security. Dr. Bejar then presented a graph that summarized the approach taken. For the field test, 1,000 items were generated, with 60% accepted by SMEs following review. There were 10 items with negative biserials. Difficulty prediction was based on an r-squared of 0.26 using gradient boosting machines (GBM). Overall, 75% of items had discrimination values above .80.

There is a long history of relying on linear regression in AIG, but in this instance GBM was used. Dr. Behar showed a graphic representing the application of GBM, before turning to improved features of version 2 of the WK AIG. These include predictors that incorporate word difficulty (familiarity), inter-word level (depth of word familiarity), and variable importance. Cross-validation considered the word level, the inter-word level, and the variable importance. Prior to the field test, version 1 of the AIG yielded an R-square of .26. This increased in version 2 to .34. Dr. Bejar then displayed a table listing the difficulty predictors by relative importance, showing that corpus-based work familiarity was the strongest predictor of difficulty. Dr. Bejar then presented additional data demonstrating the improvement of prediction from version 1 to version 2, and the additional gains that could be achieved by incorporating SME input.

Dr. Bejar continued by presenting several graphs showing the results of CAT simulations. He concluded that, for estimating ability, an approach to compensate for imprecise parameter estimates is to lengthen the

> pool, but not the test. When assembling 75-item pools from the degraded 339 field tested items, the imprecise parameters were not taken into account, although there is a methodology for that purpose. For WK, it seems plausible to avoid or greatly reduce pretesting. However, items with negative biserials need to be avoided or removed.
>
> Dr. Bejar then turned to the generation of contextual items, which was on hold to finish the definitional item system. The approach taken is to use the key and distractors from a definitional item. Templates inspired by successful WK items from Forms 7 and 3 were created. The item stem is first placed in the template, with the remainder of the template filled in by fitting n-grams. Although the resulting sentences are grammatically correct, often they do not make sense. It is expected that the yield from this system will be low, but the items generated will perform well.

As Dr. Bejar described the field test results (slide 8), he responded to a question about the meaning of the negative biserials, explaining that they indicated discrimination in the negative direction. Shortly thereafter, another committee member asked about the distinction between "familiarity" and "depth of familiarity" introduced on slide 12. Dr. Bejar said that familiarity is based on how often a word occurs, whereas depth of familiarity captures the level of reasoning that is required to answer the item. He clarified further, explaining that familiarity was determined by how common a word is in the corpus, or word frequency, and that depth of familiarity is the degree of relatedness between the stem word and the response options. A committee member related depth of familiarity to a sort of "within item" familiarity, and Dr. Bejar agreed.

When a committee member asked if the items developed by word generation performed more on the basis of depth of familiarity and SME-developed items performed more on the basis of stem familiarity, Dr. Bejar again agreed. The committee member then asked if the generated items tended to be more difficult than the SME-developed items. Dr. Bejar, replied that they were, in fact, easier. He said that this would be a great finding if it stands up to review; that is, it would indicate that item generation is a more efficient way to create items of various item difficulties.

Dr. Segall asked Dr. Bejar about how the discrimination parameter was specified in the CAT simulations. Dr. Bejar responded that the discrimination parameter was sampled from a range of values and that this approach led to higher score information than when the discrimination parameter was set to some constant. Dr. Segall said he did not think this approach (using randomly generated discrimination parameters) provided an accurate representation of the precision of the AIG item approach (as opposed to using a constant discrimination parameter in the simulations). Dr. Segall said he would follow up with Dr. Bejar to discuss in more detail.

As Dr. Bejar described the deliverables, he asked Dr. Segall if DPAC uses Windows, explaining that the item generation program is Java-based and runs on a Windows system. Dr. Segall said that they do use Windows, and Dr. Bejar replied that, in any case, there is another year of support left on the contract.

The discussion concluded with a committee member's request for clarification of the scatterplots shown on slides 24 and 25. Dr. Bejar explained that the color of the dots indicated parts of speech. The committee member then asked if each dot was an item, and Dr. Bejar said that was true. He added that difficulty was presented for definitional and contextual items. He also explained that the plots showed a strong relationship between discrimination and difficulty of the original definitional items and the subsequently developed contextual items. The committee member asked if, when

generating context, it was hard to find an instance of where the contextual statement is found in the population of books that were searched, and if there was a wide range of sources examined. Dr. Bejar replied that the sample of sources used was a well-known collection of copyrighted books. The committee member then asked if test-takers might have read some of the books; s/he also noted the availability of other sources, such as magazines. Dr. Bejar clarified that the list of sources only included books that had lost their licensing rights and could thus be used in this type of work. Hearing this, the committee member suggested that range restriction might have resulted in an underestimate of the occurrence of contextual statements. Dr. Segall then asked if the "book checking" was conducted only to determine how often the contextual statements occur in order to assess the continued requirement for SME review, as opposed to helping determine item difficulty. Dr. Bejar said that was correct. When a committee member asked if all the items had been reviewed by SMEs and field tested, Dr. Bejar said they had. He also noted that only one contextual statement had been found in the sources examined, though he said a more sophisticated search that employed synonyms might have produced different results. He said that he had conducted the search using the stem word, which was the less frequently used word in comparison to the words used in the response options. The committee member closed the discussion by saying that it was nice to see how this work had progressed.

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

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

Dr. Salyer began presenting ASVAB CEP numbers and metrics for school years 2012-2018. These showed the number of students tested ranging from 670,836 in 2013 to 713,777 in 2018. The percentage of schools tested has ranged from 55% in 2018 to 57.2% in 2016. Additional data showed that students taking the paper-and-pencil ASVAB decreased from 670,886 in school year 2016-1017 to 662,564 in school year 2017-2018. At the same time, the number of students taking the computer-based iCAT rose from 14,011 to 51,213. Regarding recruiting, 29,017 students used their CEP ASVAB scores to enter the military in 2018, and 433,317 leads were developed through the program.

Dr. Salyer continued by showing a screen shot from the Texas Education Agency announcing that the ASVAB will be made available to all Texas public school students in grades 10 through 12. A table showed the increase in participation rates in Texas as a result of this legislation. An additional table showed the states and MEPS that have engaged with the program following the passage of the Every Student Succeeds Act (ESSA).

Dr. Salyer then turned to ongoing program initiatives. An Expert Panel was convened in 2017 to examine program components and make recommendations for improvement. They are conducting analyses of Find Your Interests (FYI) data to identify the factor structure of the instrument and any gender differences that are still present in the population. An expert panel member is also reviewing state initiatives and legislation that may be relevant to the CEP. A final effort recommended by the panel is to provide additional training to Educational Service Specialists (ESS) who market the program, with the goal of offering a certification from the National Career Development Association. Training development and certification efforts are underway.

A needs assessment is being conducted, including MEPS visits and school observations of testing sessions and posttest workshops. The goal is to identify best business practices, efficiencies that can be realized, and develop a model of program delivery. At the same time, Caveon Test Security, LLC, has been reviewing web patrols to determine if the degree to which ASVAB forms currently in use in the CEP may be compromised. Dr. Salyer then detailed some updates that have been carried out the CEP iCAT to increase its efficiency and reach.

Dr. Salyer then displayed a chart showing CEP website utilization data from 2016-2017 and 2017-2018. This showed significant increases in unique and returning visitors, page views, and tablet/mobile visitors, with a decrease in the bounce rates. Corresponding figures for the Careers in the Military (CITM) website showed significant increases in returning visitors, page views, average time per session, and number of pages viewed per session, with a decrease in bounce rates. Data on access code utilization from July 1, 2017 to June 30, 2018 indicate that website utilization and access code utilization show great return on investment for conferences and marketing efforts.

Dr. Salyer then presented data on the number of users who selected the Contact Us option on the CEP website, which totaled 3.533. She noted that this represents the amount of work that requires manual labor on the part of DPAC and MEPCOM personnel. Similar data for the CITM website showed that 353 students contacted one of the military branches for additional information through the contact us option on that site, also requiring personnel time to respond. Dr. Salyer then showed a sample of inquiries received through the CITM site.

Turning to new features of the program, Dr. Salyer indicated that students now have the ability to make notes about colleges viewed and save those they are interested in. College details provided include acceptance and retention rates, average test scores, and ROTC programs offered by Service. Students can also merge accounts to include their portfolios, FYI results, saved occupations and notes, and favorites. Improvements have also been implemented in the ASVAB Summary Results, and Service line scores are now available so students can have meaningful conversations with recruiters about occupational options.

Dr. Salyer then presented navigation data for the CITM website showing pages viewed, search options employed, and additional analytics. A chart showed the data that are now available for each of the Services thanks to input from Service representatives. Regarding communication efforts, Dr. Salyer explained that a teacher engagement campaign has been launched in which individual teachers are sent information about the CEP and the website and encouraged to take advantage of classroom activities provided to assist students in exploring their future options. Additional communication efforts include attendance at relevant national conferences, which Dr. Salyer listed. She also provided a list of CEP marketing efforts through various state and association websites and print materials. An additional communication effort involves creating and managing Facebook groups for each of the 65 MEPS to share local and region-specific information regarding the ASVAB CEP.

An additional initiative involves automating score reports so that DPAC personnel can instantaneously generate a score report using minimal information. Upon request, DPAC can now provide states and schools with reports about their populations' ASVAB scores and FYI results. Another feature allows authorized users to generate website access codes.

Dr. Salyer then addressed an effort to maximize career exploration by improving the ASVAB CEP cross-links between military and civilian occupations. The purpose of this project is to update the links using a task-analytic approach and predictive analytic software to identify potential matches. OPA is collaborating with other key stakeholders involved in similar efforts, including the Department of Labor and the Transitioning Veterans Program.

Dr. Salyer then turned to a new program initiative called UNIFORM, the goal of which is to develop a web-based application to house all Service-provided occupational information and streamline data collection, manipulation, and distribution in a unified format. This allows for the seamless production of a comprehensive representation of military careers accessible to all government and civilian entities. Currently, each Service submits unique data for over 8,000 active military jobs, and DMDC manually performs analyses to code them based on commonality of skills, duties, and training. This process inhibits the timely analysis and dissemination of military occupational information for career planning. The UNIFORM application is currently populated with data provided by DMDC, and the reporting functionality is under development. The team is working with service IT representatives to facilitate automation.

As Dr. Salyer briefed the accessions by Service (slide 4), Mr. Arendt explained that the numbers of students using their ASVAB CEP score for enlistment were not proportional to the total number of enlistees by Service. A committee member asked if any one Service was more likely than the others to benefit from the CEP, and Mr. Arendt responded that the rate of CEP enlistees for the Army is roughly double that of the other Services. He added that the AF and Navy are underrepresented in CEP enlistees and said that the Coast Guard is the least represented.

Regarding how States use the CEP, Dr. Salyer cited the Every Student Success Act (ESSA), which requires states to set college- and career ready standards. Texas legislature's Senate Bill 1843, uses ASVAB CEP as method to satisfy this requirement. They require each school district and open-enrollment charter school to provide the opportunity for students in grades 10 – 12 to take the ASVAB and consult with military recruiters. Mr. Arendt said that the educational specialists cite this law when promoting the CEP.

As Dr. Salyer described the expert panel's work (slide 10), a committee member asked for clarification on the bullet dealing with "Training for ESS Community." This item deals with a recommendation to develop centralized training and certification (nationally recognized career development credential) for ESSs. Dr. Salyer said that the government would be providing the training so that the specialists' organizations would not have to pay out of pocket.

On the topic of test security (slide 12), Dr. Velgach asked if the contractor was monitoring only for CEP item compromise. Dr. Salyer said that was correct, but that there will be a discussion about whether the search should be broadened to include items in the other ASVAB pools.

When Dr. Salyer said that students were taking the time to find out more about the military via the inquiry feature of the Careers in the Military (CITM) website, a committee member asked if she knew if and how quickly the Services were responding to the inquiries. Dr. Salyer replied that they are responding, and in a timely fashion. Mr. Arendt added that these inquiries are likely to result in recruiting leads, and so the Services are likely to respond quickly.

Dr. Velgach asked how Dr. Salyer updated the Service Line (SL) scores. Dr. Salyer said that she obtains information from classification, recruiting, and the MAPWG to ensure that the information provided on the CITM website is accurate.

As Dr. Salyer presented the national events in which her program has been able to participate (slide 35), she thanked the committee and AP for their efforts to helping to make it happen. A committee member asked Dr. Salyer how many people were in her program, and Dr. Salyer said it is really just her. The committee member said s/he thought Dr. Salyer was doing a remarkable job and asked what she could accomplish if she had another staff member. Dr. Salyer then said that she has a group of six contractors, but that she is required to do site visits. She also said that Dr. Segall has been amazing in assisting her efforts. She said it would be great to have another person, but that she makes it work. Mr. Arendt mentioned that she has "an army of people behind her," which are the MEPCOM education service specialists (ESSs). Dr. Salyer then said that she thinks the ESSs could do a more "professional" job in schools with the correct training. She added that the credentialing program should help in this regard. She also said that she provides videos, but that virtual training should be coming online this year.

To continue the work in marketing the program, Dr. Salyer described advances using social media, specifically, the new features available in Snapchat to promote the CEP. She said this will provide students the opportunity to use pre-defined filters and elements to promote the program.

As Dr. Salyer discussed the Military to Civilian Crosswalk (slide 39) and the UNIFORM website (slides 40-42), Mr. Arendt said that employers are asking how they can figure out what people in the military can do. He said the employers do not like O*NET. Dr. Salyer said they probably avoid O*NET because it is hard to navigate. She also said that the service credentialing people are trying to pull that type of information together from the numerous databases in which it currently exists. She said they have quarterly meetings to identify the variables that users will require.

A committee member then described how some States are interested in ESSA (Every Student Succeeds Act) allowances. S/he cited one State that is interested but said that it does not have a system to track the performance information that ESSA requires. Linking this back to the CEP, s/he stated a concern that the States would try to harvest CEP scores to use as a school quality indicator and suggested that Dr. Salyer keep an eye out for the inappropriate use of the CEP. Mr. Arendt replied that OPA is the guardian for inappropriate use of the CEP and that they have been fighting off States, much more than local municipalities, by providing essential data in ways that make it very difficult from them to use for alternate purposes (e.g., to generate school quality metrics). Dr. Salyer said that she cannot pull the data easily. She also said that they had sent a letter to encourage States to use the program for career exploration and to check with the Services about the minimum requirements for enlistment to better understand the meaning of the scores. She said that her program provides military information that other programs do not, and that they do it for free. She also said that her program sends scores to schools, who can then send them to the State if they want, but that she did not endorse that. She said that, in any event, it would be hard for the States to use the data that are provided. The committee member then said that States do strange things, citing Indiana and Pennsylvania. Mr. Arendt responded that the downloads that are provided to the schools make the jobs of career counselors much easier. Dr. Salyer concluded the discussion by saying that the concern about the inappropriate use of data is on a list of topics to be addressed.

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, iCAT 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 (Arithmetic Reasoning, Math Knowledge, and other tests)
- Web and Cloud efforts

Upon reviewing the proposed list of future topics, a committee member asked for updates on the P*i*CAT, VTEST, and APT. S/he also asked for an update on the TAPAS panel if sufficient information would be available by the next DACMPT meeting. Mr. Arendt replied that OPA could do an IPR on any ongoing project. The committee member also mentioned that s/he would like to see the revised MCt demo and a report on the validity framework. Mr. Arendt asked if the committee would like to see the report on the validity framework when it is available, and the committee member replied that anything they could see in advance of the next meeting would be helpful. Additionally, a committee member requested an update on the device evaluation study.

A committee member asked about the scope of Caveon's test security work. Dr. Segall replied that, for a year or two, they were reviewing ASVAB items, but that their recent work has been limited to monitoring for the misuse of CEP items. He said they had also been involved with the language program and had visited testing sites to audit procedural compliance. Another committee member responded that a briefing on test security/compromise would be welcomed. Another committee member mentioned that a test security consultant s/he had dealt with in the past used to like to talk about what he had learned. Dr. Segall replied that Caveon had offered to audit the testing program from start to finish, including physical test security (e.g., buildings, computers, and flow of items through the development process). The committee member reported being worried about the move to the Cloud. Dr. Segall said that he did not know if Caveon had experience with cyber security. The committee member said s/he was limited in that respect as well. Dr. Segall responded that DoD is very attuned to cyber security, and he cited the example of how DPAC was getting a Level 4 security rating, which is sufficient for the use of PII data. He said that this is one of the reasons he is not too concerned about the move to the Cloud.

Dr. Velgach suggested that the Joint Advertising Market Research & Studies (JAMRS) should brief at the next meeting. Mr. Arendt said they can do that as part of the AP update. A committee member then asked if there would be enough information available to have another briefing on AVID. Dr. Kirkendall replied that ARI should have something by then. The committee member then asked if they could be briefed on the expert panel recommendations for the CEP, and Dr. Salyer replied that she should be ready for that. Another committee member asked if Dr. Bejar would know if his deliverables were working. Dr. Segall said that should be the case. He added that the Educational Testing Service (ETS) will be starting on General Science (GS) item automation, but that he is feeling less confident about the prospects of that effort. The committee member also said that s/he would like to hear more about test administration via the Cloud.

To close the meeting, the committee chair asked if there were any more comments. Hearing none, Mr. Arendt thanked everyone for their participation.

Tab A

LIST OF ATTENDEES

Defense Advisory Committee on Military Personnel Testing (DACMPT)
September 20-21, 2018, The Hyatt Place Downtown Minneapolis, MN

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	Senior Research Psychologist	Defense Personnel Assessment Center
Dr. Tia Fechter	Personnel Research Psychologist	Defense Personnel Assessment Center
CDR Hank Phillips	MILDEP for Research and Engineering	Naval Air Warfare Center
Dr. Donna Duellberg	Voluntary Education Program Manager	US Coast Guard
Dr. Mark Rose	Personnel Research Psychologist	US Air Force Research Laboratory
Dr. Cristina Kirkendall	Research Psychologist	Army Research Institute
Mr. Ken Schwartz	Air Force Enlistment Policy	Headquarters, Air Force Personnel Policy

Mr. Brad Tiegs	Testing Director	Headquarters, U.S. Military Entrance Processing Command
Mr. David Davis	Chief, Testing Division	U.S. Military Entrance Processing Command
Dr. Isaac Bejar	Psychometrician	Educational Testing Service
Dr. Cheryl Paulin	Vice President	Human Resources Research Organization
Dr. Arthur Thacker	Principle Scientist	Human Resources Research Organization
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
Dr. Sagar Ruby	Technical Architect	Accenture
Dr. Linda Aaker	Attorney at Law	University of Minnesota

Tab B

DEFENSE ADVISORY COMMITTEE ON MILITARY PERSONNEL TESTING
AGENDA

September 20-21, 2018
The Hyatt Place Downtown
Minneapolis, Minnesota

September 20, 2018

0800-0830	Complimentary Buffet Breakfast in Dining Room	
0830-0900	Executive Session	Dr. Michael Rodriguez, Chair
0900-0915	Welcome and Opening Remarks	Mr. Chris Arendt, OASD (M&RA)/AP*
0915-0945	Accession Policy Update	Mr. Chris Arendt, OASD (M&RA)/AP
0945-1015	ASVAB* Milestones and Project Matrix	Dr. Mary Pommerich, DPAC/OPA*
1015-1030	*Break*	
1030-1100	Next Generation ASVAB and ETP* Update	Dr. Mary Pommerich, DPAC/OPA
1100-1130	Validity Framework Update	Dr. Art Thacker, HumRRO*
1130-1215	Mental Counters	Dr. Ping Yin, HumRRO
1215-1315	Lunch	
1315-1345	CAT*-ASVAB Form 10 Equating Study	Mr. Matt Trippe, HumRRO
1345-1415	Development of New Cyber Test Items and Pools	Dr. Matt Trippe, HumRRO
1425-1500	Sparse Data Dimensionality Assessment with Application to the Cyber Test	Dr. Furong Gao, HumRRO
1500-1515	*Break*	
1515-1545	TAPAS Expert Panel Update	Dr. Tim McGonigle, HumRRO
1545-1630	Adverse Impact	Dr. Greg Manley, DPAC/OPA
1630-1730	Executive Session	Dr. Michael Rodriguez, Chair

September 21, 2018

0800-0830	Complimentary Buffet Breakfast in Dining Room (Prior to Meeting)	
0830-0900	Executive Session	Dr. Michael Rodriguez, Chair
0900-0945	Device Evaluation	Dr. Tia Fechter, DPAC/OPA
0945-1030	WK* Automated Item Generation	Dr. Isaac Bejar, ETS*
1030-1045	*Break*	
1045-1130	CEP* Update	Dr. Shannon Salyer, DPAC/OPA
1130-1145	Future Topics	Dr. Dan Segall, DPAC/OPA
1145-1200	Closing Comments	Dr. Michael Rodriguez, Chair
1200-1500	Committee Working Lunch	

*** KEY:**

APT = Armed Forces Qualification Test (AFQT) Predictor Test
ASVAB = Armed Services Vocational Aptitude Battery
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
ETP = Enlistment Testing Program
ETS = Educational Testing Service
HumRRO = Human Resources Research Organization
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
WK = Word Knowledge Test

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

December 23, 2018

Mr. Christopher Arendt
Deputy Director, Accession Policy
Pentagon, Washington DC, 20301

Mr. Arendt:

The Defense Advisory Committee on Personnel Testing (DACMPT) is pleased to provide this committee report of our recent meeting on September 20-21, 2018, in Minneapolis, MN. 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. Rodriguez (chair), Dr. Sofiya Velgach, and Mr. Arendt. Also, Drs. Neal Schmitt, Barbara Plake and Kevin Sweeney were in attendance. In addition, staff and representatives from DPAC and various military units.

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

Accession Policy Update

Mr. Arendt provided an update regarding the special reassignment of Ms. Stephanie Miller, Director of Access Policy, and his current role. Regarding the Accession Policy update, Mr. Arendt provided an organization chart for Accession Policy. He summarized the current recruiting success of each military unit, noting the continuing challenge facing the Army regarding mission attainment (at 90.7% through August 2018) – noting the positive employment rate as one source of the challenge. This also affects Reserve recruiting, which is locally based and tied to local employment rates.

ASVAB Milestones and Project Matrix

Dr. Pommerich briefed the DACMPT on the milestones and project schedules for the major ASVAB research and development efforts. Many of the projects were also included as DACMPT agenda items and were not discussed.

The automatic generation of AR and MK items topic garnered some discussion. Dr. Pommerich noted that tryout systems are currently being developed, as well as the systems to identify item enemies, as there are many item clones in the pools. In addition, the DACMPT asked about the

gaming approach mentioned in the Cyber Test topic. It was noted that the Air Force is developing performance-interactive tasks that employ a gaming format.

The AFQT Predictor Test work has continued. This will be an item for future briefings. The efforts to expand test availability through web delivery of special tests continues to progress, where service-specific special tests have priority. The deployment of iCAT and TAPAS to production in the cloud continues to require a significant amount of effort, but holds promise for the long-term accessibility and success of these projects.

The DACMPT expressed a concern regarding the volume of psychometric work being contracted to HumRRO. On one hand, the work needs to be done and HumRRO is fully capable. On the other hand, contracting out a substantial amount of work requires substantial monitoring.

Recommendation(s): The DACMPT encourages the DPAC to ensure comprehensive documentation of processes with quality metrics by HumRRO in all contracted work.

Next Generation ASVAB and ETP Program

Dr. Pommerich briefed the DACPMPT on plans for the next generation ASVAB including the special tests administered in the Enlistment Testing Program. Included in her briefing was information on the recommendations by an expert panel convened in 2005-2006. Most of these recommendations have been implemented into the ASVAB testing program. Of those not implemented, the inclusion of a non-verbal reasoning test is currently under consideration. New tests have been considered for inclusion on the ASVAB platform, including TAPAS, Cyber Test, Mental Counters, and a test of Abstract Reasoning. Some of the obstacles to bringing new tests onto the ASVAB platform include restriction in total time available for administration. In addition, there is an on-going deliberation about the "philosophy of the test". There is an effort underway to develop a Validity Framework which will provide some clarity on issues surrounding the philosophy of the ASVAB assessments. DPAC has also initiated an extensive plan to evaluate the current ASVAB tests in order to determine their desirability/expendability using a wide variety of criteria.

The DACMPT was supportive of these efforts and believes this is a promising start in addressing issues related to both the philosophy of the ASVAB and providing clear rationales for adding, continuing, or removing assessment from the ASVAB platform.

ASVAB Validity Argument

Dr. Thacker updated the DACMPT on progress HumRRO has made in the articulation of a validity argument framework for the ASVAB for both subtests that accrue to AFQT and the other tests that comprise the ASVAB platform. Dr. Thacker started with identification of the main uses of ASVAB scores: (a) admission into military branches and (b) placement into training programs or advanced educational opportunities. Dr. Thacker then clarified how a theory of action is related to the interpretative argument, which then leads to gathering evidence to support this interpretative argument. Through several examples of draft Theories of Action, more clarity was provided to the steps needed to complete a validity argument. In the discussion

around these draft theories of action, the Committee suggested that the final outcome in the theory should be "readiness for service", not just success in jobs/training. Dr. Thacker then provided several claims that could be investigated using a validity argument framework, including that AFQT measures "G", "G" is broadly predictive of performance, Candidates categorized based on AFQT are sorted according to likelihood of success in military occupations, and claims about Utility and Implementation factors. Each of these claims had sources of evidence identified to support the validity of these claims. This work represents a promising effort to clarifying the philosophy of the ASVAB.

Mental Counters

Dr. Ping Yan delivered a presentation on efforts to revise and streamline instructions for the Mental Counters test. Scores on this test have yielded a distribution that includes between 5 and 9 percent of examinees who receive scores of zero while the remainder of the scores appear normally distributed. It is thought that those receiving scores of zero do not understand the instructions. The revised instructions include step by step solutions to several of the item types included on the test. This is followed by several practice items and the provision to provide monitor help should an examinee not understand the instructions. Software will also require that the examinee complete at least one item correctly before proceeding to the test.

Recommendation(s): The DACMPT feels that these modifications will help to solve or eliminate this problem. They encourage a separate set of practice items be used if the first set is not answered correctly so that examinees cannot just provide the right answer they received in feedback to the first set of items. They also recommend that the test of the new instructions be proceeded by a "think aloud" session or cognitive lab with a small number of examinees to identify any remaining issues with the revised instructions.

CAT-ASVAB Form 10 Equating Study

Dr. Trippe presented the status of the CAT-ASVAB Form 10 equating study. Form 10 was developed from previous P&P forms, intended to be used in the CEP iCAT. To ensure interchangeability of scores with operational form scores, item parameters were calibrated and scaled through linear transformation (per subtest) to the scale of operational CAT-ASVAB forms 5-9. In addition, the form 10 theta scores will be equated to the theta scores on CAT-ASVAB form 4, the reference form. Three phases of equating involved operational administration of Form 10 to successively larger samples through random groups design, involving equal probability of being assigned Forms 4, 6, 5, 8, 9, or 10. Several analyses were presented evaluating the quality of the equating.

The DACMPT acknowledged the careful and comprehensive attention to the equating, including the additional step of equating the new form scores on the operational form scale after the transformation of item parameters. The DACMPT also noted the high level of quality in equating results regarding the AFQT, which is an important component of the CEP use of the test. In addition, the DACMPT was concerned about the use of equating error in understanding the differences between original and equated scores – of which Dr. Segall described as within the

amount of variance expected through the use of provisional scores in the midst of the three phases.

Recommendation(s): The recommendation was made to consider the estimated equating error at this point to be provisional as well.

Development of New Cyber Test Items and Pools

Dr. Trippe provided the briefing on Cyber Test (CT) item and form development. He noted the current effort to develop 200 new items targeted toward the middle and low end of the ability distribution. The DACMPT noted that a target test information function should drive efforts to develop items in ability ranges needed to achieve the target. It appears that the inherent complexity and technical nature of the content areas results in items that are too difficult, leading to a low survival rate of items (48% of the field test items were retained).

The high-end relatively narrow band of specialized skills being addressed in the CT create a mismatch with typical high school curricula. Such skills are unlikely to be developed in high schools.

Recommendation(s): One possible source of more general (less complex and technical) CT items could be found in the related yet more general content of the NAEP Technology and Engineering Literacy (TEL) assessment; however, the DACMPT acknowledges that such general items may not be informative of success in more technical jobs. These NAEP items may provide some information as to the kinds of tasks and item features that produce easier items. The question remains as to whether easier items are needed for the specific use of the CT at this time.

Sparse Data Dimensionality Assessment with application to the Cyber Test

Dr. Gao introduced the briefing on dimensionality assessment in the context of sparse data, noting that the DACMPT raised concerns about the potential impact on the CAT functioning as new items are developed and introduced into the item pool. The sparse data matrix is a challenge in the context of seeded trial items, as each candidate receives a small number of such items. Dr. Gao presented a bifactor model for the four content areas of the Cyber Test, including one general factor. The model was fit using iFACT, an IRT model-based assessment of dimensionality. The primary focus is on the explained common variance (ECV), an indicator of essential unidimensionality as the proportion of variance accounted for by the general factor. Based on the analysis of the CT forms 1 and, with 29 items each, the overall ECV was .87 and .90. When including the 117 seeded items, the ECV was .92. Dr. Gao concluded that the CT was essentially unidimensional, as it fit a bi-factor structure well.

The DACMPT noticed in the bi-factor graphs (slides 14 and 16), there appeared to be a downward trend in the associations between the secondary factor loadings and G – the items that load higher on the G factor load negatively on the second factor. This might be something worth investigating. It calls into question the nature of unidimensionality and the possibility that a different structure exists among items, perhaps as a function of item difficulty, rather than content areas.

TAPAS Expert Panel Update

A report by RAND that is critical of the research and potential contribution of the TAPAS measure developed by Dr. Drasgow and his colleagues and evaluated and used in some form by all services stimulated the request by DPAC that a panel of five distinguished and knowledgeable researchers be formed to evaluate research conducted on the TAPAS, make recommendations for future research and comment on the readiness of TAPAS for operational use. The DACMPT feels that the panel with research support from HUMRRO is competent to provide this evaluation and looks forward to the report. Examination of TAPAS incremental validity, test-retest reliability and its susceptibility to faking as well as other possible problems should inform future use of the TAPAS and the required research support.

Adverse Impact

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

Recommendation(s): The DACMPT recommends continued monitoring of subgroup scores for adverse impact and changes in mean subgroup scores.

DPAC Device Evaluation for ASVAB

Dr. Fechter provided an overview of a research effort to understand if it would be acceptable from a measurement and logistic perspective to expand the types of devices on which the ASVAB could be delivered. Currently the CAT ASVAB is delivered using a laptop computer. Other testing programs have considered and sometimes expanded the devices for their test delivery. DPAC plans to engage in a staged device delivery study, considering first those subtests that either are highly unlikely to be impacted by the use of a variety of devices (WK) and those that it is possible will have some interaction with performance, such as AO and PC. Devices to be considered in the experimental design include tablets, notebooks, and phones. Familiarity with these devices will also be gathered in the study.

Recommendation(s): The Committee recommended that the research by Laurie Davis and others also be considered as there has been quite a bit of attention to device comparability recently. These studies may provide information on whether all of the devices under consideration (especially phones) should be retained if there is evidence provided by these recent studies about their usability in assessment settings.

WK Automated Item Generation

Dr. Bejar, of Educational Testing Service, provided a briefing on the ASVAB word knowledge (WK) automated item generation project. He noted that several others at ETS have contributed to the effort. This is anticipated to be the final update on this project. The project has resulted in 3000 WK items; although the WK generator has been developed, it has not yet been evaluated or tested operationally in DPAC systems.

With respect to contextual items, the contextual generator has been on hold to complete the WK generator and other system improvements. The current challenge is that although sentences generated are grammatically correct, they do not always make sense – the yield is expected to be low. One possible limitation is the source of materials to identify the existence of such sentences and phrases (contexts) in authentic written materials. Currently, the analysis of WK contextual items is driven by a set of ebooks. Although more commonly available publications, such as newspapers or popular magazines may be more appropriate for the intended population, these are not easily compiled online for the purpose of the phrase finder system to validate the generated context sentences.

CEP Update

Dr. Salyer provided an update to the DACMPT on the ASVAB Career Exploration Program (CEP). DACMPT members were provided with a number of ASVAB CEP items, including pens, technology tools, and others, all bearing the ASVAB CEP name and logo. In addition, the DACMPT reviewed the 2018 Annual Report and Recruiter Guide. The CEP is on record pace regarding the numbers of students and schools tested. Online tools continue to be updated and expanded, linking across platforms and systems. The Expert Panel process was summarized – to be reported in future meetings. The DACMPT acknowledges the many benefits that will likely follow the establishment of the CP certificate (NCDA certification for education service specialists).

The DACMPT congratulates the ASVAB CEP staff and all members of the services that have supported the expansion of the components of the CEP. This is likely to become a key component of accession efforts in the future. The DACMPT looks forward to hearing about the Expert Panel results in the near future.

Future Topics

In addition to the topics that were planned for briefings at the next meeting, the DACMPT requested updates on PiCAT, the V-Test, the APT, and the CEP Expert Panel results. In addition,

the DACMPT requested an update on how the automatic item generation programs function, once implemented in the DPAC systems.

On behalf of Drs. Plake, Schmitt, Sweeney, and myself, we greatly appreciate the commitment and effort of the staff from DPAC and service personnel regarding the wide ranging projects to continually improve personnel testing and services. All staff present were willing to answer questions and were forthright in their responses. DPAC is to be commended for the quality of their research effort as it relates to test development, maintenance, and security. We are encouraged with the current level of funding and hope they continue to receive the resources necessary to continue their important work, as it directly impacts DOD readiness.

Sincerely,

Michael C. Rodriguez, Ph.D.
Professor and Campbell Leadership Chair in Education & Human Development

Tab D

OASD Manpower & Reserve Affairs

Military Personnel Policy
(Accession Policy)

Mr. Chris Arendt, Deputy Director

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)

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
Ms. Gail Lovisone

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
CDR Mark Monturo

Enlisted Program Analyst– COL Ginger Norris

Recruiting Analyst – Lt Col Kelli Beaty

Active Components Recruiting thru August 2018

Strength Data

	Strength Measures	Army	Navy	USMC	USAF	Status/Rationale/Comments
Strength	Overall End Strength Posture (P&R)					
	FY18 NDAA Authorized End Strength	483,500	327,900	186,000	325,100	
	Current Strength	468,331	328,244	185,219	325,222	
	Current Strength Percent	96.86%	100.10%	99.58%	100.04%	

Recruiting Data

	Recruiting Measures	Army	Navy	USMC	USAF	Status/Rationale/Comments
Recruiting Metrics (Percent)	Overall Recruiting Posture (P&R)					
	Monthly Mission Attainment (August 2018)	90.2	100.1	99.8	100.0	Army missed every month but December
	6-Month Average - Contract Mission	80.4	102.8	100.0	112.8	Contracts remain low for Army
	6-Month DEP Attrition Rate	6.9	14.0	18.6	7.6	
	YTD Mission Attainment	90.7	101.1	100.2	100.0	
	Annual Mission Attained (Shipped + DEP)	89.0	99.9	102.5	100.9	
	New Recruit Quality - Tier 1 (HSDG)	95.3	97.7	99.8	98.5	DoD Benchmark Not Less Than 90 percent
	New Recruit Quality - Cat I-IIIA (AFQT)	64.1	75.8	70.4	82.9	DoD Benchmark Not Less Than 60 percent
	FY 2018 - NPS Accession w/Waivers	13.4	10.2	18.4	12.1	Thru 3rd Qtr of FY2018
	DEP Posture for FY 2019	5.0	38.9	53.6	27.8	This is a very fluid number

	Recruiting Lever Measures	Army	Navy	USMC	USAF	Status/Rationale/Comments
Recruiting Levers	Lever Utilization					
	Recruiter Strength (6-Month Average)	7,954	3,143	2,129	1,169	
	Recruiter Strength (Deviation from 10-year Average)	920	-87	**	41	
	*Marketing Dollars	-3.9%	-7.1%	-5.0%	50.4%	
	* Bonus Dollars	118.3%	27.8%	-6.3%	42.5%	

* Percentages are based on Service provided data and may not align completely with Budget Book data.

** Marine Corps changed recruiter count methodology making comparison difficult. Over all number has not changed much over time.

Reserve Components Recruiting thru August 2018

Strength Data

	Strength Measures	ARNG	Army Reserve	Navy Reserve	USMC Reserve	ANG	USAF Reserve	Status/Rationale/Comments
Strength	Overall End Strength Posture (P&R)							P&R Assessment
	FY18 NDAA Authorized End Strength	343,500	199,500	59,000	38,500	106,600	69,800	
	Current Strength	334,459	189,387	57,645	38,233	106,912	68,431	Data lags 1 month /EOM July 2018
	Current Strength Percent	97.37%	94.93%	97.70%	99.31%	100.29%	98.04%	

Recruiting Data

	Recruiting Measures	Army Guard	Army Reserve	Navy Reserve	Marine Reserve	Air Guard	Air Reserve	Status/Rationale/Comments
Recruiting Metrics (Percent)	Overall Recruiting Posture (P&R)							P&R Assessment
	Monthly Mission Attainment (August)	83.7	67.7	100.0	98.6	93.7	1223.6	Depicts recent production
	FYTD Mission Attainment (Oct - August)	78.4	73.3	100.3	103.4	92.4	125.1	Cumulative results
	Annual Mission Attained	72.0	66.4	91.5	98.2	82.6	123.8	Projected success
	New Recruit Quality - Tier 1 (HSDG)	96.8	92.7	96.8	99.5	100.0	99.8	DoD Benchmark Not Less Than 90 percent
	New Recruit Quality - Cat I-IIIA (AFQT)	62.8	63.8	80.8	75.8	78.4	77.1	DoD Benchmark Not Less Than 60 percent
	FY 2018 - NPS Accession w/Waivers	8.7	9.0	19.0	18.6	28.5	16.2	Thru 3rd Qtr of FY2018

Recruiting Levers

	Recruiting Levers Measures	Army Guard	Army Reserve	Navy Reserve	Marine Reserve	Air Guard	Air Reserve	Status/Rationale/Comments
Recruiting Levers	Lever Utilization							
	Recruiter Strength (6-Month Average)	3,344	1,266	503	2,129	388	215	
	Recruiter Strength (Deviation from 10-year Average)	-215	-176	-198	-1,087	28	-5	
	Marketing Dollars	6.5%	**	**	**	263.0%	50.0%	
	Bonus Dollars	18.0%	37.0%	**	**	-30.0%	56.0%	

* Marine Corps changed recruiter count methodology making comparison difficult. Over all number has not changed much over time.

** These dollars are a subset of the Active Components budget and cannot be separated

Questions?

Tab E

Major ASVAB R&D Efforts

Milestones and Project Schedules

Mary Pommerich

Briefing presented to the DAC

Minneapolis, MN

September 2018

Projects

- **ASVAB Development**
  - New CAT-ASVAB Item Pools
  - Developing New CAT Item Pool for CEP*
  - Automating Generation of WK Items/AR and MK 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 at this meeting.

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

New CAT-ASVAB Item Pools

- **Objective**
  - Develop CAT-ASVAB item pools (designated as Pools 11–14) from new items
- **Projected Completion**
  - New item pool implementation: November 2019
- **Subtasks**
  - Write items ✓
  - Pretest, calibrate, and screen items (Summer 2018) ✓
  - Identify item enemies (Aug 2018–Sep 2018)
  - Complete preliminary and final form assembly (Aug 2018–Nov 2018)

New CAT-ASVAB Item Pools (continued)

- **Subtasks** (continued)
  - Modify, test, and deliver CAT-ASVAB software and item pools to MEPCOM (Dec 2018–Jan 2019)
  - Collect and analyze IOT&E data (Feb 2019–Sep 2019)
  - Implement operationally in WinCAT and iCAT (Oct 2019–Nov 2019)
- **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 (Fall 2018)
- **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 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)

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 (Sep 2018–Dec 2018)
  - Assess MK item quality and parameter accuracy (Jan 2019–Feb 2019)
  - Conduct AR data collection (Jan 2018 – Apr 2019)
  - Assess AR item quality and parameter accuracy (May 2019–Jun 2019)
  - Provide final generator, interface, and documentation (Sep 2019)

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) ✓
  - CAT-ASVAB Pool 10 for CEP *i*CAT (Fall 2018)
  - CAT-ASVAB Pools 11–14 (Dec 2019)
  - APT (Fall 2018)
  - 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 (FY 2017)
  - 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 (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) ✓
  - Conduct predictor and criterion data analyses (TBD)
  - Examine projected impact of operational use of MCt scores for selected jobs (2018/2019)

- **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) ✓
  - 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 (TBD)

Evaluating New Cognitive Tests: Cyber Test*
(continued)

- **Subtasks** (continued)
  - Phase IV: MEPS Data Collection II: Operational Support/Adv. Development Continued
    - Develop CAT item pools (Oct 2018)
    - Evaluate feasibility of CAT-Cyber Test (Feb 2019)
    - Conduct additional validation studies (TBD)
    - Program multiple versions of the AF Electronic Data Processing Test and selected Center for Applied Study of Language 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)
      - Administer with CATA and operational EDPT for evaluation of incremental validity (FY19)

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 (TBD)
    - Examine project impact of operational use of CT scores for selected jobs (2018/2019)
  - 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)

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 (in progress)
      - Summarize literature & recommendations on how serious gaming could be used to enhance assessment of cyber aptitude (in progress)
    - Phase II: Develop serious gaming approach (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 data & report results (Dec 2018)
    - 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 (in progress)
    - Administer TAPAS/NCAPS/SDI tests to Basic Recruits to examine construct validity (in progress)
    - 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 applicants on the ASVAB platform (FY15–ongoing)
    - TAPAS pilot testing of Marine Corps officers using paper & pencil (FY17–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)
    - Collect validation data (Jul 2017–TBD)
    - Conduct initial validation study (Summer 2018)

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 (Oct 2018)
    - Develop new items and validate DNA (Q4, 2018)
    - Proof of Concept for gaming environment vice self report format (Q4, 2019)

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 (TBD)

Air Force Compatibility Assessment (AFCA)

- **Objective**
  - Program the Air Force Compatibility Assessment for WinCAT administration
- **Projected Completion**
  - Fall 2018†
- **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 package to MEPCOM (Sep 2018–Sep 2018) ††
  - 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)

Expanding Test Availability: Web Delivery of Special Tests

- **Objective**
  - Transition delivery of special tests from Windows-based platform to web-based platform
- **Projected Completion**
  - Aug 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 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 to the cloud, including Tests 1-3 (Jul 2019–Mar 2020)
  - Decommission WinCAT (Mar 2020–Mar 2020)

† 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 Delivery of Special Tests (continued)

- **Subtasks** (continued)
  - Deploy iCAT to Production in the cloud (Mar 2020–Mar 2020)
  - Deploy TAPAS 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)
  - Migrate DLAB2 to the cloud platform (Jan 2021–Apr 2021)
  - QA DLAB2 on the cloud platform (May 2021–Aug 2021)
  - All special tests operational in the cloud (Aug 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**
  - Aug 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)
  - Migrate iCAT to the cloud, including Tests 1–3 (Jul 2019–Mar 2020)††
  - Deploy TAPAS to Production in the cloud (Mar 2020–Mar 2020)
  - Deploy iCAT 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)
  - Migrate DLAB2 to the cloud platform (Jan 2021–Apr 2021)
  - QA DLAB2 on the cloud platform (May 2021–Aug 2021)
  - All special tests operational in the cloud (Aug 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
List of Acronyms

List of Acronyms

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
CAT-ASVAB	Computerized Adaptive Testing version of the ASVAB
CEP	Career Exploration Program
CIO	Chief Information Officer
CS	Coding Speed
DHRA	Defense Human Resources Agency
DIF	Differential Item Functioning
DLAB	Defense Language Aptitude Battery
DLPT	Defense Language Proficiency Test

List of Acronyms (continued)

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
OCCU-Find	Occupational Finder
OMB	Office of Management and Budget

List of Acronyms (continued)

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
STEM	Science, Technology, Engineering, and Mathematics
STP	Student Testing Program
TAPAS	Tailored Adaptive Personality Assessment System
WinCAT	Windows-based CAT-ASVAB
WPA	Work Preferences Assessment

Tab F

OFFICE OF PEOPLE ANALYTICS

Next Generation ASVAB and ETP Update

Mary Pommerich

Defense Personnel Assessment Center

DAC Meeting
September 20-21, 2018
Minneapolis, MN

Purpose and Overview

- Provide background and update on history, status, and plans for the next generation of ASVAB and the special tests administered in the Enlistment Testing Program (ETP).
  - Document history to date.
  - Review ASVAB expert panel review recommendations.
  - Provide update on new tests of interest and status.
  - Review the philosophy of the ASVAB question—history and hang-ups.
  - Review plans for evaluating current tests on the ASVAB.
  - Review next steps.

OPA
OFFICE OF PEOPLE ANALYTICS

History

- An expert panel was convened in 2005–2006 to consider the status of the military enlistment testing program and to make recommendations for improvements and enhancements.
- The panel:
  - Reviewed ASVAB content, methodology, and use.
  - Discussed problems associated with the current battery.
  - Reviewed new types of cognitive and non-cognitive skills not currently measured by ASVAB that might prove valid for selection and classification.
  - Developed recommendations for potential changes to the battery.

Expert Panel Recommendations

- Expert panel recommendations [with MAPWG rank prioritization]:

1.	Implement CAT at MET sites	[1]
2.	Consider classification accuracy when evaluating content changes	[1]
3.	Re-evaluate the contents of the ASVAB	[1]
4.	Examine validity regularly	[5]
5.	Increase time for seeding new items and measures	[5]
6.	Include validated non-cognitive measures in job classification composites	[7]
7.	Include nonverbal reasoning test on ASVAB	[8]
8.	Develop standardized data banks on Service member performance	[8]
9.	Relax the requirement for criterion validity of new measures	[8]
10.	Implement content controls in CAT	[8]
11.	Continue utility research on non-cognitive measures	[12]
12.	Develop IT/communications technology test	[13]
13.	Review test specifications on a regular basis	[13]
14.	Evaluate WK and PC for ESL examinees	[13]
15.	Consider the multidimensionality of the ASVAB	[13]
16.	Evaluate Spanish verbal test for ESL examinees	[17]
17.	Use automatic item generation	[17]

New Tests of Interest and Status

- Tailored Adaptive Personality Assessment System (TAPAS)
  - Extensive research has been conducted by the Services on the usefulness of TAPAS as a screening instrument for military applicants.
  - Some concerns about the seemingly low stability of test-retest scores over time have recently been raised by RAND following an independent review.
  - A TAPAS expert panel has been formed to review TAPAS research in light of the recent technical critiques.
- Cyber Test (CT; formerly called the Information Technology and Literacy Test)
  - Extensive research has been conducted by the Services on the usefulness of CT as a screening instrument for military applicants.
  - Some concerns about vulnerability to compromise associated with the two 29-item fixed forms currently in use.
  - Some concerns about the content being too difficult for the applicant population and the lack of moderately difficult items.
  - Feasibility of CAT-CT will be revisited after development of new item pools.

New Tests of Interest and Status

- Mental Counters (MCt)
  - A working memory test currently being administered to Navy applicants on the ASVAB platform.
  - Very promising characteristics, including high reliability and very short testing times.
  - There is a persistent floor effect, with approximately 4–9% of examinees receiving a score of zero each year.
  - Options for eliminating the floor effect are being considered.
- Abstract Reasoning (ART)
  - A test of nonverbal reasoning currently being administered to DLI applicants as part of DLAB2.
  - A planned research study of nonverbal reasoning tests has been deferred indefinitely.
  - Analyses of ART and MCt (also administered in DLAB2) could give some information about the desirability of investigating ART further.

WHY NO MODIFICATIONS TO ASVAB TO DATE?

- In response to a DPAC briefing discussing considerations in adding or deleting ASVAB tests, the DAC encouraged DPAC to determine the "uses that each Service requires the ASVAB to meet, in order to establish the philosophy of the test" [June 2011].
- The MAPWG has had numerous discussions since 2011 about potential modifications to the ASVAB. These discussions have been hung up by several key issues:
  1. The unresolved question of what should the philosophy of the ASVAB be?
  2. Concerns about insufficient resources to accommodate a revised ASVAB that takes more time than the current battery.
  3. The logistical difficulties associated with making changes that would impact existing composites and systems set up to operate on those composites.
- DPAC is now hopeful that application of an argument-based approach to validation of the ASVAB [validity framework effort] will help answer the question of what the philosophy of the ASVAB should be.

PLANS TO EVALUATE ASVAB TESTS*

- 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.

*Briefed to the MAPWG/DAC in 2015.

Next Steps

- Continue efforts to evaluate and resolve issues/concerns pertaining to the new tests of interest (TAPAS, Cyber Test, Mental Counters, Abstract Reasoning).
- Continue efforts to evaluate tests currently in the ASVAB.
- Complete effort to apply argument-based approach to validation of the ASVAB.
- Stakeholders develop a shared vision that defines the purpose and general makeup of the next generation ASVAB.
  - Revisit the question of the philosophy of the ASVAB as needed, following establishment of a validity framework.
- Establish a systematic process to follow for evaluating potential changes and making decisions regarding tests in the ASVAB.
  - Recommended by the DAC following the last revision to the battery.
  - DPAC presented a proposed process for potential changes to the ASVAB in 2014.

Next Steps

- Review and update the psychometric checklist, as needed, for the purpose of evaluating tests to be administered as part of the ASVAB.
  - Current checklist was developed for making decisions about adding tests to the ASVAB platform, not the battery.
- Services/proponents complete the updated psychometric checklist for new tests of interest, documenting all new information since the last checklist was completed.
- Revisit logistical questions with stakeholders, including the feasibility of lengthening the ASVAB and the feasibility of dropping existing tests.
- Stakeholders summarize impact of potential modifications to the battery and identify resources to support a revised battery.
- Compile all information, then identify and discuss potential changes to the contents of the ASVAB and tests administered in the ETP.
  - Given the complexities associated with making changes to the battery, DPAC believes it is best to consider all new and existing tests at once, rather than on a case-by-case basis.

Tab G

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

ASVAB Validity Argument Briefing

September 20, 2018

Presented to: *Defense Advisory Committee on Military Personnel Testing*

Arthur Thacker

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

Presentation Overview

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

Validity Argument Overview

- The validity of an assessment cannot be summarized via a single statistic or coefficient. Validation depends on the assessment's purpose, the inferences made from assessment results, and the uses of those results.
- Argument-based validation tests the underlying claims that must be true to support the inferences made from assessment information (scores).
- An assessment score may be valid for multiple purposes.
- When assessments are used for multiple purposes, it is rare that the assessment is equally valid for all of them.
- Evidence is collected for a validity argument to support claims in a chain of reasoning, where any claim in the chain found to be weak may undermine subsequent claims.
  - Example 1—Poor item quality can undermine all results from an assessment
  - Example 2—Even if all aspects of a test seem supported and strong validity evidence for use of scores is available for a given year, poor year-to-year equating can undermine cross-year comparisons of scores.
- If multiple inferences are drawn from a single assessment score (or event), each inference may have its own unique validity argument.

Where Do We Start

- What are the most important inferences we want to make?
  - Admission into military branches
  - Placement into training programs or advanced educational opportunities
  - Prioritization of recruiting efforts

- Establishing Draft TOAs for ASVAB
  - ASVAB primarily relies on an informal reasoned approach
  - Evidence is not typically tied to organized claims or assumptions
  - A TOA (or similar) is required to frame interpretive and validity arguments

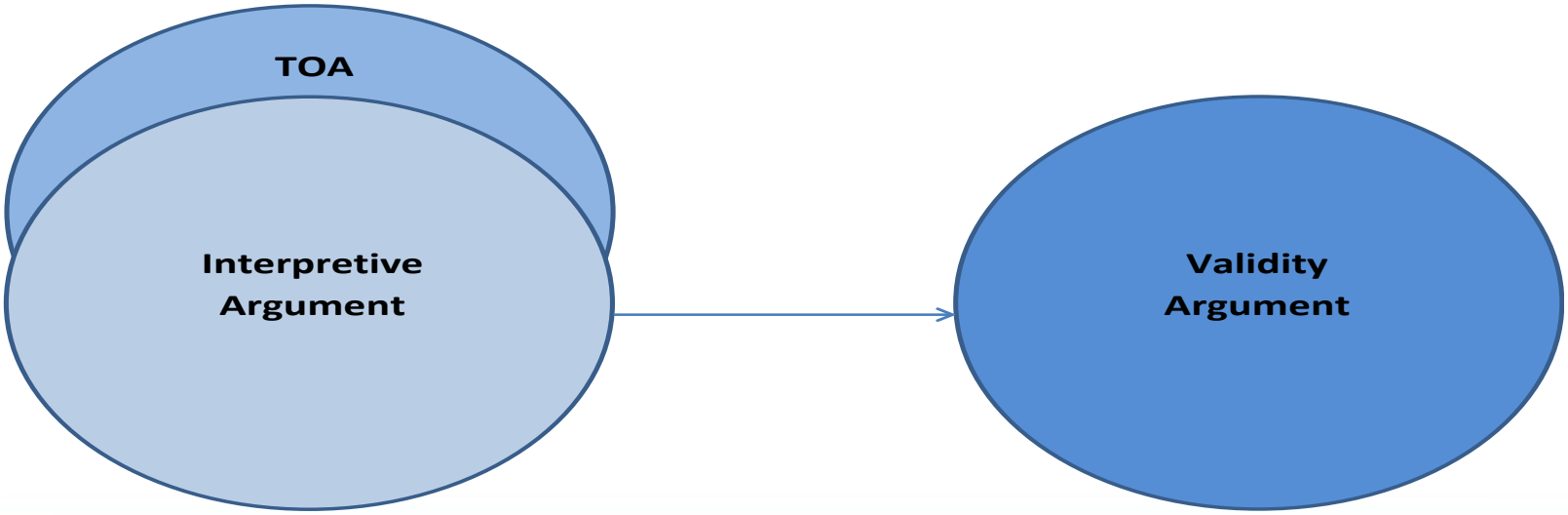
- Bounding the Argument (Limitations)
  - Will not address admittance to specific training programs or occupations (each would require its own body of evidence which is beyond the scope here)

Validity Argument Illustration

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

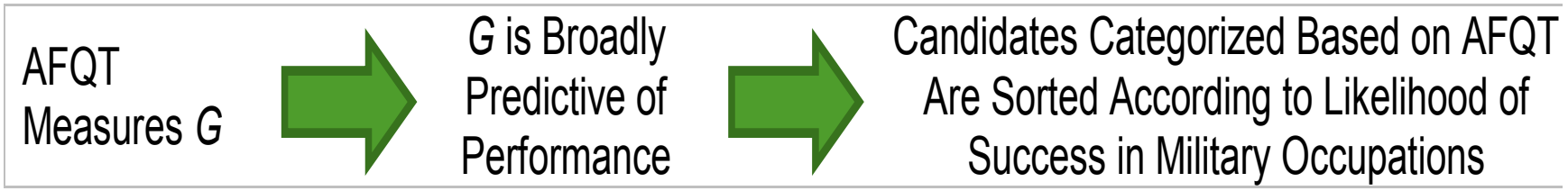
Interpretive Argument—a description of the inferences (and uses) that the test scores support

Validity Argument—evidence providing justification for the inferences (and uses) in the interpretive argument.



Draft AFQT Theory of Action

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

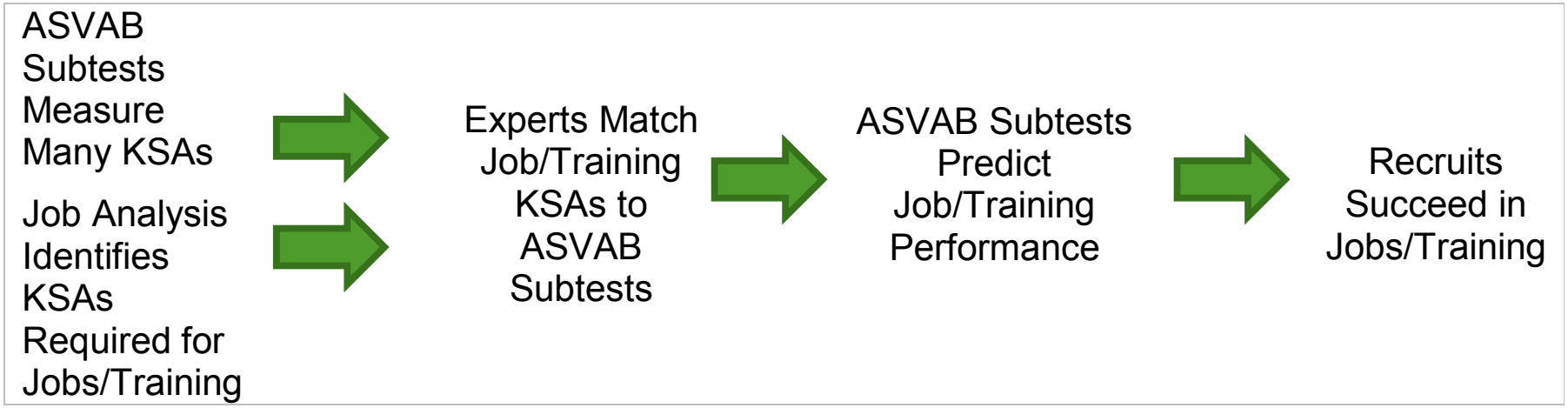


We can then develop claims that must be supported for each step in the TOA to be true.

The AFQT's primary function is selection.

Draft ASVAB Theory of Action #1

Job Analysis Model

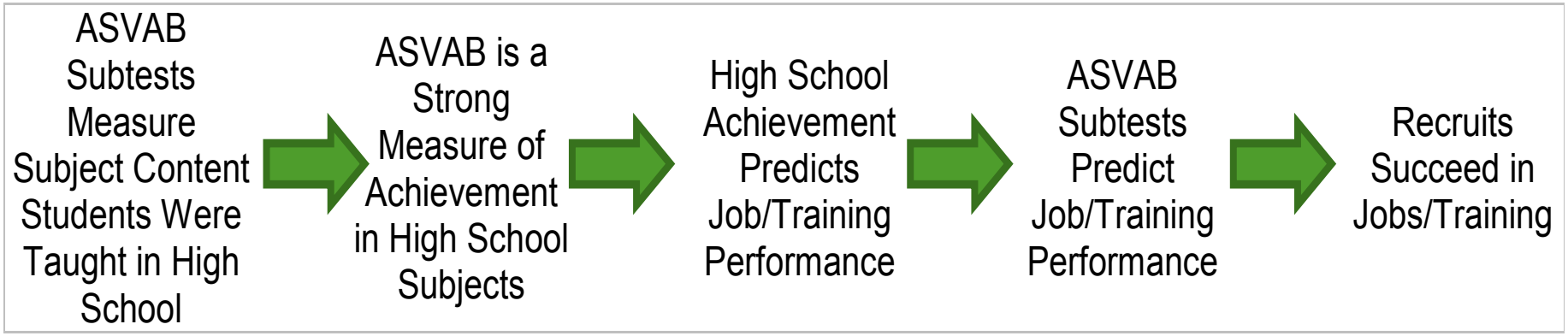


Model relies on clear linkages between KSAs required for military jobs/training and KSAs measured by ASVAB.

The ASVAB's primary function is classification.

Draft ASVAB Theory of Action #2

Prior Training Success Model



Model relies on prior educational/training success being predictive of future success in military jobs/training.

Discarding this model because this is not how the Services conceptualize the ASVAB.

AFQT Claims

- **AFQT Measures G**
  1. A candidate's score on the AFQT is an estimate of that candidates true G.
  2. The predictive nature of G is continuous for nearly the full scale for the AFQT (i.e. a higher score always results in a better prediction of outcome, irrespective of the area of the scale the score falls in).
  3. The AFQT categories represent important differentiators among candidates.
  4. AFQT scores have high overall reliability, especially near the cut points for the categories.
  5. AFQT scores have high classification accuracy.
  6. AFQT scores are largely free from construct irrelevant variance.
- **G is Broadly Predictive of Performance**
  7. Other G measures are used to predict performance broadly in non-military contexts, and these measures correlate positively and strongly with AFQT.

AFQT Claims (continued)

- **Candidates Categorized Based on AFQT Are Sorted According to Likelihood of Success in Military Occupations**
  8. AFQT scores correlate positively and strongly with success in military careers.
  9. AFQT scores are unbiased with regard to race/ethnicity, gender, etc.

The TOA may also include some uses of the AFQT scores. These may not fall under the heading of inference, but are vital to the success of the assessment. Examples might include the following.

- **Utility and Implementation Factors**
  10. Users of the AFQT scores can interpret and understand score reports.
  11. Users of the AFQT scores are sufficiently trained to help candidates understand their options based on the AFQT performance.
  12. Factors outside of AFQT scores that contribute to the decision to enlist a candidate enhance predictions based on AFQT alone.

AFQT Draft Validity Argument Excerpt

Assumption	Claim	Evidence	Specific Evidence (Citations or Links)
AFQT Measures G			
The 4 AFQT components selected are the best options for a G proxy.	If verbal and quantitative ability are strong proxies for G, then AR, WK, PC, and MK are the best subtests for estimating G from ASVAB.	1. Evidence that items reflect G stemming from item writing efforts.	
		2. Correlation studies linked to other measures of G (e.g. IQ, ACT, SAT).	
		3. Literature linking math and verbal tests to G.	
		4. Dimensionality studies among ASVAB sub-tests.	
	If AR, WK, PC and MK are the best subtests for predicting G, then information from other subtests should add minimal improvement in prediction strength.	1. Regression-based or similar studies indicating the added prediction strength gained by including potential additional subtests.	
		2. Subtests should be considered as candidates for AFQT as they are introduced or substantively revised.	

Next Steps

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

Challenges

1. Lack of models from comparable assessment systems
2. 50 years of history
3. Multiple users
4. Varied score information
5. Multiple inferences need to be supported
6. Discerning which ASVAB literature is relevant for the validity argument is not straightforward
7. ASVAB literature is not always unbiased

Thank you!

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

Tab H

OPA

OFFICE OF PEOPLE ANALYTICS

Mental Counters 4.0

Eliminating the Floor Effect in Operational Testing

Ping Yin, HumRRO

Gregory Manley, DPAC

DACMPT

September 20, 2018

OVERVIEW

- **Background**
- **What do we know about the floor effect?**
- **Moving forward, what can we do to eliminate the floor effect?**
  - Option 1: Change the test
  - Option 2: Without changes to the test
- **More detailed discussion and demo for Option 2**
  - CAVEAT: Not all proposed recommendations may be feasible to implement at this time due to practical constraints.
- **Recommendation for Mental Counters 4.0**

MENTAL COUNTERS: BACKGROUND

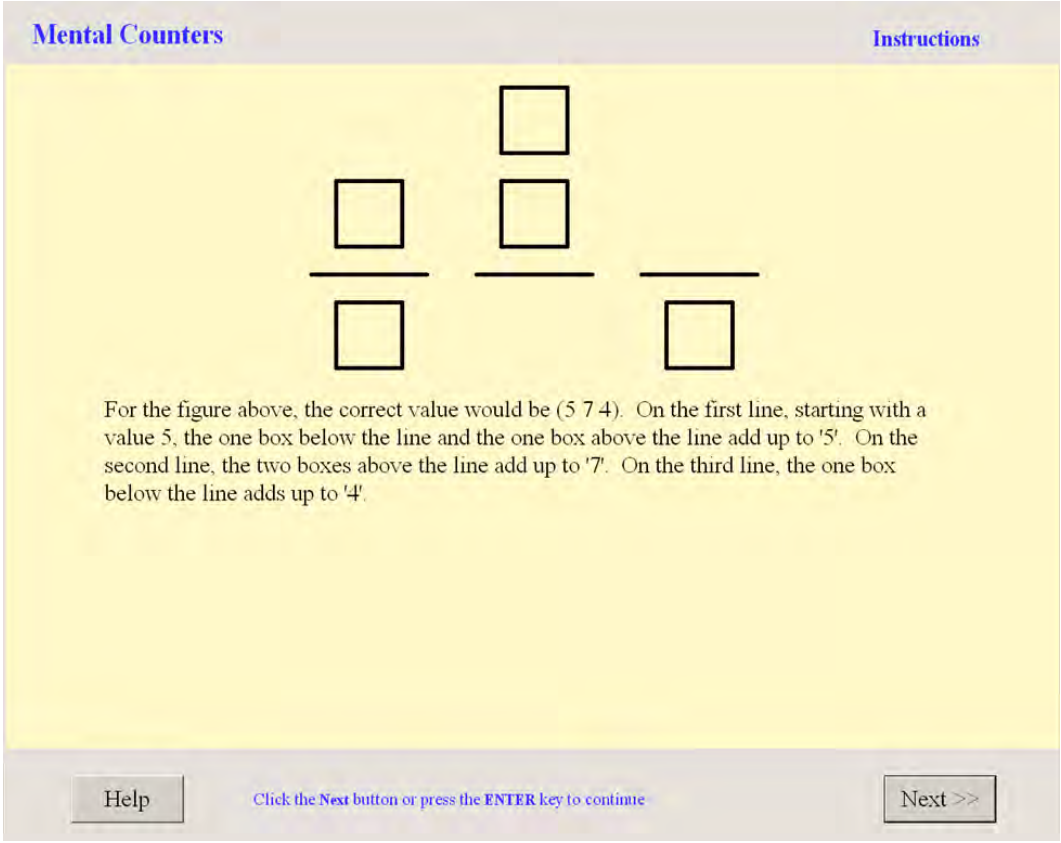
- **Mental Counters (MCt) is a test of working memory originally developed by the Navy and studied as part of the Enhanced Computer-Administered Test (ECAT) battery evaluation.**
  - 32 items
  - Currently administered to Navy applicants on the CAT-ASVAB platform
  - Measures a unique domain not represented on the ASVAB
    - Evidence of incremental and predictive validity
    - Evidence of classification efficiency
    - Evidence of excellent reliability
    - No adverse impact for gender
    - Small effect for practice
    - Excellent candidate for automatic item generation

MENTAL COUNTERS: BACKGROUND

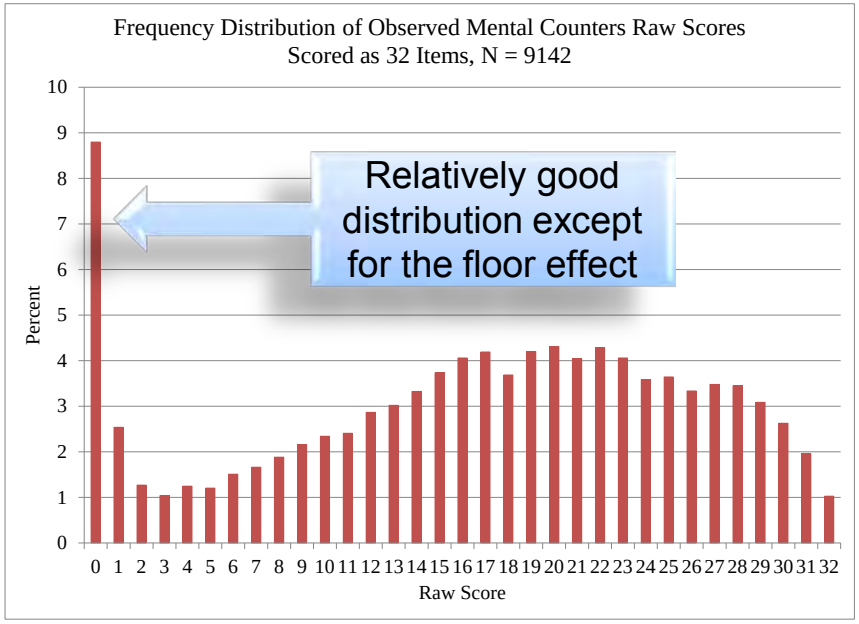
- **The MCt test requires the examinee to count the number of boxes that flash above or below one of three stationary lines on the computer screen.**

MENTAL COUNTERS: BACKGROUND

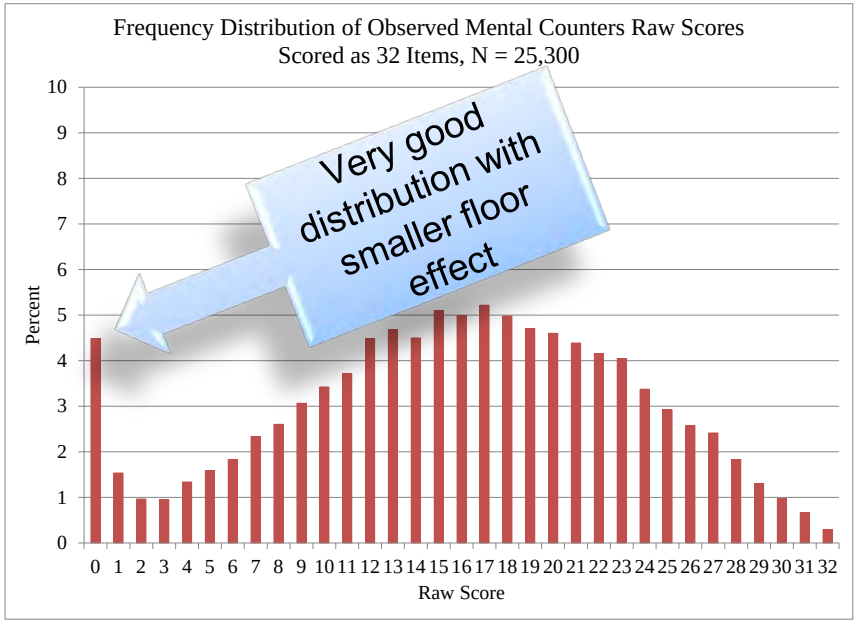
- **Counters for each line start at 5. A value of 1 is [added to]/[subtracted from] the counter for a line if a box appears [above]/[below] the line.**



WHAT DO WE KNOW ABOUT THE FLOOR EFFECT?



Version 2.0 (2013)

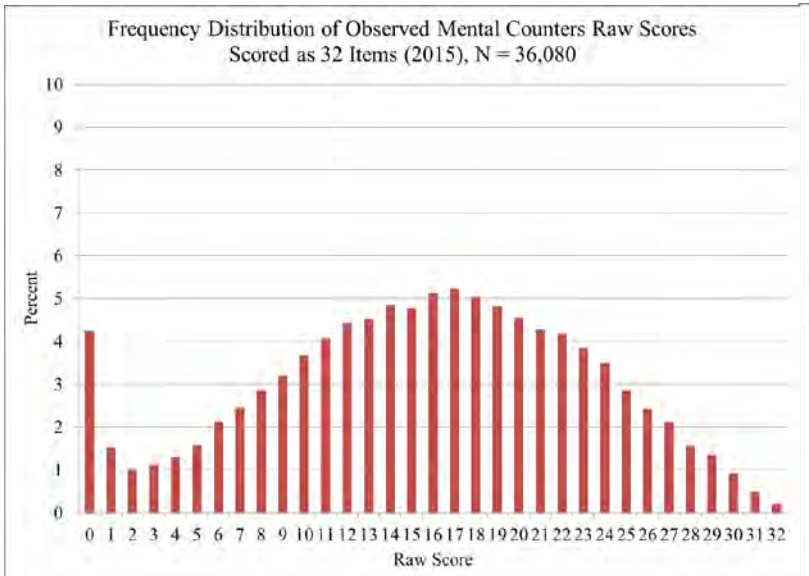


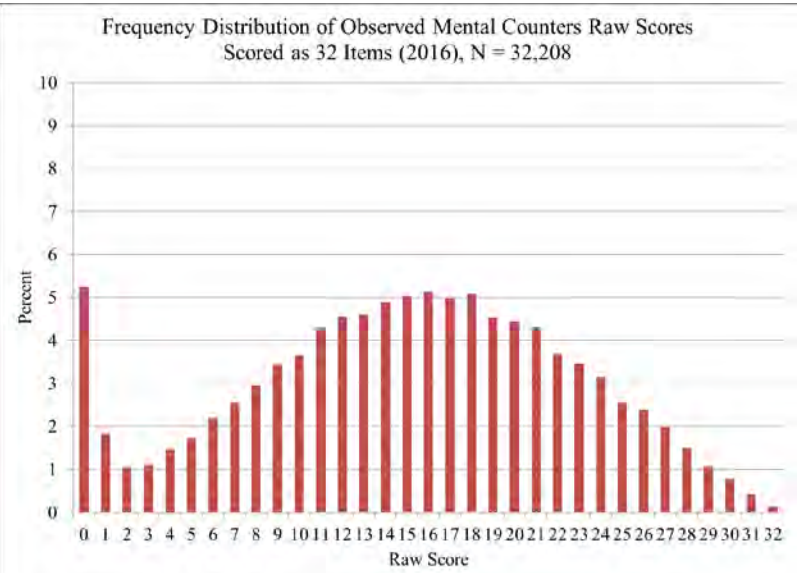
Version 3.0 (2014)

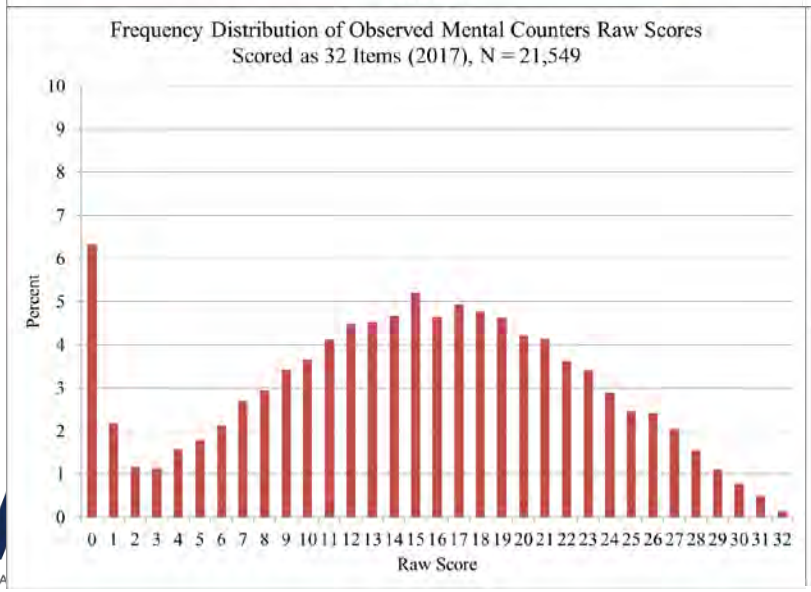
Minor clarification of instructions to emphasize that the counter starts at 5.

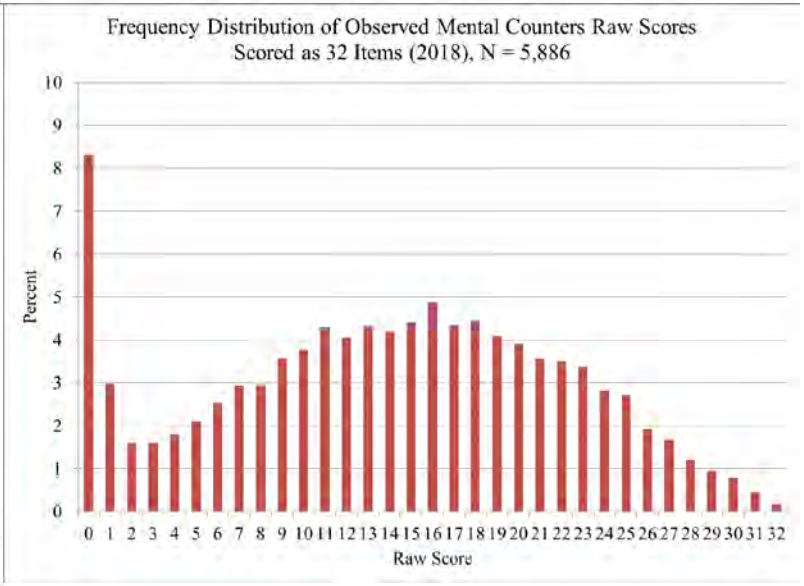
OPA
OFFICE OF PEOPLE ANALYTICS

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



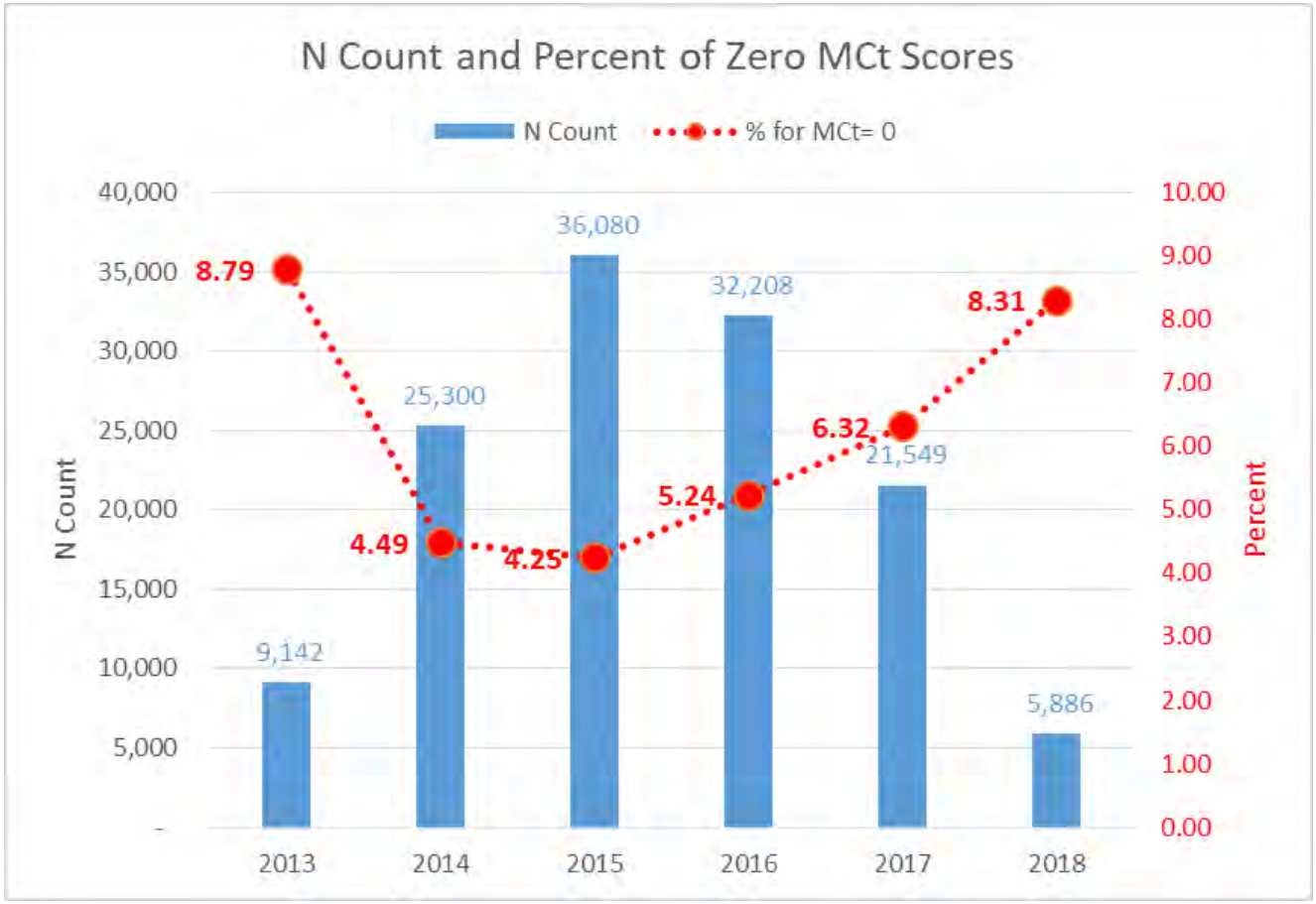






OPA
OFFICE OF PEOPLE ANA

FLOOR EFFECT OVER TIME: 2013–2018



WHAT DO WE KNOW ABOUT THE FLOOR EFFECT?

- **We were hoping that the floor effect would be reduced over time, but this is not the case.**
  - Over time, test usage has gone down, and the floor effect has gotten worse.
- **We need to think proactively about how to eliminate the floor effect.**

WHAT CAN WE DO TO ELIMINATE THE FLOOR EFFECT?

- **Option 1: Change the test**
  1. Order items from easiest to hardest
  2. Add easier items to the test
     - Rationale: The current administration design is unbalanced
     - Item difficulty is related to two factors:
       - Number of counter adjustments
       - Delay between counter adjustments
     - Add 5 easier items (5 adjustment, 830 MS delay)

		Delay		
		830 ms	500 ms	Total
# Adjustments	5	0 Items	5 Items	5 Items
	6	8 Items	6 Items	14 Items
	7	8 Items	5 Items	13 Items
	Total	16 Items	16 Items	32 Items

WHAT CAN WE DO TO ELIMINATE THE FLOOR EFFECT?

- **Option 1 Pro:**
  - May effectively eliminate the floor effect
  - Can balance the administration design
- **Option 1 Con:**
  - Could change the test
  - Could introduce potential ceiling effect
  - Could change the overall difficulty of the test, which will make it difficult to compare performance from previous administrations without an equating study
  - Would require equating if test/score is changed
  - Would require a conversion table of MCt scores if test is changed

WHAT CAN WE DO TO REDUCE THE FLOOR EFFECT?

- **Option 2: Without changes to the test**
  - Make instructions clearer
  - Add an animated demo of the task
  - Have examinees cycle through instructions and demo
  - Require that an examinee answer at least one practice item correctly in order to start the operational test
- **Option 2 Pro:**
  - Very likely to reduce the floor effect
    - Minor clarification between v2 and v3 reduced the floor effect by half
  - No need to equate
- **Option 2 Con:**
  - Practical constraints
    - Update instruction, demo, and practice items and sequencing
    - Minor adjustment during operational test

OPTION 2

- **Possible factors contributing to the floor effect:**

Possible factors	What we can do about it
Applicants with limited working memory	MCt is a valid measure for working memory
Lack of motivation	Make sure examinees believe their scores will count and they are taking the test seriously
Mental exhaustion	Provide a longer break after the ASVAB
	Add a pause between screens to allow for additional time between answering an item and starting the next item
Not able to understand the task before taking the test	Clarify instructions, add a visual demo, cycle examinees through the instructions and demo

OPTION 2: AREAS OF POTENTIAL UPDATES

- Instruction and demo
- Practice items
- Minor update during operational testing

OPA
OFFICE OF PEOPLE ANALYTICS

Instruction and Demo

INSTRUCTION AND DEMO

- **Currently, there are at least 50 screens for instruction, demo, and practice**
- **Recommended updates:**
  - Reduce the amount of text/screens for instruction
  - Provide a video-like visual demo to aid instruction
    - Based on recommendations made by Held and Carretta for revisions to MCt Version 3.0
  - Utilize TA assistance to ensure a clear understanding of the test
    - Update or provide TA manual so that better assistance can be provided to examinees who need help understanding the test
    - Provide guidelines when examinees fail to answer any practice items correctly
- **New instruction and demo:**
  - MCt_demo_Do_Not_Rename.pptx

INSTRUCTION AND DEMO

- **Pilot test the new instruction and demo**
  - Select a few volunteers to think aloud (cognitive lab) as they go through the instruction and demo
  - Divide volunteers randomly into two groups:
    1. Current MCt instruction
    2. New instruction and demo
  - Will provide useful information before finalizing the new instruction and demo

Practice Items

CURRENT PRACTICE ITEMS

- **Presented in two groups:**
  - 1st group (easy to moderate items):
    - N=2 if answered all correctly
    - N=5 if answered all incorrectly
  - 2nd group (harder items):
    - N=2 if answered all correctly
    - N=4 if answered all incorrectly

Examinees are given more practice items if they didn't answer the first two items correctly

Examinees are given more practice items again if they didn't answer the two harder items correctly

- **The total number of practice items varies between 4 (if answered all correctly) and 9 (if answered all incorrectly).**
- **Test will start even if an applicant did not answer any of the practice items correctly**

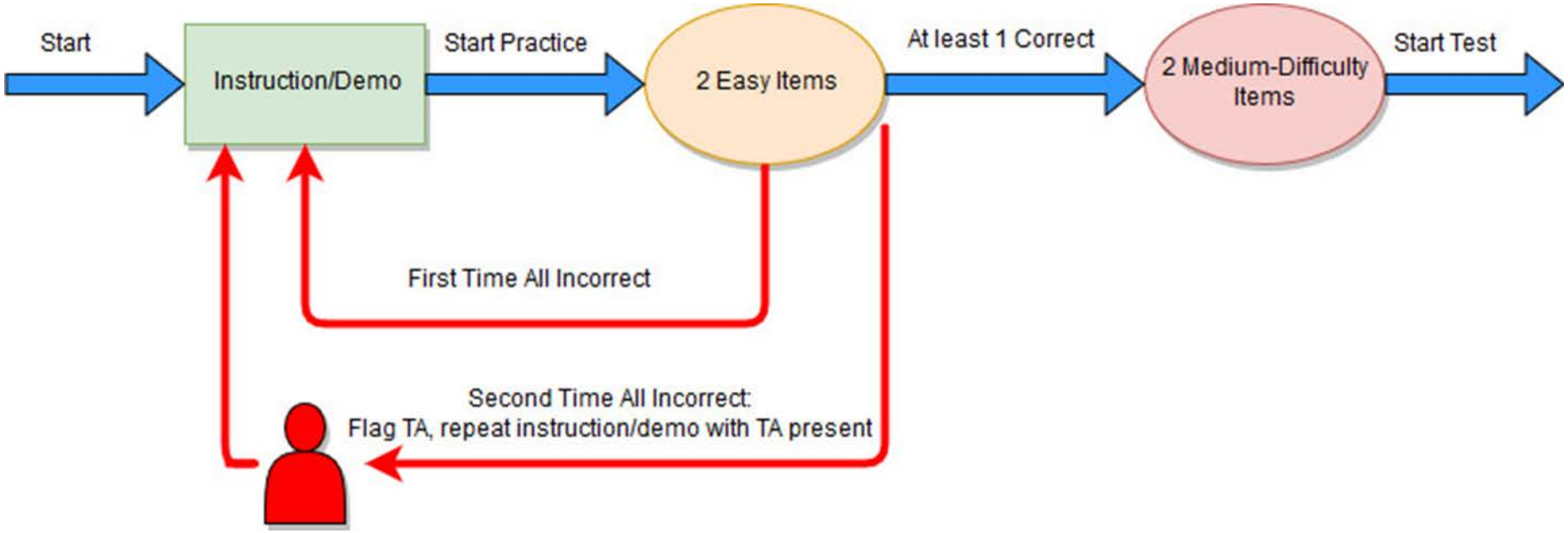
RECOMMENDED UPDATE FOR PRACTICE ITEMS

- **Reduce the number of screens for practice items**
- **Reduce the difficulty of the 2nd group of practice items (from hard to moderately difficult)**
- **Increase the delay for the easiest practice items**
- **More detailed instructions from TA, if needed**
- **Specifically require that an examinee answer at least one practice item correctly in order to start the operational test**
  - Simply providing more (and harder) practice items is not enough to ensure a clear understanding of the test. To reduce the floor effect, we need all examinees to have a clear understanding of the task before taking the test.
  - Any practical or logistical concerns?
    - For example, if an examinee failed to answer any practice item correctly after N (e.g., 3) attempts, we will allow the examinee to take the test but assign a code in the examinee's record for identification.

OUTLINE OF THE NEW PRACTICE ITEM SEQUENCE

- **Practice items are still presented in two groups**
  - 1st group (easy items): 2 items
  - 2nd group (moderately difficult items): 2 items
- **Practice items**
  - Easy: 5 adjustment, 830 MS
  - Moderately difficult: 5 adjustment, 500 MS
- **All examinees start with two easy practice items**
- **If an examinee answers one or both easy items correctly:**
  - The examinee will be given the 2nd set of practice items. Regardless of whether they answer any of the 2nd set correctly, the test will start once the 2nd set of practice items is complete.
- **If an examinee fails both easy items:**
  - The first time, the examinee will review the instruction and demo and start the practice again.
  - The second time the examinee fails both easy items, the TA will be signaled to help, and then the examinee will review instruction/demo and restart with the 1st set of practice items with TA's assistance.

FLOWCHART OF THE INSTRUCTION, DEMO, AND PRACTICE ITEM SEQUENCE



Minor Updates During Operational Testing

MINOR UPDATES

- **Add a pause between screens to allow for additional time between answering one item and starting the next item**
- **Remove the halfway break during the test**
  - Currently during MCt, there is a break between items 16 and 17:
    - "*You are half-way through the Mental Counters test. The test will pause for a 15 second rest period*."
  - The average time for the MCt is less than 4 minutes, and a break is not necessary
- **Equating may be considered if there is any concern of "changing the test" due to the minor updates**

RECOMMENDATIONS FOR MENTAL COUNTERS 4.0: OPTION 2

Instruction/Demo	Practice	Operational
Provide clear, simple, and easy-to-follow instruction and demo	Provide targeted practice items to ensure that an examinee must answer at least one item correctly before proceeding further	Provide a testing envirement with minimim distraction to help examinees focus and complete the task

OPA
OFFICE OF PEOPLE ANALYTICS

RECOMMENDATIONS FOR MENTAL COUNTERS 4.0: OPTION 2 ADDITIONAL STEPS

- **Solicit short, post-MCt feedback from examinees (similar to the DLPT) to guide future changes for improvement:**
  - MCt
    - Instruction
    - Demo
    - Practice
    - Operational test
  - TA
  - Motivational level
  - Fatigue
  - Other factors
- **Statistically differentiate examinees who are engaged in "rapid-guessing" behavior from "solution" behavior using response time**
- **Consider Option 1 if floor effect is not reduced after implementing option 2**

Questions?
Comments?

Tab I

CAT-ASVAB Form 10 Equating

Presenters : **D. Matthew Trippe, HumRRO**

September 20, 2018

Objectives

- Form 10 background
- ASVAB equating design & procedure
- ASVAB equating results
- ASVAB equating evaluation analyses
- Questions/Discussion

Form 10 Background

- Form 10 is a new CAT-ASVAB form developed from old paper and pencil (P&P) forms, intended for use with Career Exploration Program (CEP) iCAT
- Form 10 item parameters have been calibrated and scaled (through linear transformation) to be on the same scale as operational CAT-ASVAB forms 5–9
- As an extra precaution, form 10 theta scores will be equated to theta scores on CAT-ASVAB form 4 (a reference form used for the purpose of equating analyses)
- Equating ensures that form 10 scores have the same meaning or can be treated interchangeably with operational form scores

ASVAB Equating Design & Procedure

- Rigorous equating procedures were developed by DPAC to equate forms 5–9 (most recent equating)
  - Used this as template for equating form 10
  - Also a template for new forms 11–14
- Linear equating methods were used to derive constants to transform IRT-based theta scores on form 10 to scale of the reference form 4
- Conducted at the subtest level
- Linear equating constants match the mean and variance of each subtest distribution
- Works well to the extent that subtest distributions have similar shapes
- Evaluated comparability of composite distributions to ensure subtest equating resulted in sufficient precision

ASVAB Equating Design & Procedure

- Perform equating in three phases of operational administration of form 10 to military applicants
  - Each phase includes progressively larger sample size
  - Intent of phased design was to maximize accuracy of reported operational scores
  - Random groups design
- Each applicant was assigned to a single form with 1/6 assignment probability
  - The reference form 4
  - An operational form (5, 6, 8, 9)
  - Form 10

ASVAB Equating Design & Procedure

Phase	Step	Duration	Target Sample Size
1 *	Data collection	2 weeks	1000 per form
	Equating + Replacement of Transformation Constants	2 weeks	
2	Data Collection	3 weeks	2500 per form
	Equating + Replacement of Transformation Constants	2 weeks	
3	Data Collection	7 weeks	6000 per form
	Equating + Replacement of Transformation Constants	2 weeks	

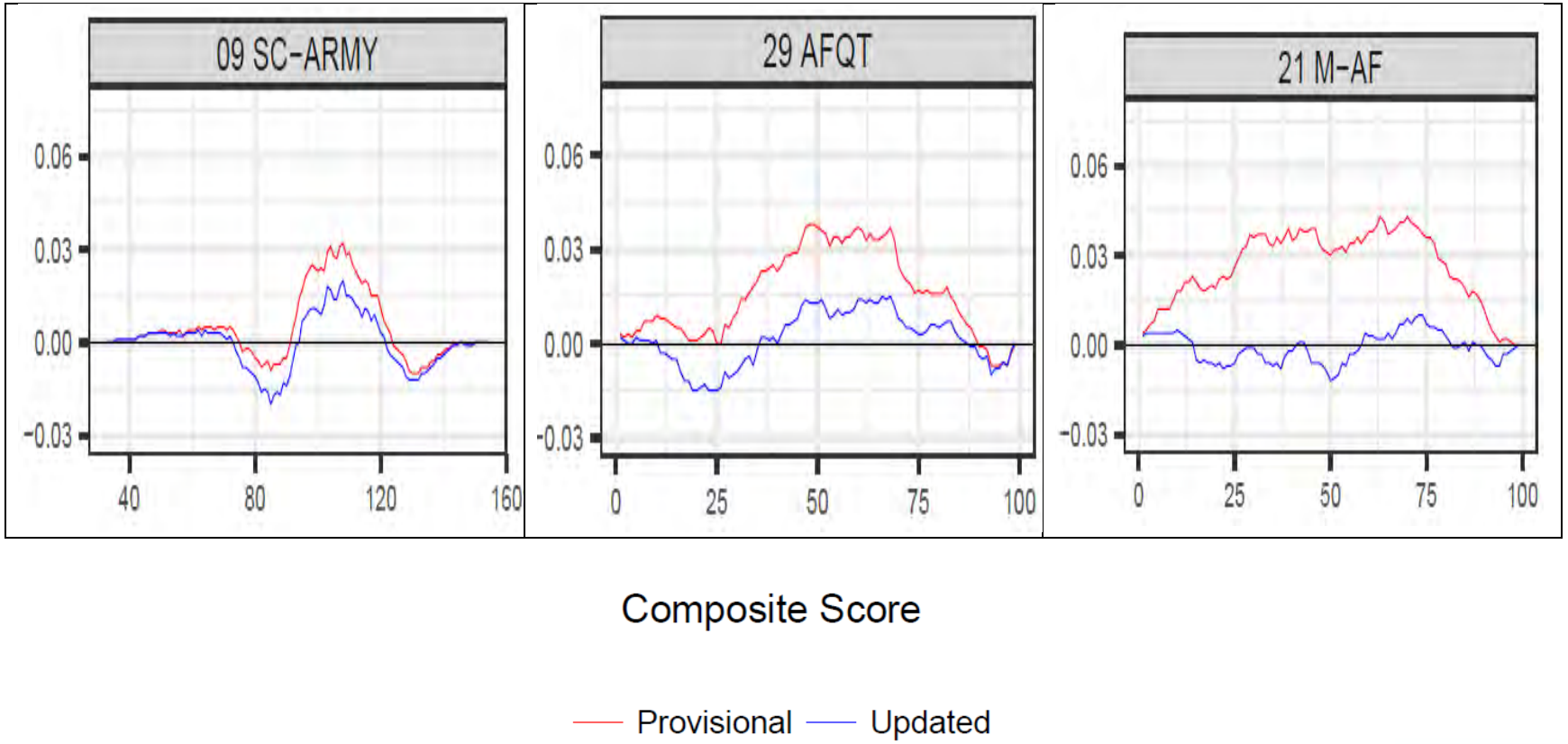
- *Rescaling/linear transformation performed on form 10 parameters prior to phase 1
- *Provisional score transformations based on the IRT property of invariance
  - Scores are defined independent of specific/common items
  - Parameters and scores are population independent

ASVAB Equating Results

- Provisional transformation constants were updated after phase 1 and phase 2 sample sizes achieved
- Evaluated differences in qualification composite cumulative distribution functions (CDFs) between reference form 4 and form 10 (examples on next slide)
- Provisional transformation constants were not replaced after phase 1 and phase 2 target sample sizes achieved as originally planned
  - Replacement of composite transformation constants is a non-trivial update to CAT-ASVAB
  - Requires change to software; incompatible with data collection schedule
  - Evidence suggests provisional constants were sufficiently accurate for reported operational scores
- Transformation constants updated after phase 3 target sample size achieved on July 23, 2018
- Operational form 10 transformation constants will be replaced based on phase 3 results

ASVAB Equating Results

- Form 10 – Form 4 qualification rate difference examples from phase 3



ASVAB Equating: Data Collection Results

Form	Description	Phase 1	Phase 2	Phase 3
4	Reference	1,108	2,547	6,020
5	Operational	1,203	2,625	6,120
6	Operational	1,154	2,562	6,039
8	Operational	1,176	2,567	6,185
9	Operational	1,194	2,598	6,138
10	New	1,154	2,611	6,141

ASVAB Equating: Phase III Analyses

- Random group equivalence
- Equating transformation constant estimation
- Form subtest intercorrelation equivalence analysis
- Composite distribution equivalence
- Subgroup performance across forms
- Operational form comparison
- Provisional equating transformation accuracy

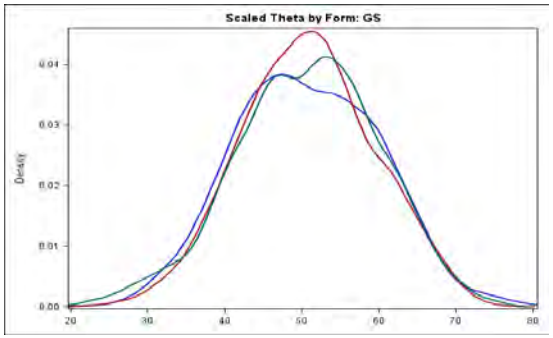
Random Group Equivalence

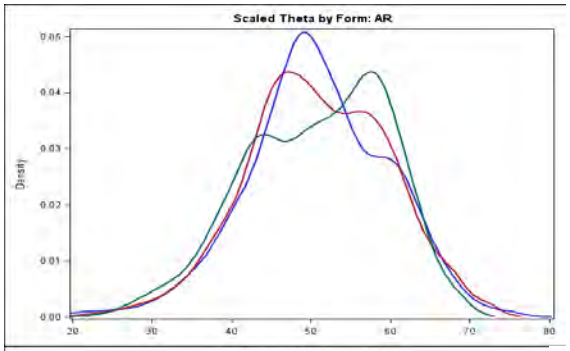
- Does assignment procedure produce equivalent groups with respect to key demographic variables?
- Compare distributions of key demographic variables across assignment to forms 4 and 10
  - Gender: χ^2 (1, N = 12,150) = 0.48, p < 0.48
  - Education: χ^2 (2, N = 10,749) = 3.3, p < 0.19
  - Race: χ^2 (2, N = 11,465) = 6.4, p < 0.04
  - Ethnicity: χ^2 (1, N = 12,142) = 1.48, p < 0.22
- Expect groups assigned to different forms to be randomly equivalent

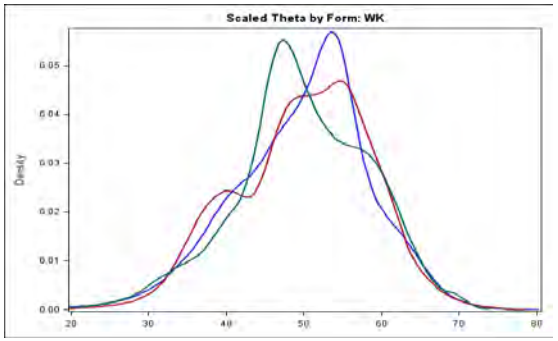
Equating Transformation Estimation

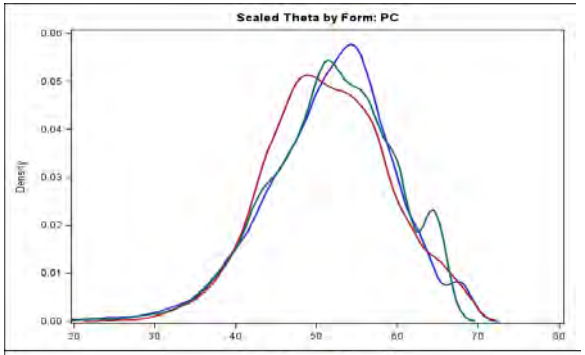
- Is linear transformation adequate?
- Do subtest distributions have similar shapes?
- Evidence of systematic difference in shapes of subtest distributions
- Qualification decisions are based on composite scores
- Composites are likely to be more normal-like

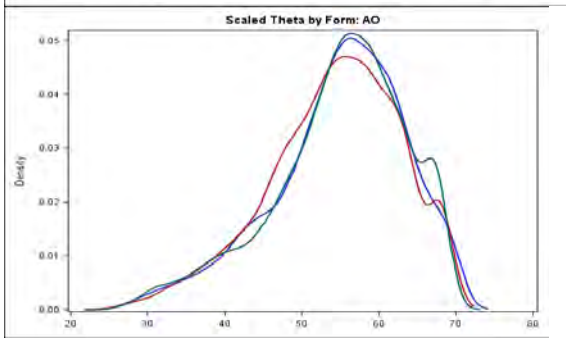
ASVAB Equating: Subtest Distributions

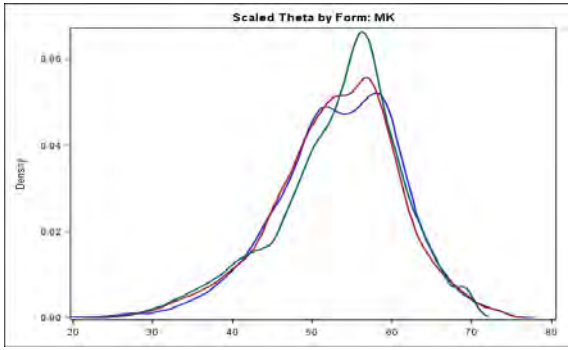


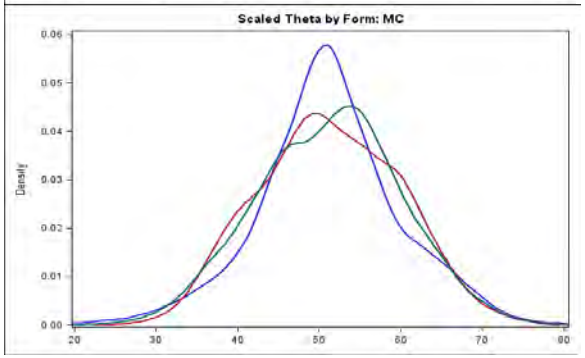


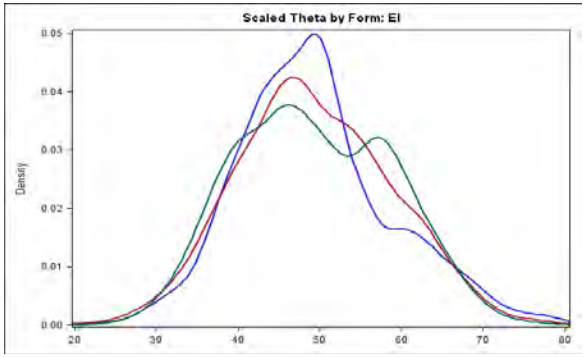






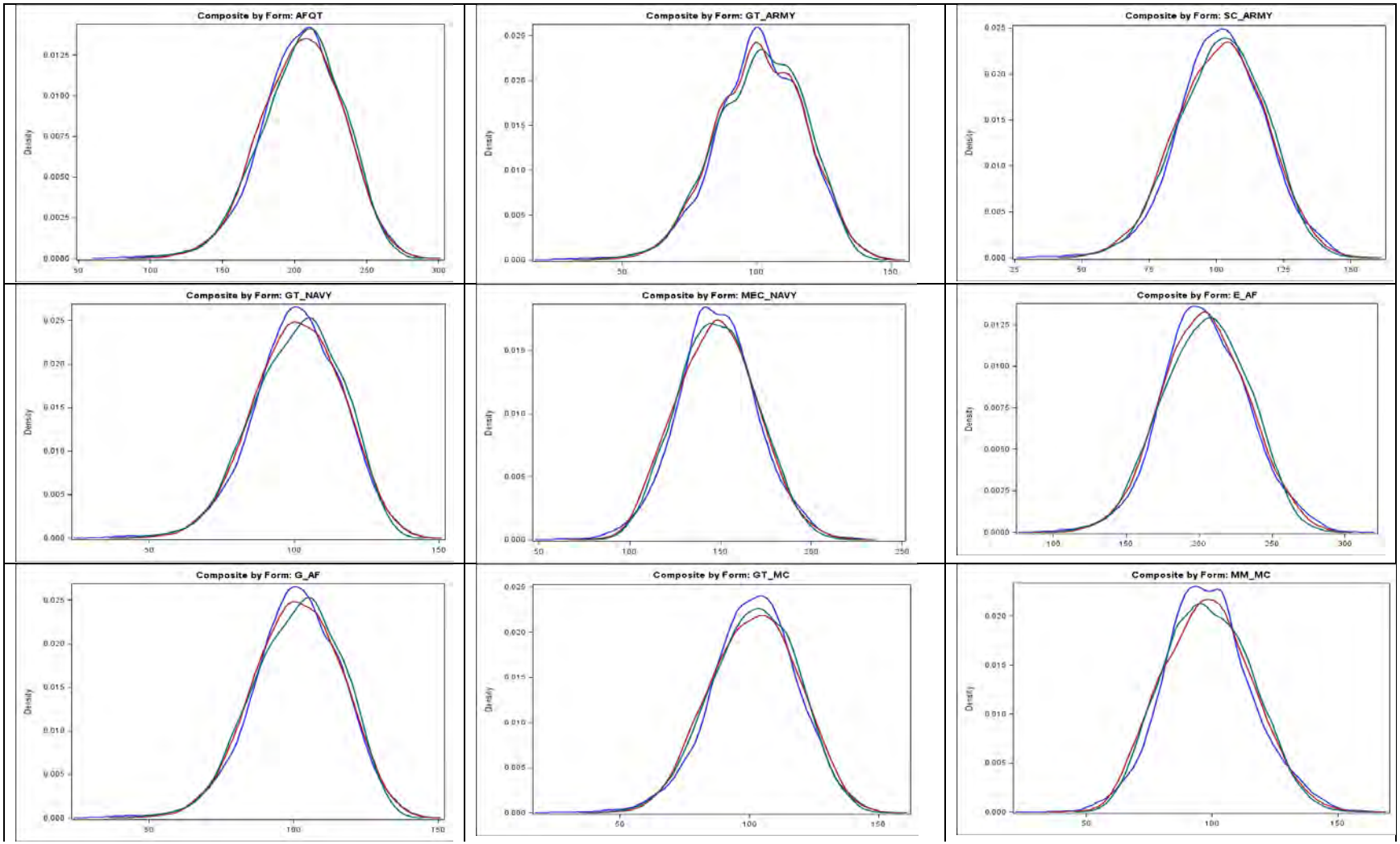






Form 4 Form 5 Form 10

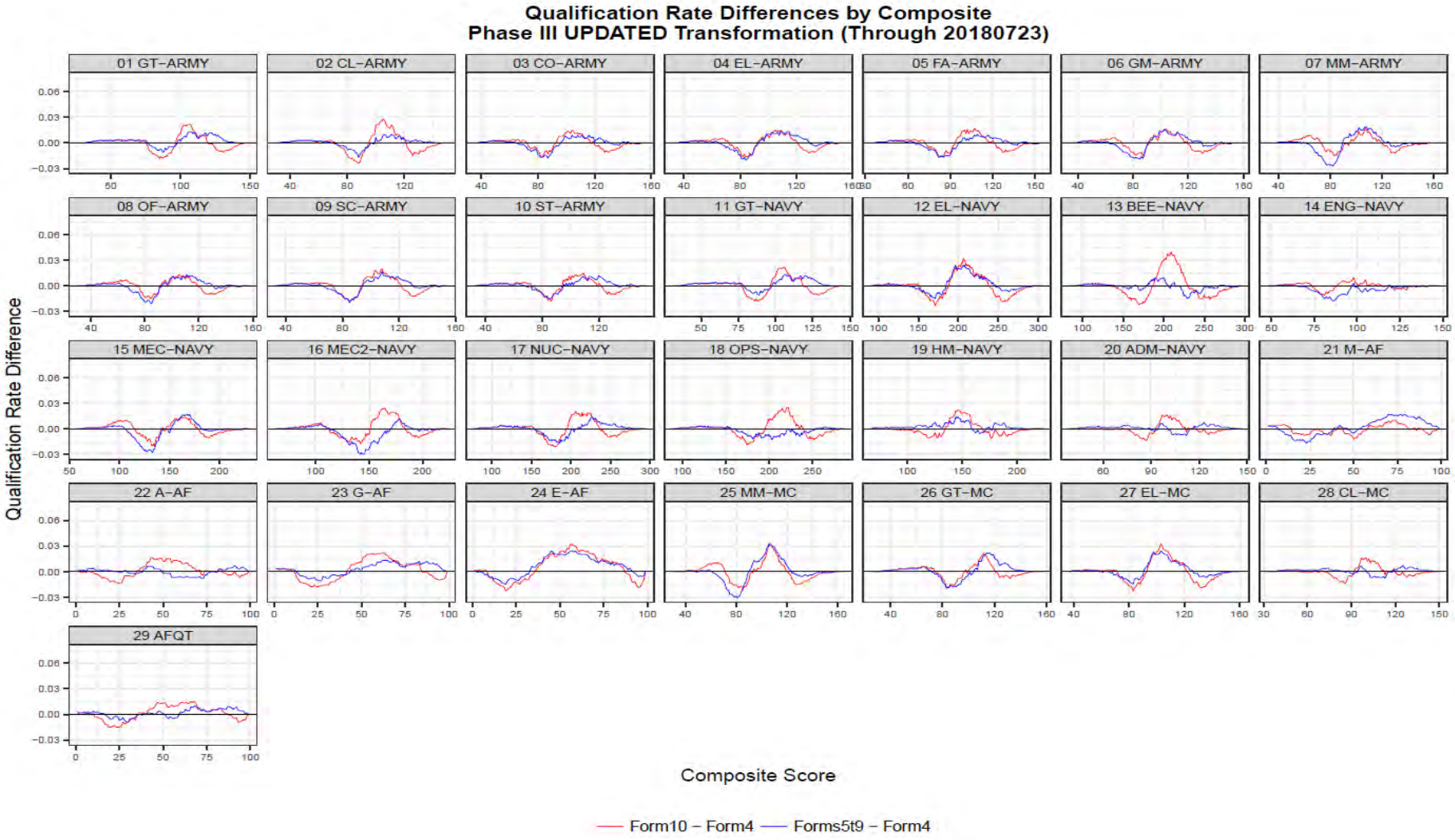
ASVAB Equating: Composite Distributions



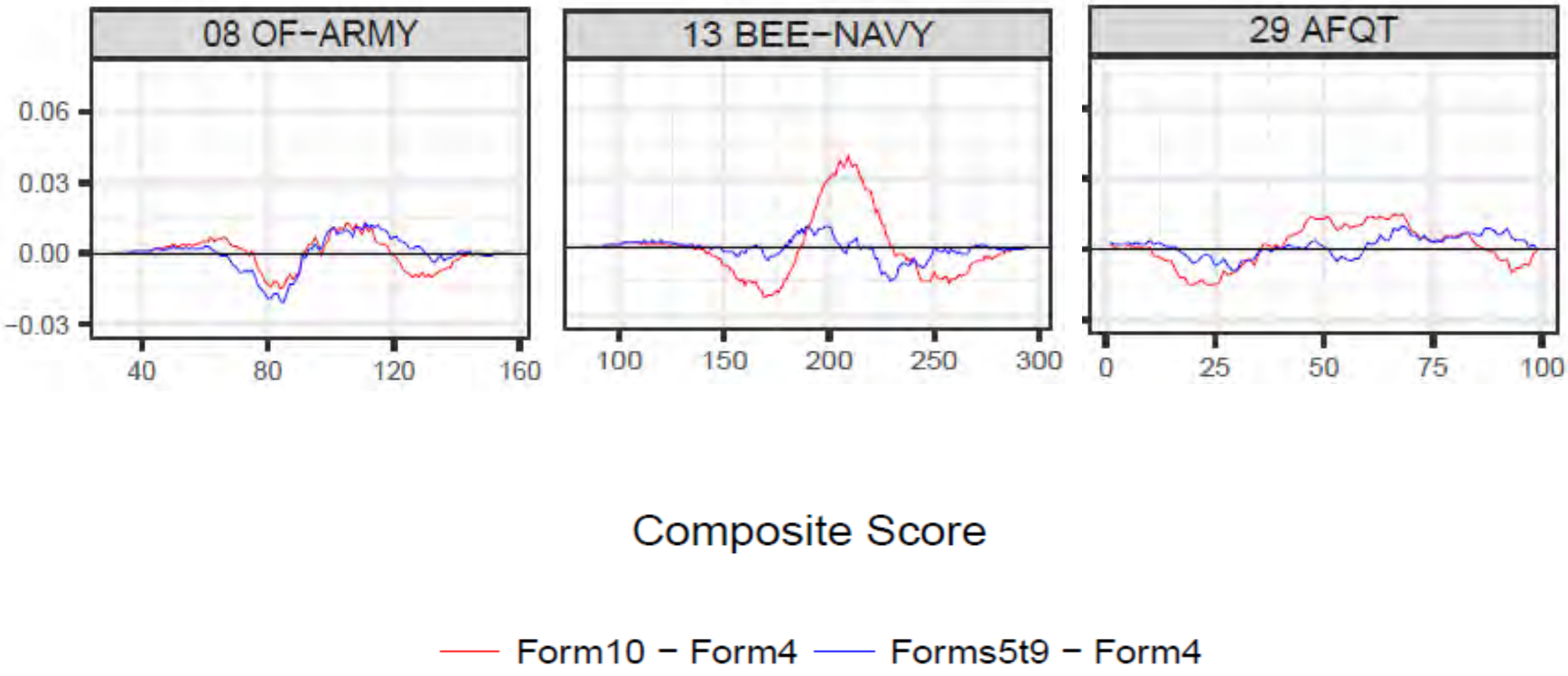
Form Composite Equivalence Analysis

- Composites can have different variances if the forms display different pattern of subtest correlations
- Evaluate differences in cumulative distribution functions
  - Kolmogorov-Smirnov (K-S) test
  - Cumulative Distribution Function (CDF) for reference group minus CDF for new form group

Composite Distribution Equivalence



Composite Distribution Equivalence



Subgroup Performance

- Do subgroups perform at the same levels across forms?
  - Females
  - Blacks
  - Hispanics
- Compare subgroup performance across new and reference forms
  - One-way ANOVA with groups defined by form
  - Statistical significance and effect size
- Analysis of form 10
  - Two statistically significant differences in female analysis
    - AS and VE
    - Small effect sizes ($\delta = 0.11, 0.08$, respectively)
  - No other statistically significant differences observed

Operational Form Comparison

- How do equated scores on form 10 compare to operational forms?
  - Compared mean differences in form 4 to forms 5, 6, 8, 9
  - Statistically significant mean differences, representing small effect sizes, in several tests
    - Form 5
      - MK ($\delta = 0.05$)
      - AO ($\delta = 0.07$)
      - AS ($\delta = 0.08$)
    - Form 6
      - EI ($\delta = -0.07$)
      - AO ($\delta = 0.07$)
      - VE ($\delta = -0.05$)
    - Form 8
      - VE ($\delta = -0.06$)
    - Form 9
      - GS ($\delta = -0.11$)
      - AO ($\delta = 0.07$)
      - VE ($\delta = -0.05$)
- No differences observed in remaining tests or AFQT

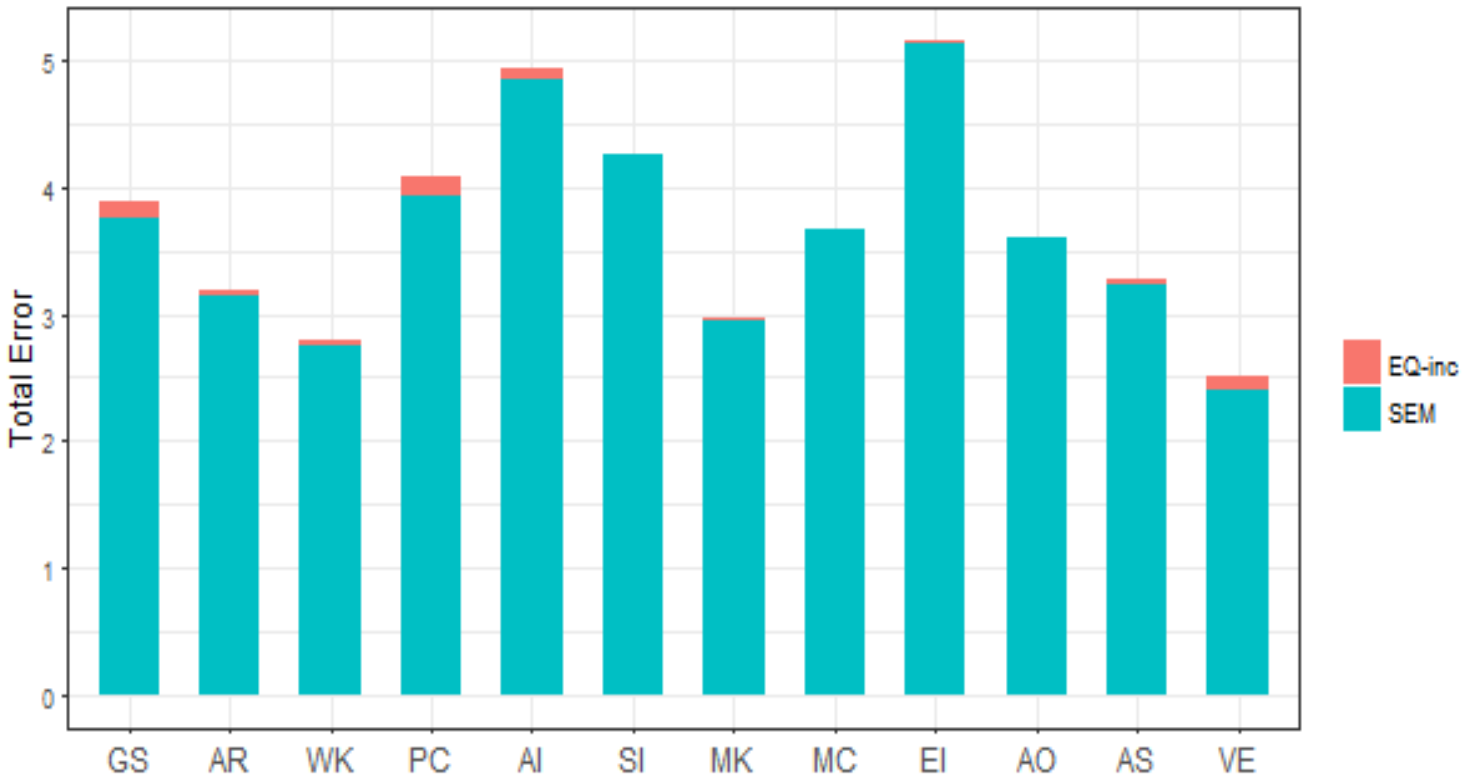
Provisional Equating Transformation Accuracy

- How closely did the provisional equating transformations match the final?
  - How different are scores based on provisional constants from what they would have been if based on final constants?
- Rescore all applicants who took form 10 using final transformation constants
  - Compare rescored values to those used operationally based on provisional constants
- Calculate total errors as the sum of equating errors and measurement errors
- Compare total error with standard errors of measurement

Provisional Transformation Accuracy

Subtest	RMSD	Bias	σ_δ	σ_E	σ_E^*
GS	1.18	0.51	1.06	3.75	3.90
AR	1.03	0.84	0.59	3.14	3.19
WK	0.95	0.86	0.41	2.76	2.79
PC	1.31	0.67	1.13	3.94	4.09
AI	1.50	1.20	0.89	4.85	4.93
SI	1.05	-1.01	0.29	4.25	4.26
MK	0.92	-0.83	0.38	2.95	2.97
MC	1.10	1.04	0.36	3.66	3.68
EI	0.41	0.02	0.41	5.15	5.16
AO	1.00	0.93	0.36	3.60	3.61
AS	0.57	0.13	0.55	3.23	3.27
VE	1.09	0.83	0.71	2.40	2.50

Provisional Transformation Accuracy



Questions?

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

HumRRO Team

- Adam Beatty
- Ted Diaz
- Amanda Koch
- Peter Ramsberger
- Matthew Reeder
- Matthew Trippe

Technical Appendix

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

Provisional Equating Transformation Accuracy

- Compute difference between Provisional and Final equating scores for each examinee $j = 1, \ldots, N$ and subtest $s = GS, AR, WK, PC, AI, SI, MK, MC, EI, AO, AS, VE$:

$$\delta_{s,j} = e_P(\hat{\theta}_{s,j}) - e_F(\hat{\theta}_{s,j})$$

- Compute RMSD, bias, and standard error of equating:

For each subtest s:

$$RMSD_s = \left(\frac{1}{N} \sum_{j=1}^{N} \delta_{s,j}^2 \right)^{1/2}$$

$$\overline{\delta}_s = \frac{1}{N} \sum_{j=1}^{N} \delta_{s,j}$$

$$\sigma_{\delta_s} = \left(\frac{1}{N} \sum_{j=1}^{N} \left(\delta_{s,j} - \overline{\delta}_s \right)^2 \right)^{1/2}$$

- Compute total error as sum of equating error and measurement error:

$$\sigma^*_{E,X,s} = \sqrt{\sigma^2_{\delta_s} + \sigma^2_{E,X,s}}$$

Provisional Equating Transformation Accuracy

- Compute average standard error of measurement for penalized theta estimate:

For subtest $s = GS, AR, WK, PC, AI, SI, MK, MC, EI, AO$

$$\sigma_{E,s} = \left[\frac{\sum^{k} w_{\theta,k}\, I(\theta_k, s)^{-1}}{\sum^{k} w_{\theta,k}}\right]^{1/2}$$

where $I(\theta_k, s)$ is the approximate theta score information function.

- Compute average standard error of measurement for standard score:

For subtest $s = GS, AR, WK, PC, AI, SI, MK, MC, EI, AO$:

$$\sigma_{E,X,s} = \sigma_{e_F,s}\sigma_{E,s}$$

where $\sigma_{e_F,s}$ is the standard deviation of the final updated standard score for subtest s.

For AS and VE:

$$\sigma_{E,X,VE} = \frac{1}{2}\sqrt{\sigma^2_{E,X,WK} + \sigma^2_{E,X,PC}}$$

$$\sigma_{E,X,AS} = \frac{1}{2}\sqrt{\sigma^2_{E,X,AI} + \sigma^2_{E,X,SI}}$$

- Compute total error as sum of equating error and measurement error:

$$\sigma^*_{E,X,s} = \sqrt{\sigma^2_{\delta_s} + \sigma^2_{E,X,s}}$$

Tab J

Cyber Test
Item & Form Development

Presenters : **D. Matthew Trippe**

September 20, 2018

Background

- Development of the Joint Service Cyber Test (CT), formerly known as the Information and Communications Technology Literacy (ICTL) test, began over 10 years ago
- The CT is modeled after ASVAB “information” tests (e.g., Electronics, Auto, Shop Information) and is administered on the ASVAB platform as a linear (i.e., non-adaptive) test
- The CT has demonstrated evidence as a valid predictor of training success in several Air Force, Navy, and Army technology-related occupations
- Like any selection test, the CT requires periodic maintenance to review and refresh the item pool

Overview

- Situation
  - The Air Force developed 251 new Cyber Test items in 2015. These items, along with the prior pool of items (*n* = 167), are intended to be transitioned to a computerized adaptive testing (CAT) platform.
  - The item pool comprises many items that provide information on relatively high-ability applicants.
- Objectives
  - Review, calibrate, and equate the items to the current operational scale, and replace the static forms with two CAT forms/pools
  - Develop 200 new items targeted toward the middle and low end of the ability distribution

Primary Tasks

- Technical review of experimental items
- Equating
- Form assembly
- New Item development
  - Blueprint validation
- Documentation

Technical Review of Experimental Items

- Review of experimental item quality
- 243 experimental items administered at MEPS
- Items screened in similar manner to ASVAB experimental items
- Empirical evidence
  - CTT (p-value, item-total, option-total *r*)
  - IRT (three-parameter logistic model; 3PL)
- SME content review guided by item statistics
  - SMEs evaluate items with potential issues identified by distractor analyses or other empirical guidance
- Based on large applicant sample
  - Total *n* = 84,988
  - Per item *n* = 3,386 (average)
- 117 (48%) items retained from screening process
  - Relatively low "survival" rate
  - Item difficulty in relation to applicant population remains primary factor in experimental item loss
  - Many content areas are inherently complex or technical

Equating & Form Assembly

- 117 new items placed on operational scale established in 2011 via Stocking & Lord (1983) procedure
- All viable CT items (*n*=284) included in form assembly
  - Automated Test Assembly (ATA) as described in van der Linden (2005)
  - Binary/integer programming approach
  - Form specifications are set up as quantities to be minimized (e.g., TCC between forms) or maximized (e.g., score information) against an objective function
  - Additional constraints such as item enemies, content % specifications, keyed response, are incorporated into the model
  - Goal is two parallel CAT forms

New Item Development

- Develop 200 new items targeted toward the middle and low end of the ability distribution
  - SMEs provided feedback on empirical difficulty of items they wrote in 2015
- Items of “easy” and “moderate” difficulty are challenging to write in this content area
- Information is concentrated at the high end of the ability distribution
- Prior to developing each new set of experimental items, we conducted a blueprint “validation” or review
- CT blueprint
  - Determines test item content
  - Comprises 40-50 knowledge, skill, ability (KSA) statements organized into four broad content areas
    - Individual KSA statements serve as stimulus for development of new items
    - For example: “Knowledge of network addressing concepts,” “Knowledge of operating system internals”

Blueprint Validation

- Assemble SMEs from Services to review the CT blueprint for relevance or potential obsolescence
- SMEs provided with
  - KSA statements in current CT blueprint (*n* = 49)
  - KSA statements from National Initiative for Cybersecurity Education (NICE) framework (*n* = 61)
- In standardized rating exercise, SMEs rate
  - Should this KSA be acquired prior to enlistment?
  - How important is this KSA for successful performance in entry-level training for enlisted cyber occupations?
  - Given ongoing technological change, how stable do you think this KSA will be over time?
  - Any KSA statements missing?
- Results in minor/marginal updates to
  - knowledge, skill, ability (KSA) statements included
  - content area weighting

Blueprint KSAs

Knowledge, Skills, and Abilities on Blueprint	2008	2011	2015	2018
Total Number of KSAs	39	46	49	41
KSAs Retained from Previous Blueprint	n/a	39	33	41
KSAs Dropped from Previous Blueprint	n/a	0	7	8
New KSAs Added	n/a	7	16	0

Blueprint Category Weights

Category	2008 Weight	2011 Weight	2015 Weight	2018 Weight
Networking and Telecommunications	25%	35%	30%	30%
Computer Operations	35%	35%	30%	30%
Security and Compliance	25%	20%	25%	25%
Software Programming and Web Design	15%	10%	15%	15%

Schedule/Status

- Technical review: complete
- Equating: in progress
- Form assembly: anticipated by October 2018
- New item development
  - 200 items written
  - Editorial review complete
  - Technical review complete
  - Anticipated completion by November 2018
- Documentation
  - Technical report
  - Anticipated completion by December 2018

Questions/Discussion?

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

HumRRO Team

- Adam Beatty
- Chris Huber
- Amanda Koch (project manager)
- Oren Shewach
- Matthew Trippe

Tab K

OPA

OFFICE OF PEOPLE ANALYTICS

Sparse Data Dimensionality Assessment

with Application to the Cyber Test Data

Furong Gao, HumRRO

Defense Advisory Committee on Military Personnel Testing
September 20-21, 2018
Minneapolis, Minnesota

OUTLINE

- **Background information**
- **Dimensionality and IRT models**
- **Dimensionality assessment approaches**
- **Application to Cyber Test data with seeded items**
  - Confirmatory analysis of unidimensionality
- **Conclusions and Discussions**

OPA
OFFICE OF PEOPLE ANALYTICS

BACKGROUND INFORMATION

- **IRT model fit**
  - Affect the accuracy of item parameter estimation, test scores, and classification of the test takers
- **CAT item selection in ASVAB tests**
  - Current algorithm assumes unidimensionality for a given test
  - Without content constraint
- **Concerns raised by the DAC members**
  - Potential content or item difficulty shift in continually developed new items
  - CAT item-rendering algorithm
    - Content constraints

DIMENSIONALITY AND IRT MODELS

- ***Strictly* unidimensional test is theoretical in nature and doesn't exist in practice**

- **Tests that are carefully constructed to measure only a single dimension (construct) often show one or more minor dimensions.**
  - Essential unidimensionality: a unidimensional model will adequately represent the test data

- **Essentially unidimensional tests usually display a bi-factor structure**
  - The intended dimension/construct with items from different content domains
  - Items in each content domain measure a secondary (minor) dimension

DIMENSIONALITY ASSESSMENT APPROACHES WITH COMPLETE DATA

- **Item score matrix *X*: *N* x *n***
  - *N:* number of examinees
  - *n:* number of items
  - x_{ij}: item score of the i^{th} examinee on the j^{th} item

- **Complete data: *X* has few or no missing data points**

- **Dimensionality assessment:**
  - Covariance $\boldsymbol{\Sigma}$-based
    - Classical factor analyses
    - DIMTEST, DETECT
  - IRT model-based
    - Item factor analysis
      - TESTFACT

OPA
OFFICE OF PEOPLE ANALYTICS

DIMENSIONALITY ASSESSMENT APPROACHES
MISSING DATA IN SEEDED ITEM COLLECTION—SPARSE DATA

- **For seeded items, each examinee gets a small set of items from the seeded (experimental) item pool, resulting in a sparse data matrix *X***

- **Missing pattern:**
  - Missing at Random (MAR)
    - The “missingness” of an item score is not related to the missing value, but related to some of the observed data.
  - Missing Completely at Random (MCAR)

DIMENSIONALITY ASSESSMENT APPROACHES WITH SPARSE DATA

- **Covariance Σ-based approaches will not work anymore**

- **Full information maximum likelihood (FIML) estimation**
  - Using only the observed data without direct imputation of the missing values
  - Under MAR or MCAR, produce unbiased estimates

- **Factor analysis**
  - Software: R-package lavaan (Rosseel, 2012)
    - FIML and EM

- **IRT model-based item factor analysis**
  - Software: iFACT (Segall, 2002)
    - MCMC
    - Application to the Cyber Test data with seeded items

Yves Rosseel (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48(2)*, 1–36.

Segall, D. (2002). iFACT computer program Version 2.0: Full information confirmatory item factor analysis using Markov chain Monte Carlo estimation "Computer Program." Seaside, CA: Defense Manpower Data Center

OPA
OFFICE OF PEOPLE ANALYTICS

DIMENSIONALITY ASSESSMENT APPROACHES WITH SPARSE DATA – CONT'D

- **Assumptions**
  - Test is designed to be unidimensional: measure a single construct but with broad content coverages that may introduce minor additional unintended dimensions to the test data
  - Items are rendered in a way 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 MAR 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

DIMENSIONALITY ASSESSMENT APPROACHES WITH SPARSE DATA – CONT'D

- **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 do not distort the meaning of the general factor/dimension that is measured generally by all the items on the test.

- **Explained common variances (ECV)**
  - An indicator of essential unidimensionality
  - Calculated using the factor loading values of the G factor and the secondary factors of the bi-factor model

EXPLAINED COMMON VARIANCE (ECV)

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

- **Where**
  - λ_g are the factor loadings on the G factor; the summation is over all items on the test
  - λ_{sj} *(j = 1, …, k)* are the factor loadings on the *k* secondary factors
    - Each item loads on only one secondary factor
    - The summation is over all the items on that secondary factor

- **Value is between 0 and 1**
  - The larger the ECV, the stronger the unidimensionality
    - 0.9 < ECV, essentially unidimensional
    - 0.7 <= ECV <= 0.9, additional information should be used (subscore, etc.)
    - ECV < 0.7, evidence of multidimensionality
  - Strictly unidimensional: ECV = 1

- **To adjust for the standard error of estimates**

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

OPA
OFFICE OF PEOPLE ANALYTICS

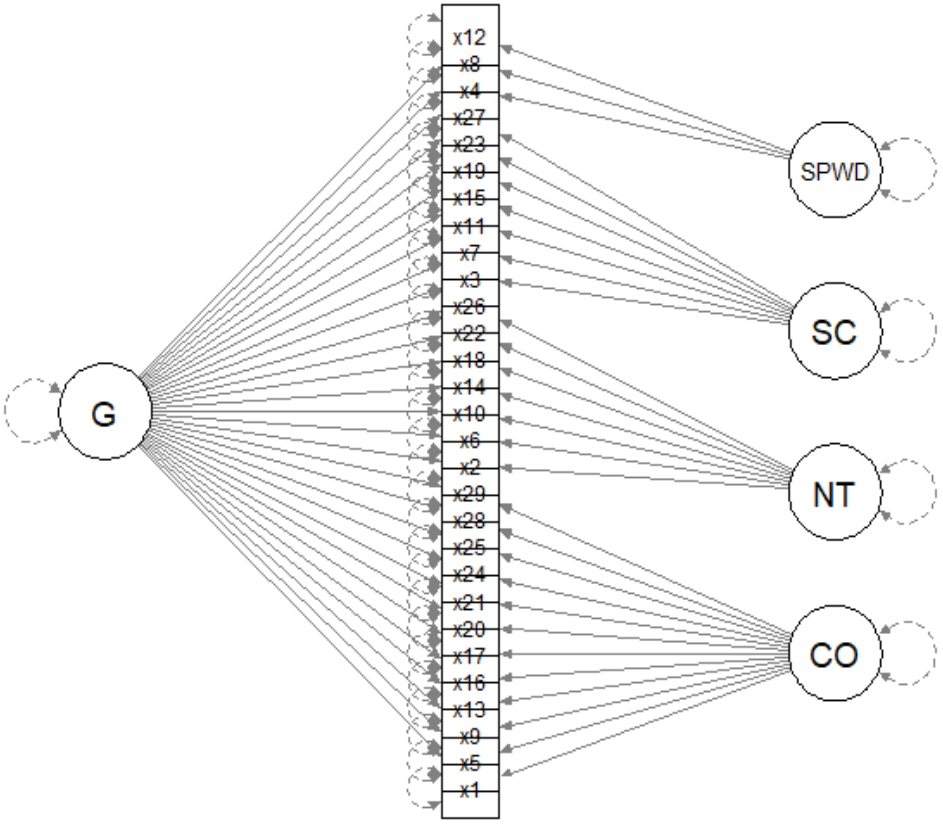
CYBER TEST

- **A test of information and communications technology literacy**

- **Four broad content areas**
  - Computer operations (CO)
  - Networks and Telecommunications (NT)
  - Security and compliance (SC)
  - Software programming & Web development (SPWD)

A BI-FACTOR MODEL FOR THE CYBER TEST

- **One general factor (G)**
- **Four secondary factors, one for each content area: CO, NT, SC, and SPWD**
- **All the factors are independent of each other**
- **Item 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: $(\lambda_g, \ 0, 0, 0, \lambda_{co})$



OPA
OFFICE OF PEOPLE ANALYTICS

DIMENSIONALITY ASSESSMENT: CYBER TEST FORM1 DATA

- **Cyber Test Form1**
  - 29 items
  - Number of items on each of the four content areas/sub-domains:
    - CO 12
    - NT 7
    - SC 7
    - SPWD 3

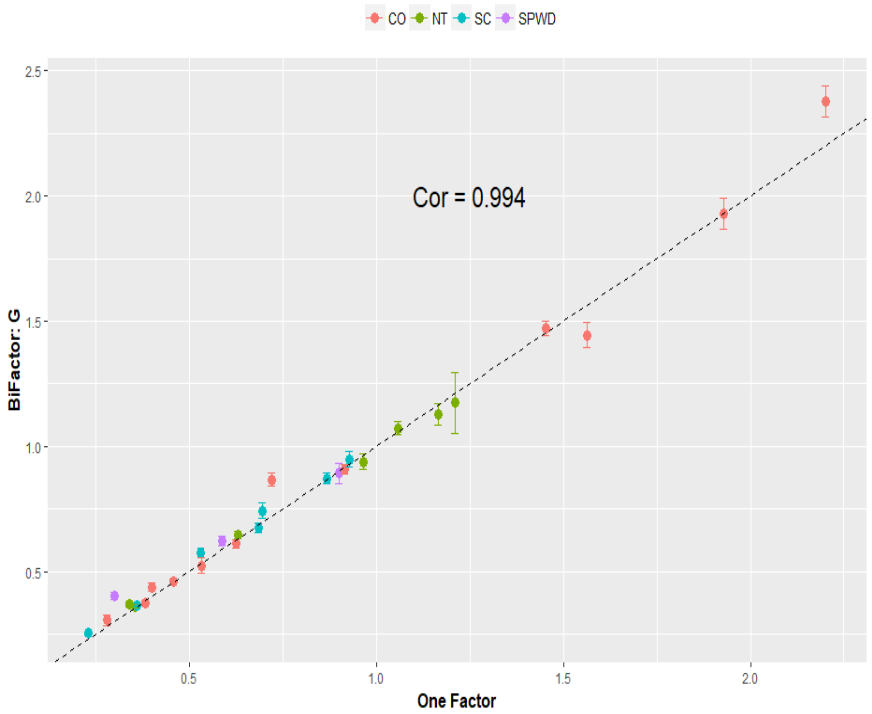
- **Total of 65,289 test-takers**

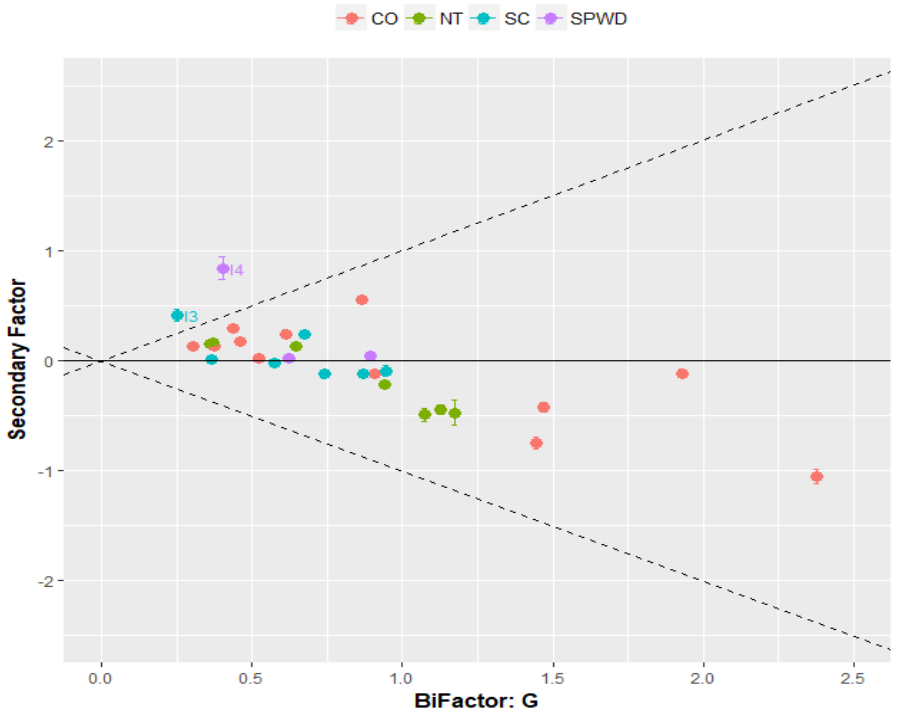
- **iFACT**
  - 2,000 burn-in cycles
  - 2,000 additional cycles for posterior summarization

OPA
OFFICE OF PEOPLE ANALYTICS

IRT MODEL-BASED ASSESSMENT—IFACT

- **Form1: ECV = 0.865; ECV.adj = 0.866**





OPA
OFFICE OF PEOPLE ANALYTICS

FORM1 + 117 SEEDED ITEMS—IFACT RESULTS

- **Seeded items**
  - 190 seeded items
  - 73 excluded from the previously evaluated CAT pool due to undesirable psychometric quality; also excluded from the dimensionality analyses here

- **146 items**
  - 29 From1 items
  - 117 seeded items
  - Number of items in each content area:
    - CO 59
    - NT 45
    - SC 31
    - SPWD 11

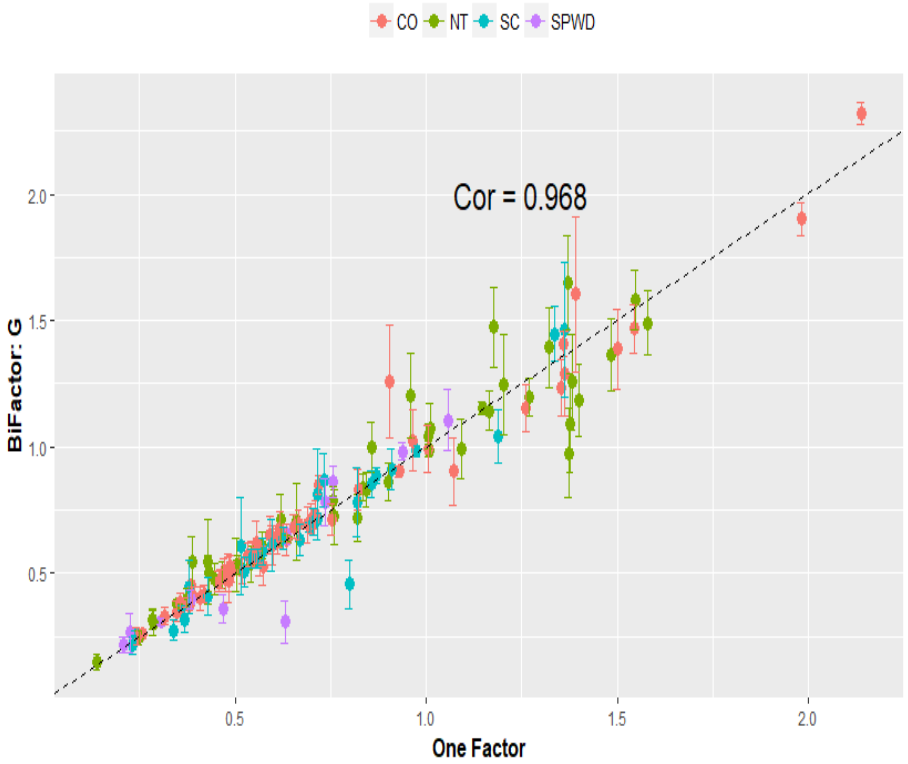
- **65,289 test-takers**

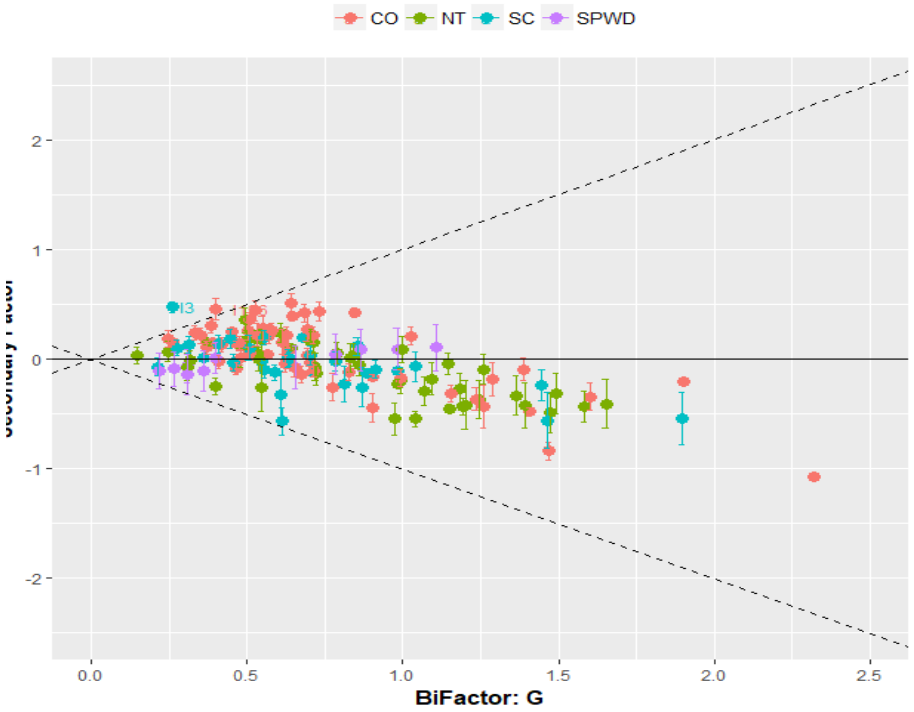
- **Case count on each of the 117 seeded items ranges from 3,050 to 3,978**

OPA
OFFICE OF PEOPLE ANALYTICS

FORM1 + 117 SEEDED ITEMS—IFACT RESULTS

- Form1 + 117 seeded items: 146 items
- ECV = 0.909; ECV.adj = 0.922

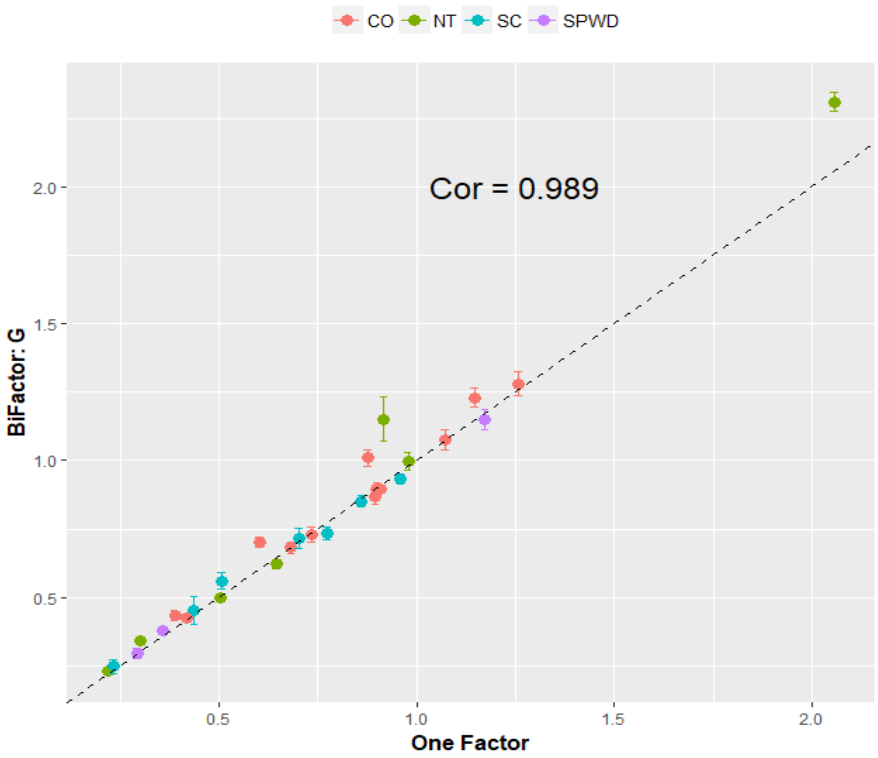


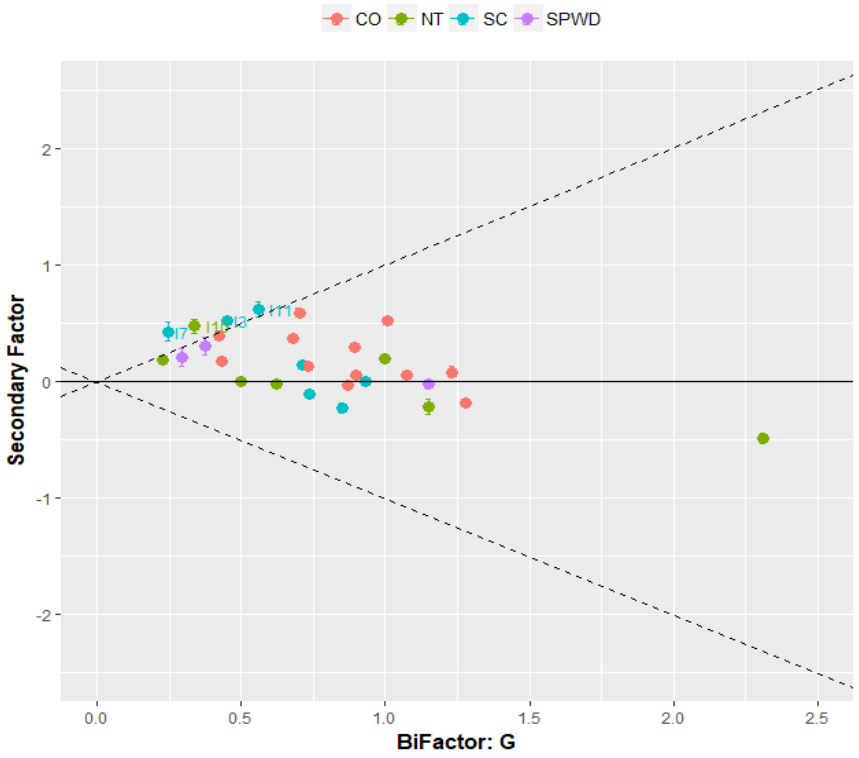


OPA
OFFICE OF PEOPLE ANALYTICS

FORM2 DATA—IFACT RESULTS

- Form2: 68928 cases
- ECV = 0.893; ECV.adj = 0.895





OPA
OFFICE OF PEOPLE ANALYTICS

FORM1 + FORM2 + SEEDED: 175 ITEMS

- **Total 175 items**
  - 29 From1 items
  - 29 Form2 items
  - 117 seeded items
  - Number of items on each content area:
    - CO 71
    - NT 52
    - SC 38
    - SPWD 14

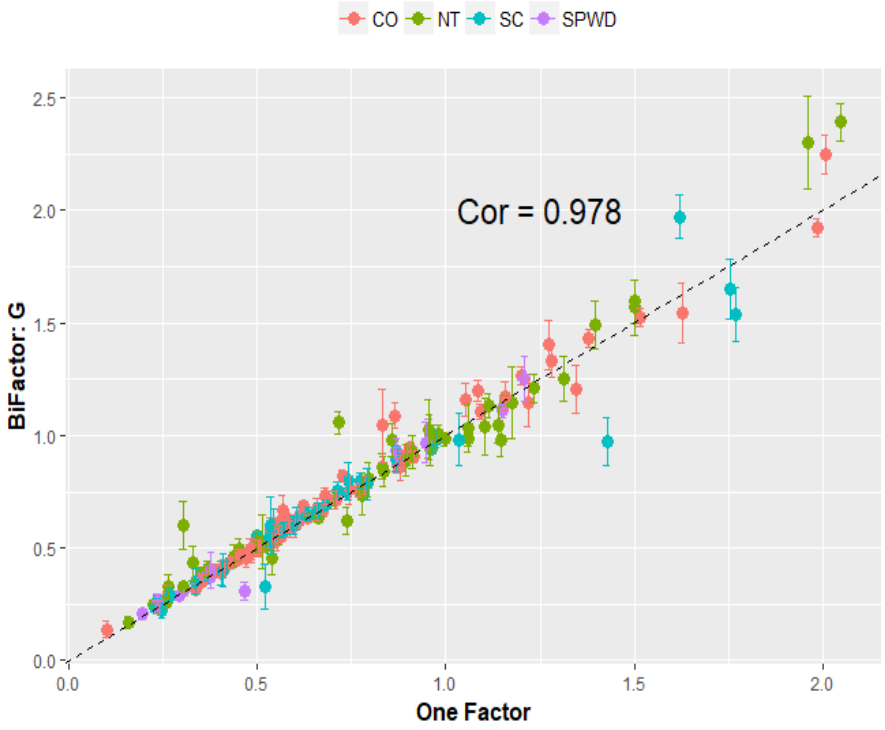
- **65,289 test-takers on Form1, 68,928 on Form2**

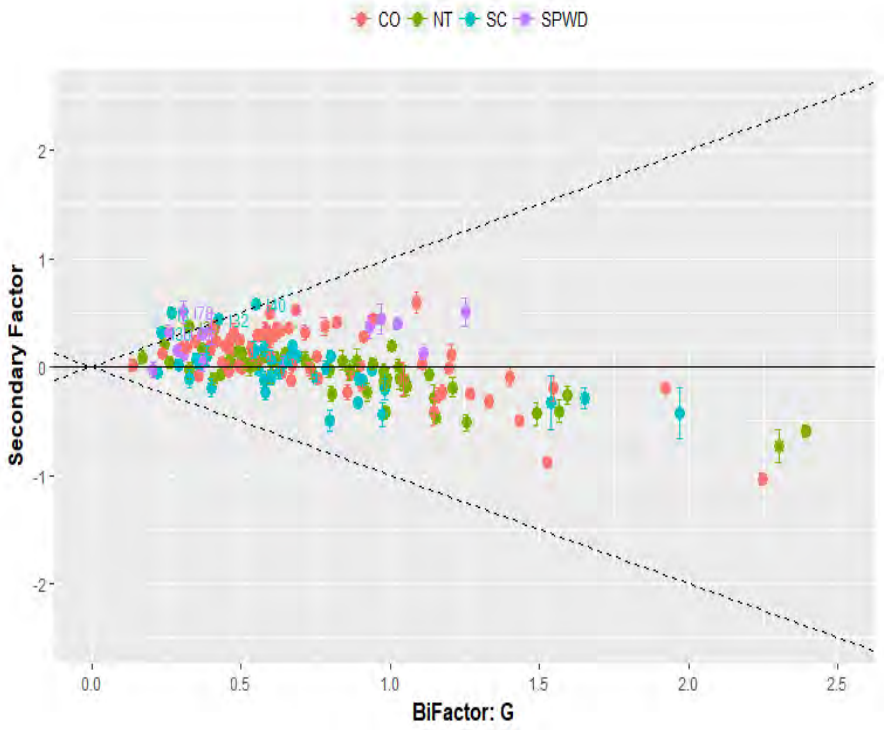
- **Case counts on seeded items range from 6,493 to 7,678**

OPA
OFFICE OF PEOPLE ANALYTICS

IFACT RESULTS

- Fom1 + Form2 + 117 seeded items: 175 items
- ECV = 0.911; ECV.adj = 0.916





OPA
OFFICE OF PEOPLE ANALYTICS

CONCLUSIONS AND DISCUSSIONS

- **Well-constructed unidimensional tests often display a bi-factor structure with one dominant general factor and minor secondary factors that are negligible**
- **These tests are essentially unidimensional, and the response data can be adequately modeled/explained by unidimensional IRT models**
- **An example:**
  - The Cyber Test with seeded item collection
- **Future further analyses**
  - On CAT-ASVAB data: simulated and operational data
  - Continue to monitor potential content or item difficulty shift

OPA

OFFICE OF PEOPLE ANALYTICS

Appendix

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

Tab L

Technical Expert Panel on the Use of Non-Cognitive Measures

Status Update

Presented to: Defense Advisory Committee (DAC)

September 20, 2018

Presenters : Tim McGonigle, HumRRO

Agenda

- Background on TAPAS
- Overview and Goals of the TEP
- Process for Selecting TEP Members
- Introduction of the TEP Members
- Potential Research Topics
- Next Steps and Intended Schedule

Background on TAPAS

- Tailored Adaptive Personality Assessment System (TAPAS)
  - Originally developed by ARI and Drasgow Consulting Group (DCG) to measure up to 27 facets of the Big Five personality dimensions
  - Uses multidimensional pairwise preference (MDPP) items
    - Generally presents two statements from different personality dimensions
    - Matched on the strength of the dimension and on the socially-desirable nature of the response options
    - Intended to make it more difficult to fake because the "correct" answer is difficult to identify
  - Items generated on-the-fly by selecting from pools of pre-calibrated personality statements that measure construct dimensions relevant to performance in the military; approximately 1M statement combinations possible
  - Scored using multi-unidimensional pairwise preference IRT (ideal point) model
- Army, Navy, Air Force, and Marines have all collected TAPAS data on applicants
  - Evidence of incremental validity beyond ASVAB for training and military success criteria (e.g., attrition)

Which of these statements is the most like you?

- People come to me when they want fresh ideas
- Most people would say I am a "good listener"

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
    - Analyzed data from candidates who completed TAPAS between March 2010 and April 2015 and subsequently completed at least six months of service
    - Found small, significant incremental validity over education credential in predicting attrition
    - Found low test-retest reliability in some conditions
      - r_{xx} = 0.07 (TAPAS 9/10/11, Army recruits who failed first test)
    - But not as low under other conditions
      - r_{xx} = 0.59 (TAPAS 5/7/8, Air Force all recruits)
- DPAC requested the establishment of a Technical Expert Panel (TEP) to independently review the body of TAPAS research and make recommendations regarding the readiness of TAPAS for operational use
  - Review related research conducted by the services, both on TAPAS and on other instruments (e.g., interest inventories)
  - Make recommendations for future research and development
  - Comment on the readiness of TAPAS for operational use

Process for Selecting TEP Members

- TEP should include five experts whose research and practice have involved personality measurement and the use of non-cognitive measures for selection
- Balance previous involvement with TAPAS
  - TEP should bring both fresh perspectives and familiarity with TAPAS research
- Developed five criteria for recruiting TEP members. The overall panel should include members with
  - Familiarity with TAPAS research and development
  - An independent perspective on personality measurement
  - Knowledge of psychometrics and IRT, particularly ideal point IRT modeling with forced-choice pair-wise comparisons
  - Gender/race diversity
  - Academic/practitioner diversity

Process for Selecting TEP Members

- DPAC and HumRRO independently identified potential candidates (N = 35)
- Grouped them into four categories
  - National-level testing experts (N = 1)
  - Psychometrics experts (N = 2)
  - Personality theory experts (N = 1)
  - Operational testing experts (N = 1)
- Used criteria to prioritize the candidates
- Recruited members top down within category
- All top choices have agreed to join TEP

Introduction of the TEP Members

TEP Member	Title	Affiliation	Primary Area(s) of Expertise			
			TAPAS	Ideal Point IRT	Personality Theory	Operational Testing
Paul Sackett	Beverly and Richard Fink Distinguished Professor of Psychology and Liberal Arts	University of Minnesota	✓		✓	
Mark Reckase	University Distinguished Professor Emeritus	Michigan State University	✓	✓		
James Roberts	Associate Professor	Georgia Institute of Technology		✓		
Winfred Arthur	Professor	Texas A&M University			✓	
April Zenisky	Research Associate Professor and Director of Computer-Based Testing Initiatives	University of Massachusetts, Center for Educational Assessment				✓

Paul Sackett

Beverly and Richard Fink Distinguished Professor of Psychology and Liberal Arts, University of Minnesota

- Relevant areas of research/practice
  - Role of personality in personnel selection, including effects of "instructed" versus "natural" faking (Ellingson, Sackett, & Hough, 1999; Ellingson, Smith, & Sackett, 2001)
  - Need for methodological rigor and psychometric sophistication in evaluating personnel decision making (Sackett & Larson, 1990)
  - Legal, psychometric, and philosophical perspectives on tension between maximizing job performance and maximizing diversity in selection systems (Sackett, Borneman, & Connelly, 2008; Sackett, Schmitt, Ellingson, & Kabin, 2001; Sackett & Wilk, 1994)
- National commissions, professional committees, and advisory boards
  - Member, Department of Defense Advisory Committee on Military Testing
  - Chair, National Research Council Committee on Physical, Medical, and Mental Health Standards for Military Recruitment
  - Member, Committee on the Revision of the *Principles for the Validation and Use of Personnel Selection Procedures*
  - Member, Joint Committee for the Revision of the *Standards for Educational and Psychological Testing*
  - Member, College Board SAT Psychometric Panel

Mark Reckase

University Distinguished Professor Emeritus, Michigan State University

- Relevant areas of research/practice
  - Unidimensional and multidimensional item response theory (IRT) models (e.g., Reckase, 2009; Reckase, Ackerman, & Carlson, 1988)
  - Computerized adaptive testing (e.g., Reckase, 2010; Reckase & McKinley, 1991)
  - Operational testing experience at ACT (Assistant Vice President of Assessment Programs, 1984-1991; Assessment Innovations, 1991-1998)
- National commissions, professional committees, and advisory boards
  - Member, Department of Defense Advisory Committee on Military Testing
  - Member, Expert Panel for Tier One Performance Screen (TOPS) IOT&E
  - Member, Technical Advisory Committee for National Assessment of Educational Progress (NAEP)
  - Member, Technical Advisory Committee for Computerized Adaptive General Aptitude Test Battery (GATB)

James Roberts

Associate Professor of Psychology, Georgia Institute of Technology

- Relevant areas of research/practice
  - Development and application of a family of unfolding item response theory (IRT) models to measure psychological constructs (e.g., Roberts, 2016; Roberts, Donoghue, & Laughlin, 2000; Roberts & Thompson, 2011)
    - Unfolding models imply higher item scores to the extent that an individual is located close to an item on a unidimensional latent continuum (similar to ideal point models)
    - Can be used to assess satisfaction, preference, personality and individual differences
    - Current research extends unfolding models to the multidimensional domain where an individual is expected to endorse an item to the extent that the individual is close to it in a latent space
  - Created GGUM2004 software for analyzing generalized graded unfolding model data

Winfred Arthur

Professor of Psychology, Texas A&M University

- Relevant areas of research/practice
  - Personality measurement (e.g., Arthur, Glaze, Villado, & Taylor, 2010; Arthur, Woehr, & Graziano, 2001)
  - Methodological issues in testing, assessment, selection, validation (e.g., Arthur & Glaze, 2011; Arthur & Villado, 2008)

- National commissions, professional committees, and advisory boards
  - Member, Technical Advisory Committee, Association of American Medical Colleges
  - Member, Technical Advisory Committee, State Department Board of Examiners for the Foreign Service
  - Member, Committee on the Revision of the *Principles for the Validation and Use of Personnel Selection Procedures,* Society for Industrial and Organizational Psychology
  - Chair, Committee on Psychological Tests and Assessment, American Psychological Association

April Zenisky

Research Associate Professor and Director of Computer-Based Testing Initiatives for the Center for Educational Assessment, University of Massachusetts

- Leads Center's research studies for large testing programs; manages psychometric activities for computerized adult education assessments; evaluates testing practices and policies
- Recent testing programs include
  - National Council of State Boards of Nursing
  - National Board of Professional Teaching Standards
  - National Science Foundation
  - American Chemical Society
- Extensive publication record with emphasis on operational decisions in computer-based testing programs (e.g., Hambleton & Zenisky, 2011; Sireci & Zenisky, 2006; Zenisky & Hambleton, 2013; Zenisky & Luecht, 2016)

Research Support

- HumRRO will provide research support to TEP
  - Focus will be mostly on synthesizing existing research
- Services have done lots of research on TAPAS
  - Requesting assistance to identify a comprehensive set of existing research
- Potential research topics
  - Meta-analysis of TAPAS validity across Services
  - Additional analysis of the test-retest reliability of TAPAS, such as examining effect of re-test interval
  - Effects of coaching, regression to the mean, random responding, and motivation on reliability
  - Research agenda for future TAPAS research
- Additional topics?

Next Steps and Intended Schedule

- Next Steps
  - Hold first meeting
    - October 22 (Atlanta, GA)
  - Identify POCs from each Service
  - Collect existing research
  - Coordinate involvement of RAND and DCG
  - Establish rules of engagement for TEP
    - Model after DAC?
- Intended Schedule
  - Four meetings, 1+ days each
    - October, January, May, August
    - Final report by October 2019

Draft Agenda for First Meeting

- Introductions
- TEP Purpose and Outcomes
- Background on TAPAS
  - TAPAS development (Steve Stark and Chris Nye, DCG)
  - RAND evaluation (Larry Hanzser, RAND)
  - Air Force faking/validity studies (John Trent, AFPC)
  - Army validity studies (Tonia Heffner, ARI)
- TEP Governance
  - Mission and goals
  - Rules of engagement
  - Prioritize research topics
  - Solicit agenda items for subsequent meetings
  - Discuss ongoing roles of Services, RAND, DCG

Questions?

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

References

- Arthur, W., Jr., & Glaze, R. M. (2011). Cheating and response distortion on remotely delivered assessments. In N. T. Tippins, & S. Adler (Eds.), *Technology-enhanced assessment of Talent* (pp. 99-152). San Francisco, CA: Jossey-Bass.
- Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93*, 435-442.
- Arthur, W., Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment, 18*, 1-16.
- Arthur, W., Jr., Woehr, D. J., & Graziano, W. G. (2001). Personality testing in employment settings: Problems and issues in the application of typical selection practices. *Personnel Review, 30*, 657-676.
- Ellingson, J. E., Sackett, P. R., & Hough, L. (1999) Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*, 155-166.
- Ellingson, J.E., Smith, D. B., & Sackett, P.R. (2001) Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology, 86,* 122-133.
- Hambleton, R. K., & Zenisky, A.L. (2011). Translating and adapting tests for cross-cultural assessments. In D. Matsumoto, & F. van de Vijver (Eds.), *Cross-cultural research methods* (pp. 46-70). Oxford, England: Oxford University Press.
- Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling, 52*(2), 127-141.
- Reckase, M. D. (2009). *Multidimensional Item Response Theory.* Springer, New York.
- Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*(4), 361-373.
- Reckase, M. D., Ackerman, T. A. & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25*(3), 193-204.
- Roberts, J. S. (2016). The generalized graded unfolding model. In W. J. van der Linden (Ed.), *Handbook of Modern Item Response Theory* (2nd edition). NY: Taylor & Francis.

References

- Roberts, J. S., & Thompson, V. M. (2011). Marginal maximum a posteriori item parameter estimation for the generalized graded unfolding model. *Applied Psychological Measurement, 35,* 259-279.
- Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3-32.
- Sackett, P. R., & Larson, J. (1990). Research strategies and tactics in I/O psychology. In M. D. Dunnette & L. Hough (Eds.) *Handbook of Industrial and Organizational Psychology (2nd)*. Palo Alto, CA: Consulting Psychologists Press.
- Sackett, P. R., & Wilk, S. L. (1994) Within-group norming and other forms of score adjustment in pre-employment testing. *American Psychologist, 49*, 929-954.
- Sackett, P. R., Borneman, M., & Connelly, B. S (2008). High stakes testing in education and employment: Evaluating common criticisms regarding validity and fairness. *American Psychologist*, 63, 215-227.
- Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative action world. *American Psychologist*, 56, 302-318.
- Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 329-348). Mahwah, NJ: Lawrence Erlbaum Associates.
- Zenisky, A. L., & Hambleton, R. K. (2013). From “Here’s the Story” to “You’re in Charge”: Developing and maintaining large-scale online test and score reporting resources. In M. Simon, M. Rousseau, & K. Ercikan (Eds.), *Improving Large-scale Assessment in Education* (pp.175-185). New York, NY: Routledge.
- Zenisky, A. L., & Luecht, R. M. (2016). The future of computer-based testing: Some new paradigms. In. C. Wells & M. Faulkner-Bond (Eds.), *Educational Measurement: From Foundations to Future* (p. 221-238). New York, NY: Guilford.

Tab M

OPA

OFFICE OF PEOPLE ANALYTICS

(Adverse) Impact of the ASVAB: Findings for Fiscal Year 2017 Applicants

Gregory Manley
Ping Yin
Mary Pommerich

Defense Personnel Assessment Center

DAC-MPT Meeting
September 20-21, 2018
Minneapolis-St.Paul, MN

WHAT IS ADVERSE IMPACT?

- ***Impact*** **can occur when groups that are not matched on ability perform differentially on an item or test.**
- ***Adverse impact*** **occurs when a group is disadvantaged by those performance differences.**
- ***Bias*** **occurs when an item or test unfairly favors one group over another.**
  - The occurrence of bias is problematic because it can negatively affect test validity.
  - The occurrence of (adverse) impact does not necessarily mean that a test is biased.

OPA
OFFICE OF PEOPLE ANALYTICS

WHO IS AFFECTED BY ADVERSE IMPACT?

- The ASVAB testing program evaluates (adverse) impact 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.

OPA
OFFICE OF PEOPLE ANALYTICS

WHEN IS ADVERSE IMPACT MEASURED?

- **Ideally, adverse impact is assessed on a regular basis.**
- **Here, adverse impact is measured for applicants testing in fiscal year 2017.**

  FY2017 = Oct 1, 2016 – Sept 30, 2017
- **Previously, adverse impact was evaluated for applicants testing in:**

  FY2015 = October 1, 2014 – September 30, 2015

  FY2013 = October 1, 2012 – September 30, 2013

  FY2011 = October 1, 2010 – September 30, 2011

  FY2009 = October 1, 2008 – September 30, 2009

  FY2005 = October 1, 2004 – September 30, 2005

HOW IS ADVERSE IMPACT MEASURED?

- **The four-fifths rule is often used to determine the occurrence of adverse impact:**

  "A selection rate for any race, sex, or ethnic group which is less than four-fifths (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact."

  -[Section 60-3, Uniform Guidelines on Employee Selection Procedures (1978); 43 FR 38295 (August 25, 1978).]

- **The ratio comparing the selection rates is called the *impact ratio*:**

$$IR = \frac{SR_{Foc}}{SR_{Ref}},$$ where SR is the selection rate

OPA
OFFICE OF PEOPLE ANALYTICS

HOW IS ADVERSE IMPACT MEASURED?

- **Statistical significance of the impact ratio can be computed, as well as confidence intervals around the impact ratio (Morris & Lobsenz, 2000):**
- $Z_{IR} = \frac{\ln\left(\frac{SR_{Foc}}{SR_{Ref}}\right)}{\sqrt{\frac{1-SR_{Tot}}{SR_{Tot}}\left(\frac{1}{N_{Foc}}+\frac{1}{N_{Ref}}\right)}}$, where *SR* = selection rate
- Z_{IR} is significant at $\alpha = .05$ if $|Z| > 1.96$
- Confidence interval $= e^{(\ln(IR)\pm 1.96 SE_{IR})}$, **where**
- $SE_{IR} = \sqrt{\frac{1-SR_{Foc}}{N_{Foc}SR_{Foc}} + \frac{1-SR_{Ref}}{N_{Ref}SR_{Ref}}}$

OPA
OFFICE OF PEOPLE ANALYTICS

HOW IS ADVERSE IMPACT MEASURED?

- **The four-fifths rule and accompanying statistics are applied to the ASVAB by comparing qualification rates across the focal and reference groups of interest with regard to:**
  - Examinees who qualify for entry into the military (i.e., those scoring in AFQT category IIIB or higher, AFQT ≥ 31).
  - Examinees who qualify for enlistment incentives (i.e., those scoring in AFQT category IIIA or higher, AFQT ≥ 50).
- **Note that adverse impact is measured using initial test scores only (i.e., scores from retests or confirmation tests are excluded from the analyses).**

HOW IS ADVERSE IMPACT MEASURED?

- **Effect sizes (i.e., standardized mean differences) provide another method of evaluating impact across individual ASVAB 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 group.

μ_F is the mean score in the Focal group.

σ_p is the pooled standard deviation across the two groups.

OPA
OFFICE OF PEOPLE ANALYTICS

HOW IS ADVERSE IMPACT MEASURED?

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

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

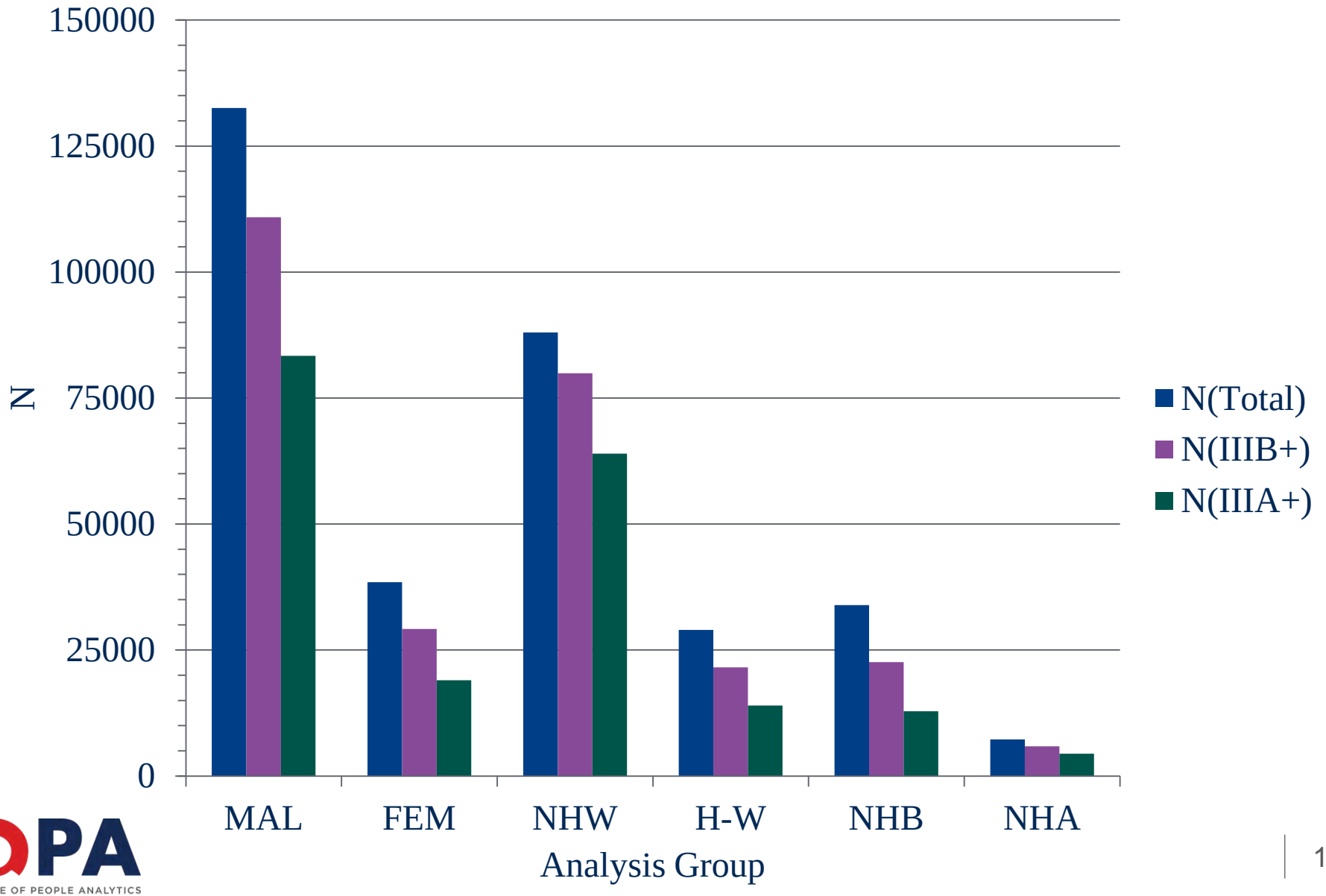
where

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

- **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.

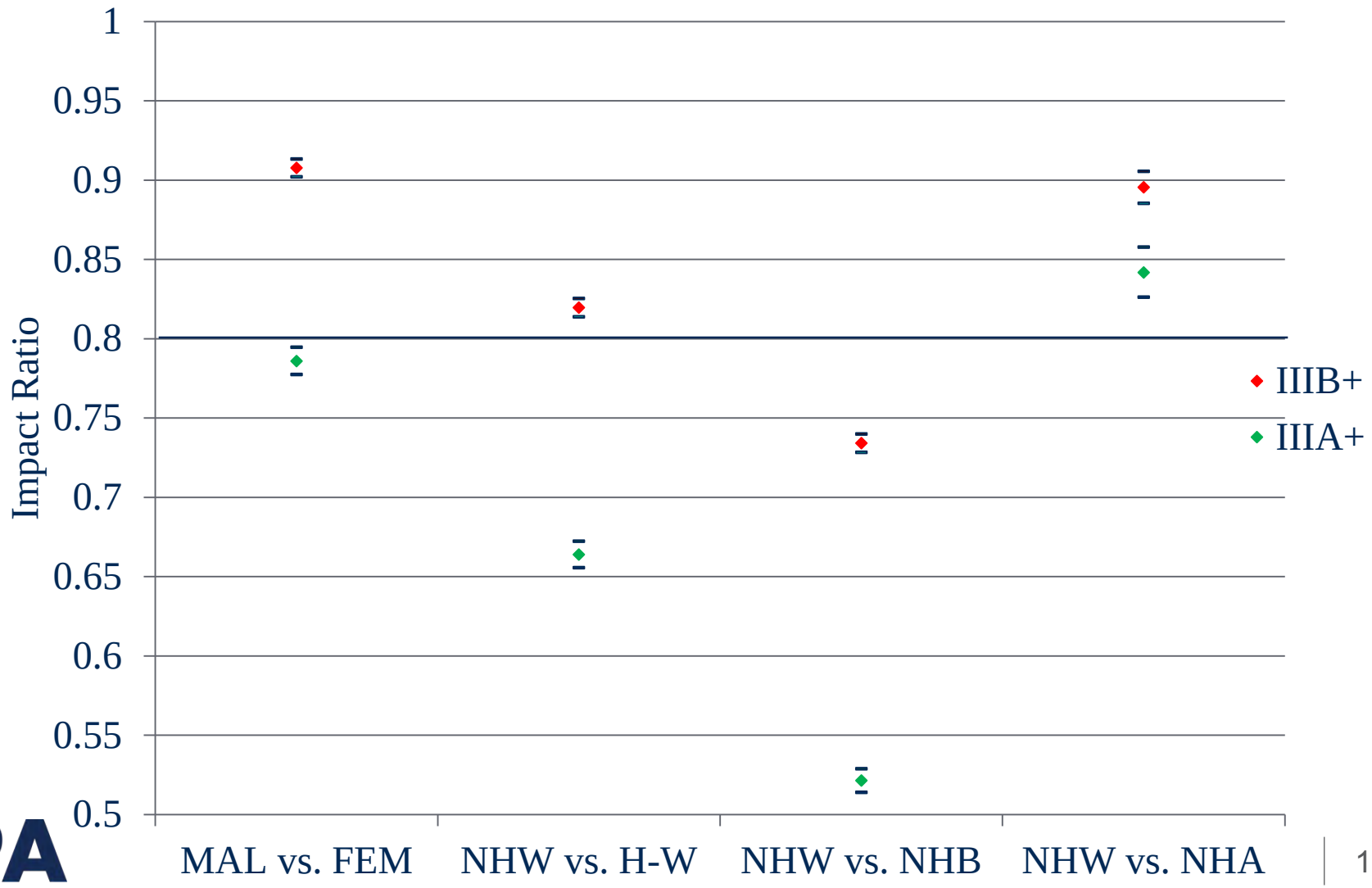
OPA
OFFICE OF PEOPLE ANALYTICS

Adverse Impact Analysis Sample Sizes - FY2017



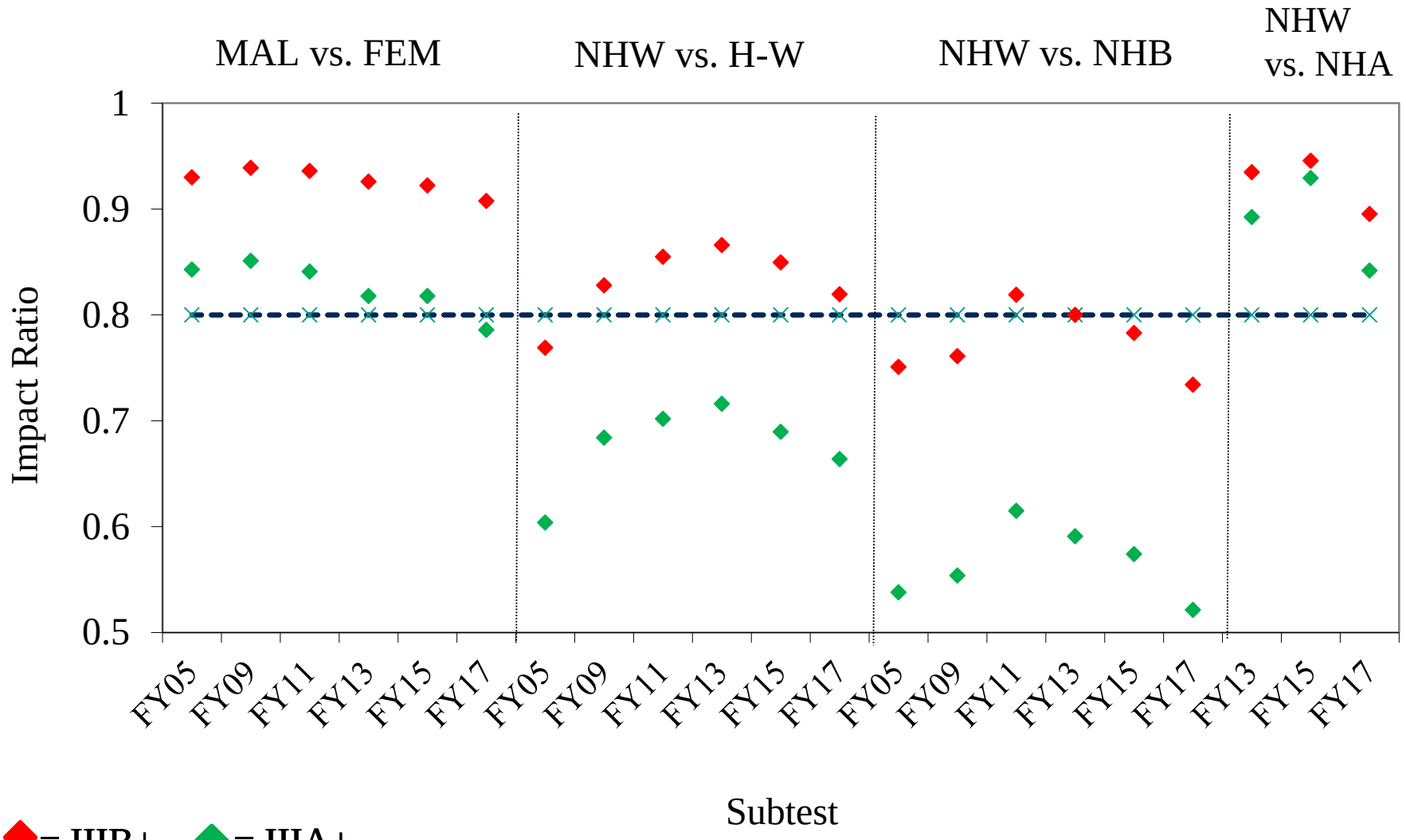
OPA
OFFICE OF PEOPLE ANALYTICS

Impact Ratio (and 95% Confidence Interval) for AFQT Cutscores
FY2017



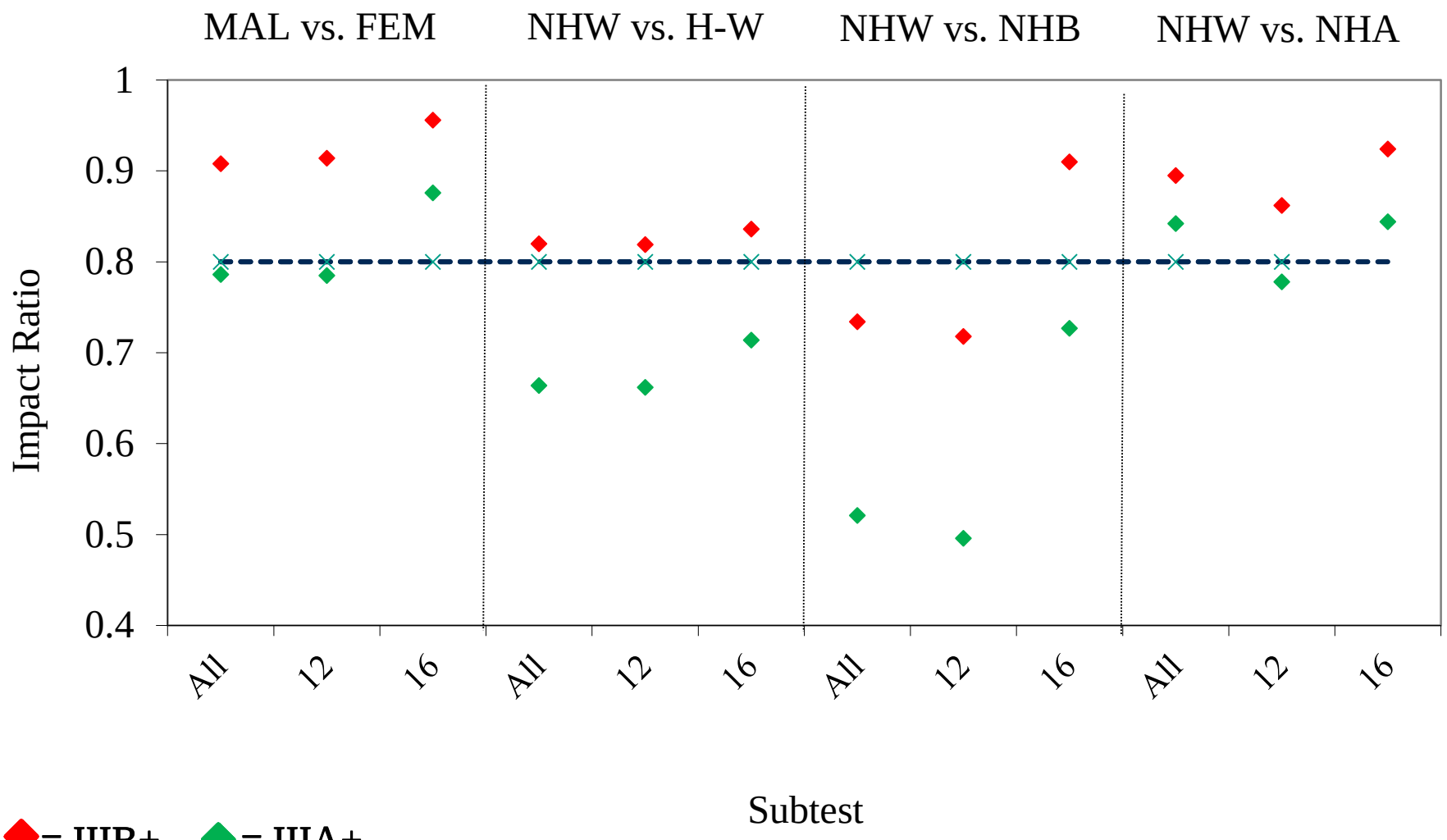
OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Impact Ratios* for FY05, FY09, FY11, FY13, FY15 & FY17



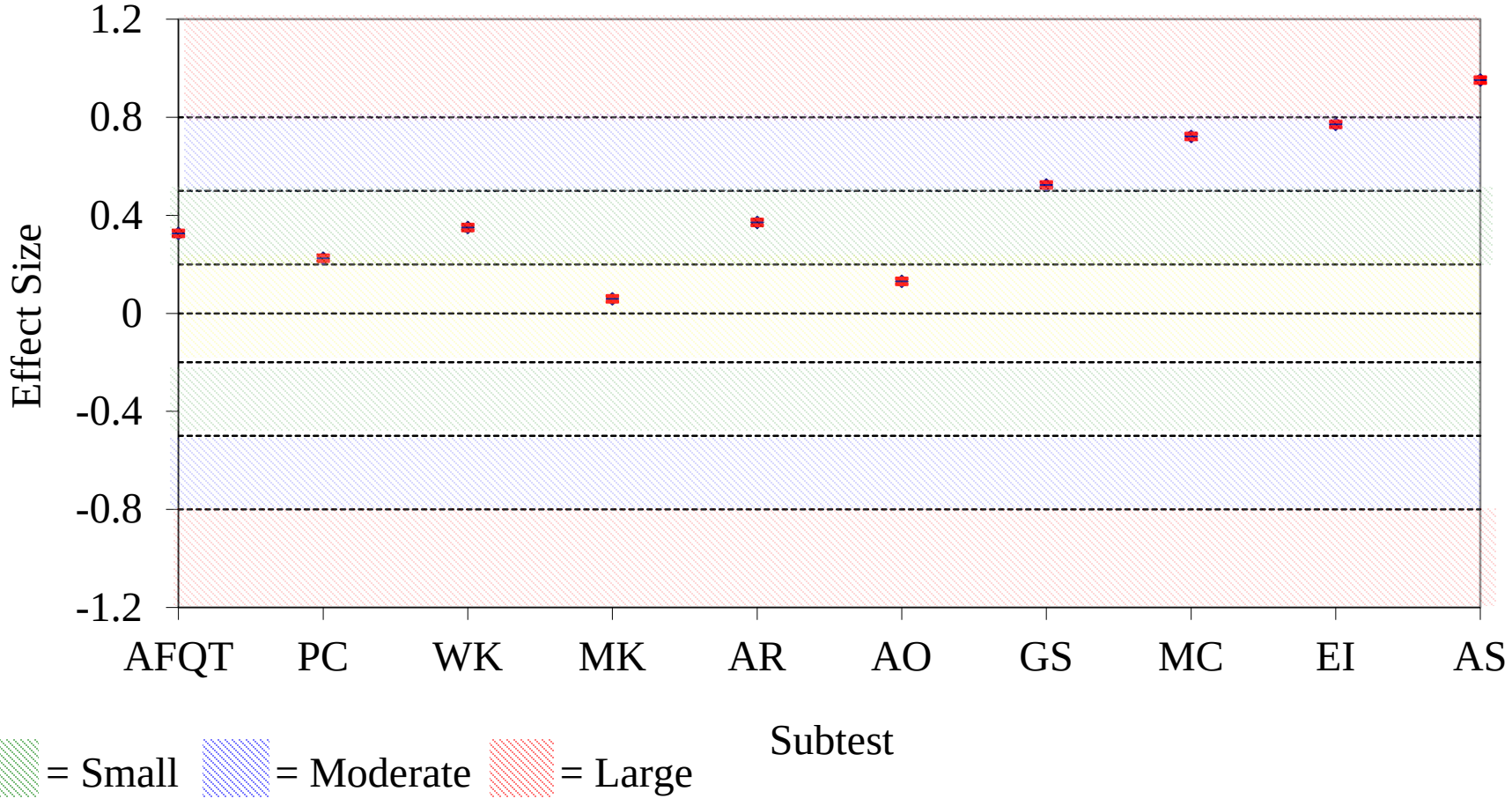
OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of FY2017 Impact Ratios for Years of Education Group



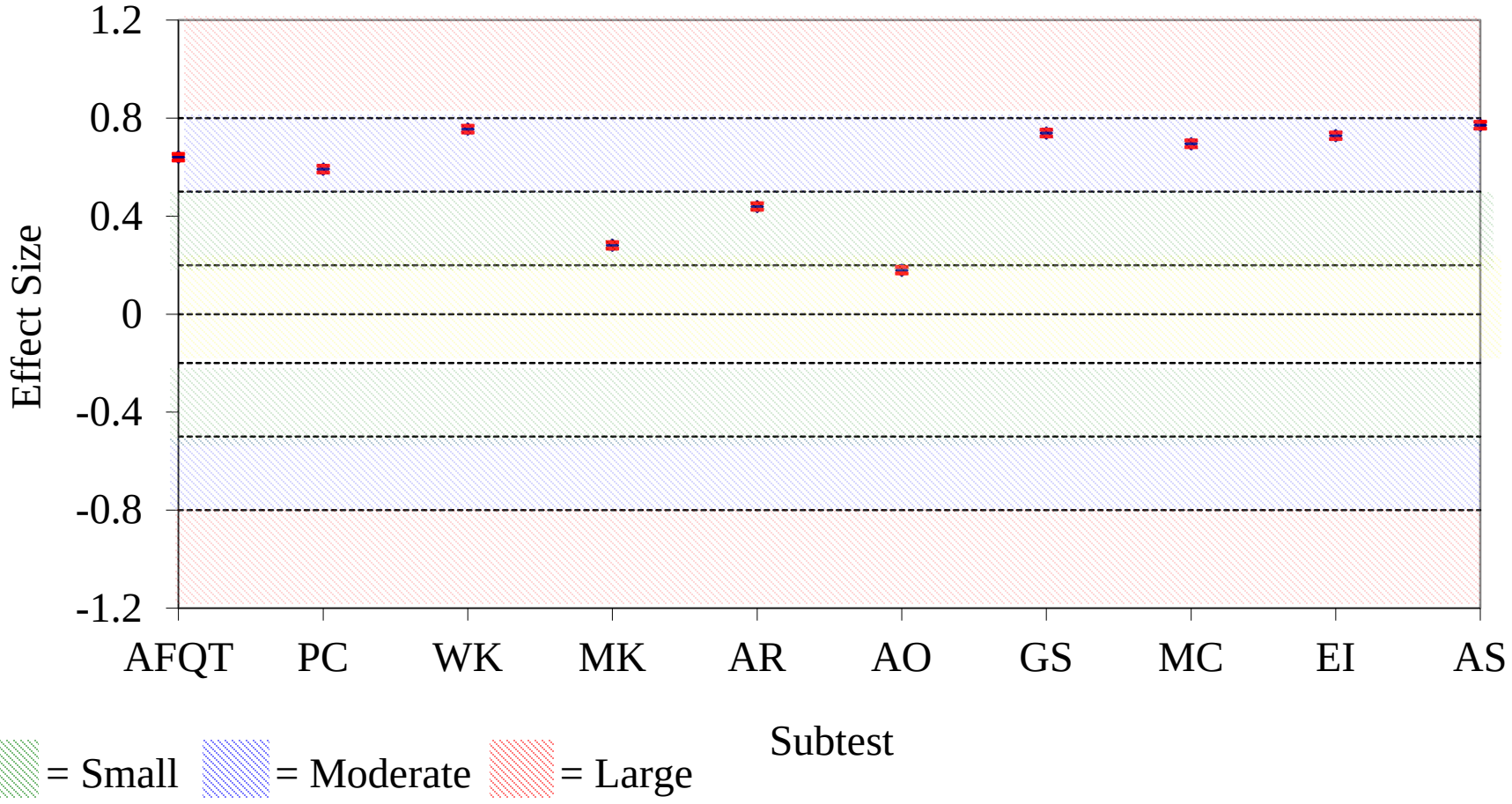
*12 = 12 years of education reported; 16 = 16 years of education reported.

Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Males Versus Females
FY2017



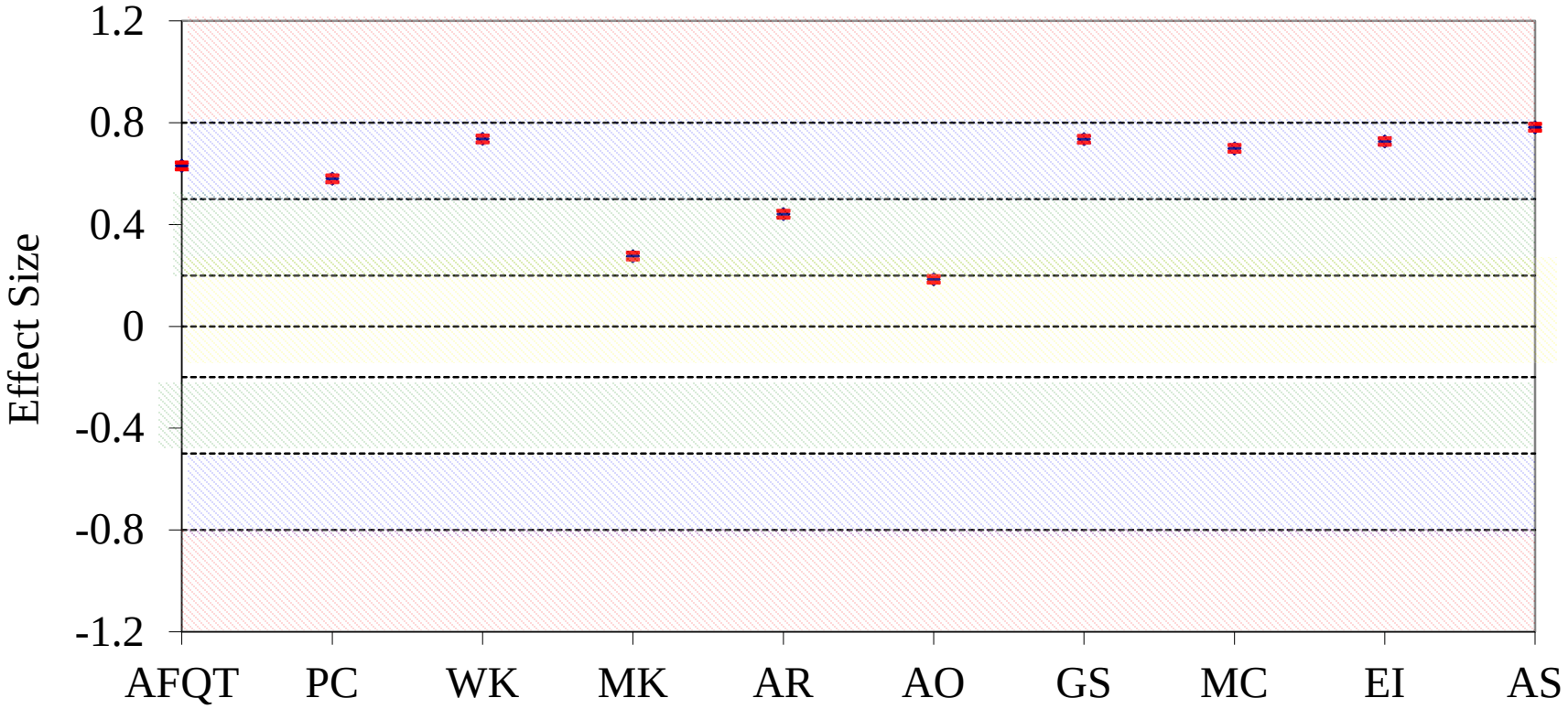
OPA
OFFICE OF PEOPLE ANALYTICS

Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Non-Hispanic Whites Versus Hispanic Whites
FY2017



OPA
OFFICE OF PEOPLE ANALYTICS

Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Non-Hispanic Whites Versus Hispanics*
FY2017

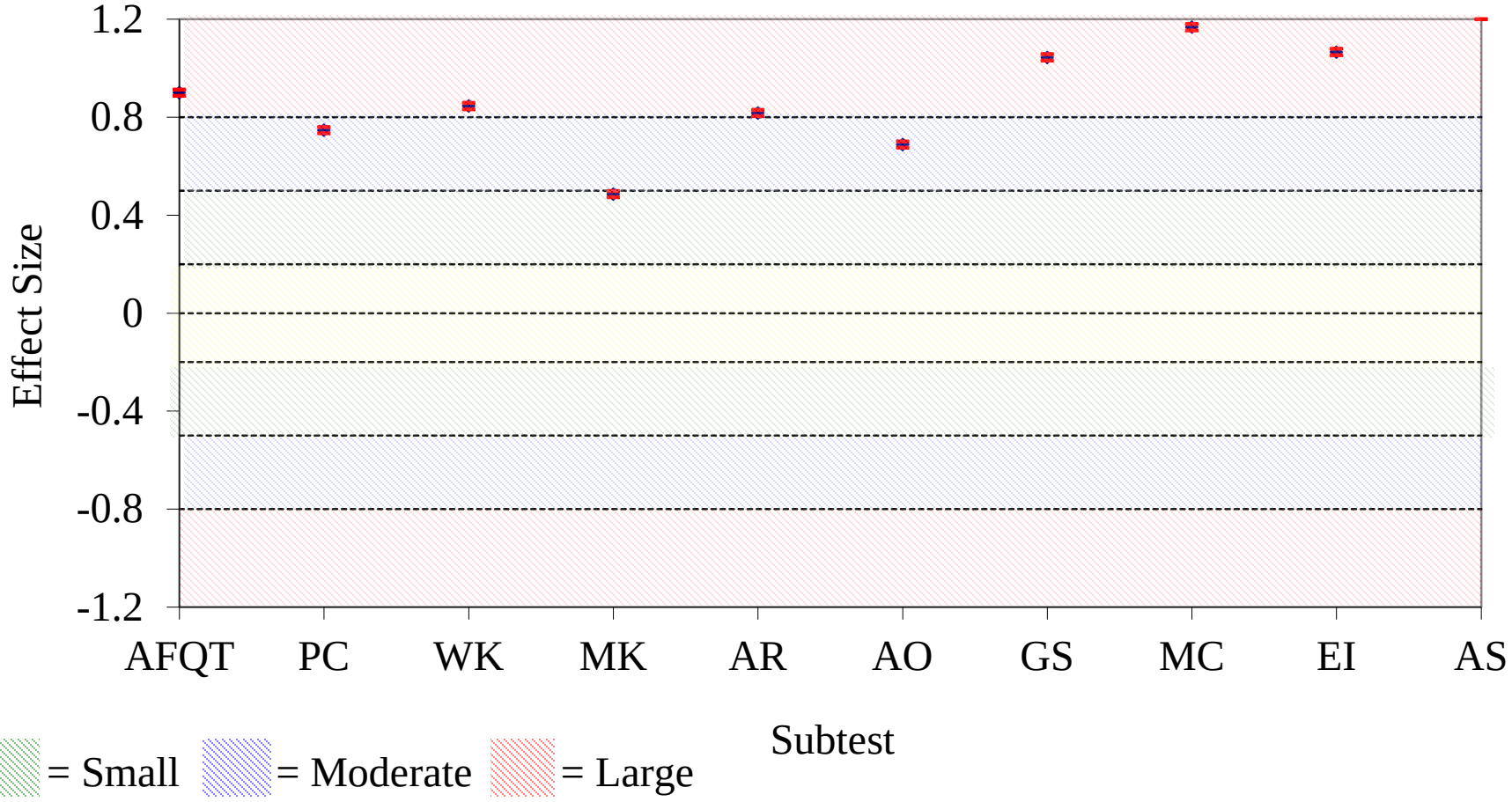


= Small = Moderate = Large

OPA
OFFICE OF PEOPLE ANALYTICS

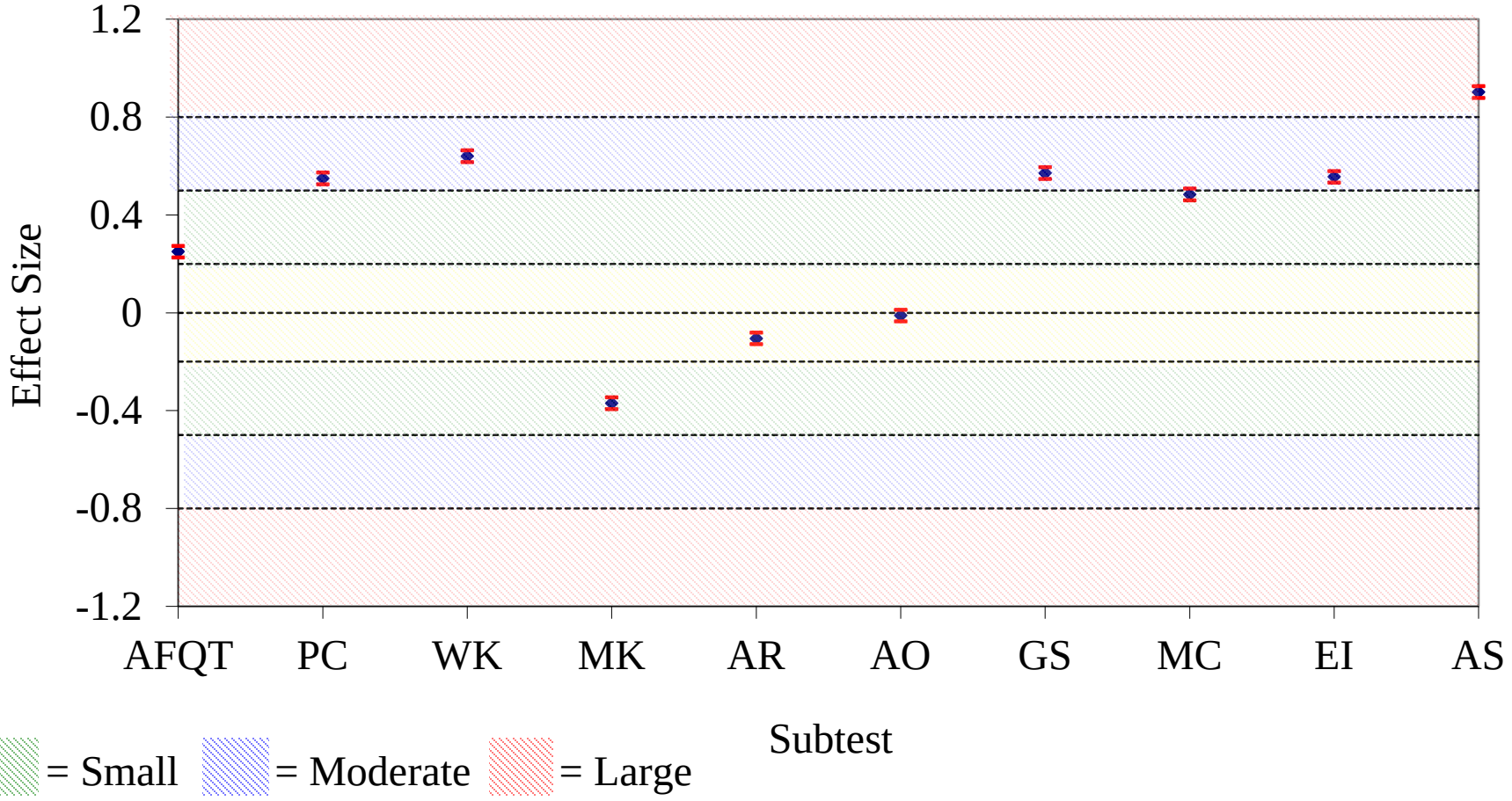
* Included for later comparisons with NAEP and SAT using Hispanics.

Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Non-Hispanic Whites Versus Non-Hispanic Blacks
FY2017



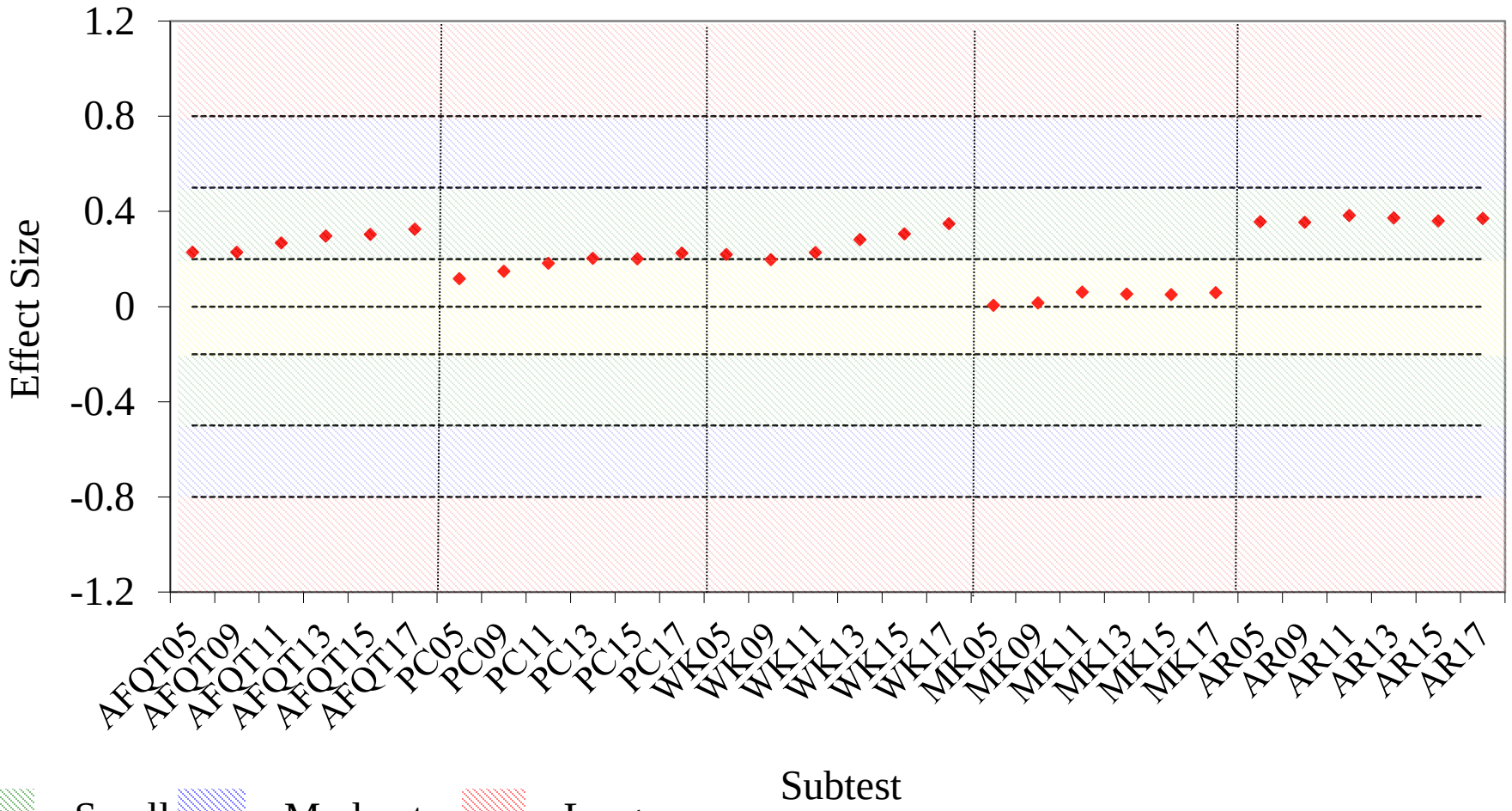
OPA
OFFICE OF PEOPLE ANALYTICS

Effect Sizes (and 95% Confidence Interval) for ASVAB Scores
Non-Hispanic Whites Versus Non-Hispanic Asians
FY2017



OPA
OFFICE OF PEOPLE ANALYTICS

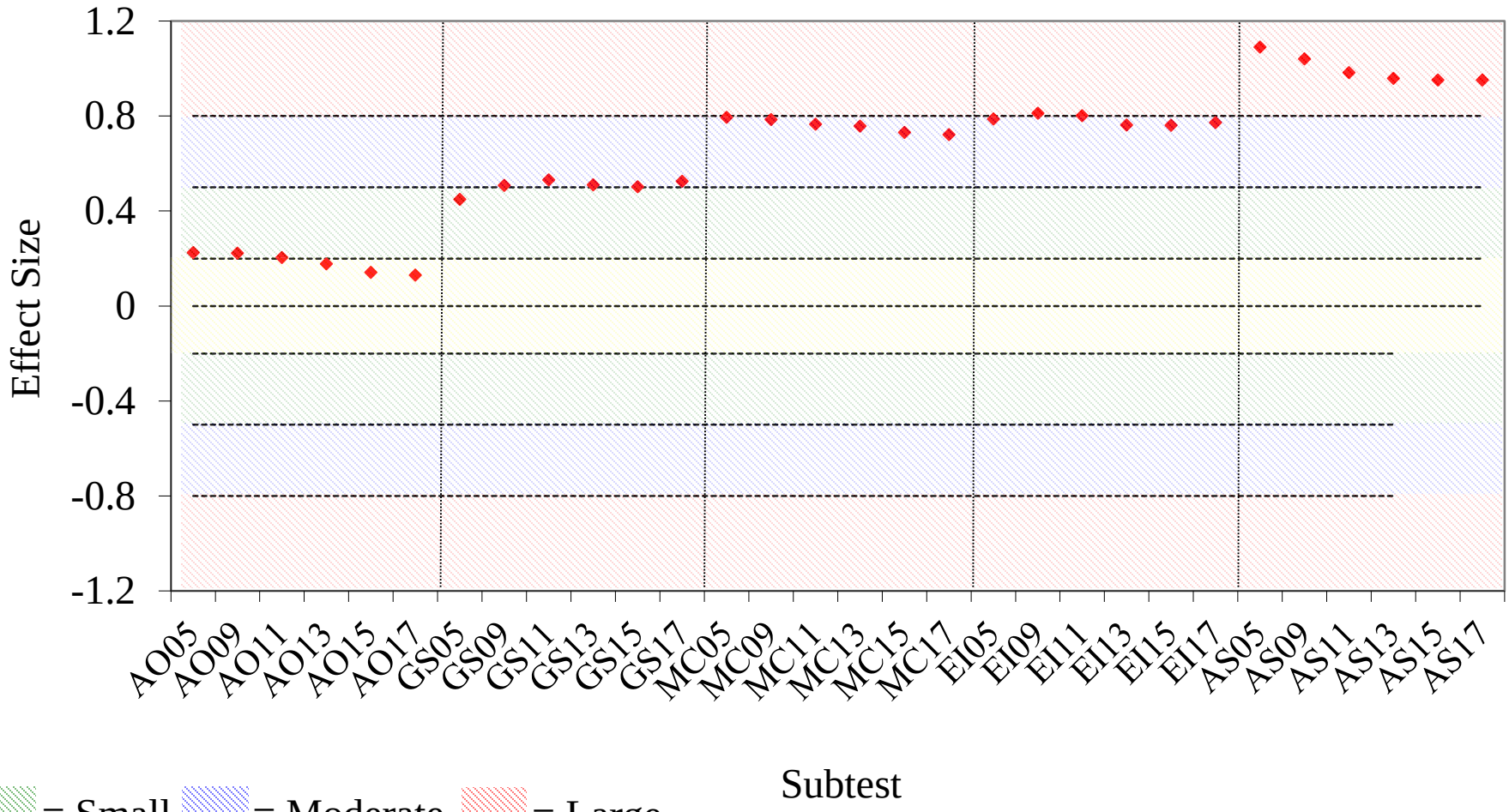
Comparison of Effect Sizes for FY05, FY09, FY11, FY13, FY15 & FY17
Males Versus Females
AFQT Tests/Score



= Small = Moderate = Large

OPA
OFFICE OF PEOPLE ANALYTICS

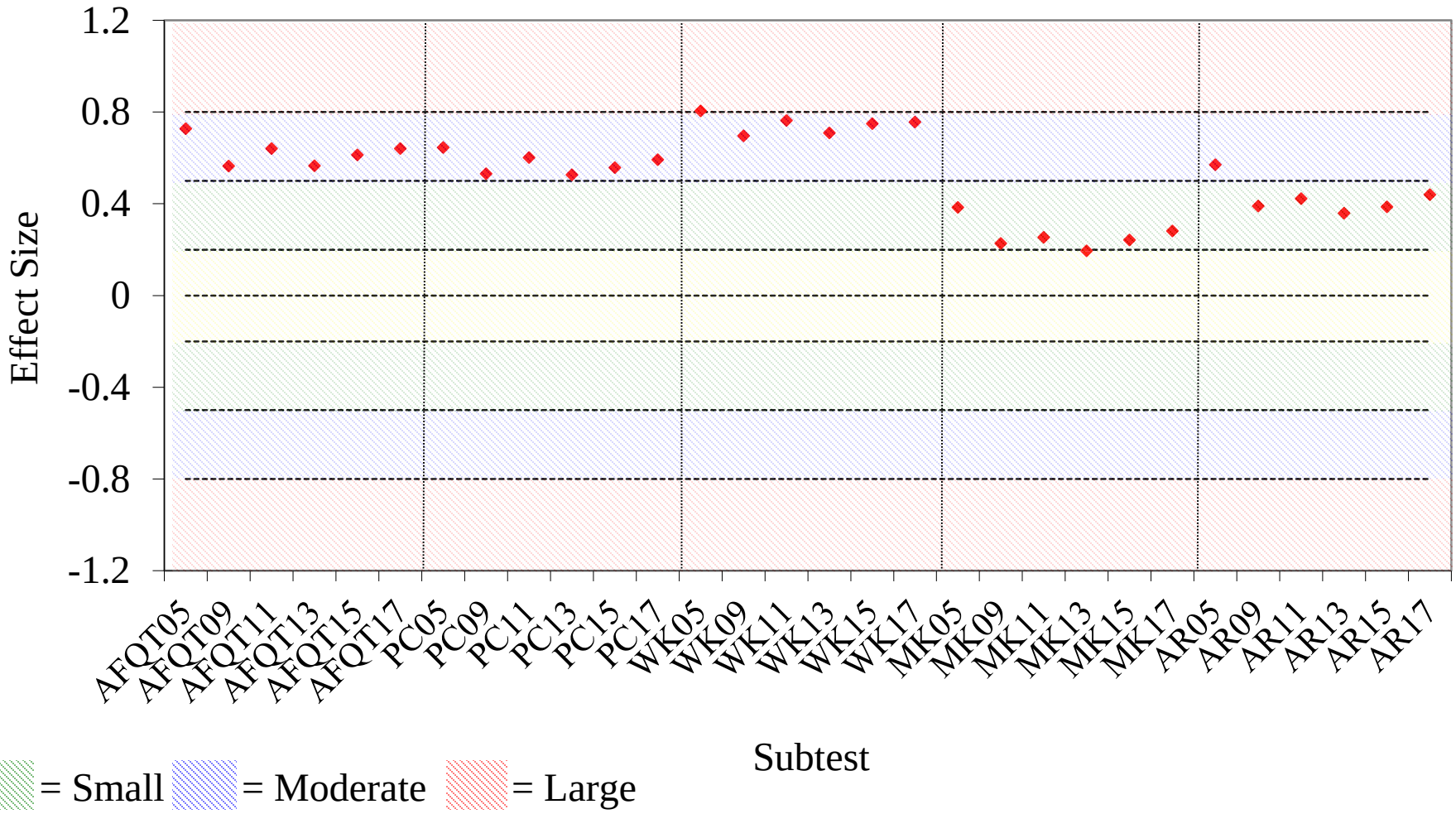
Comparison of Effect Sizes for FY05, FY09, FY11, FY13, FY15 & FY17
Males Versus Females
Non-AFQT Tests



= Small = Moderate = Large

OPA
OFFICE OF PEOPLE ANALYTICS

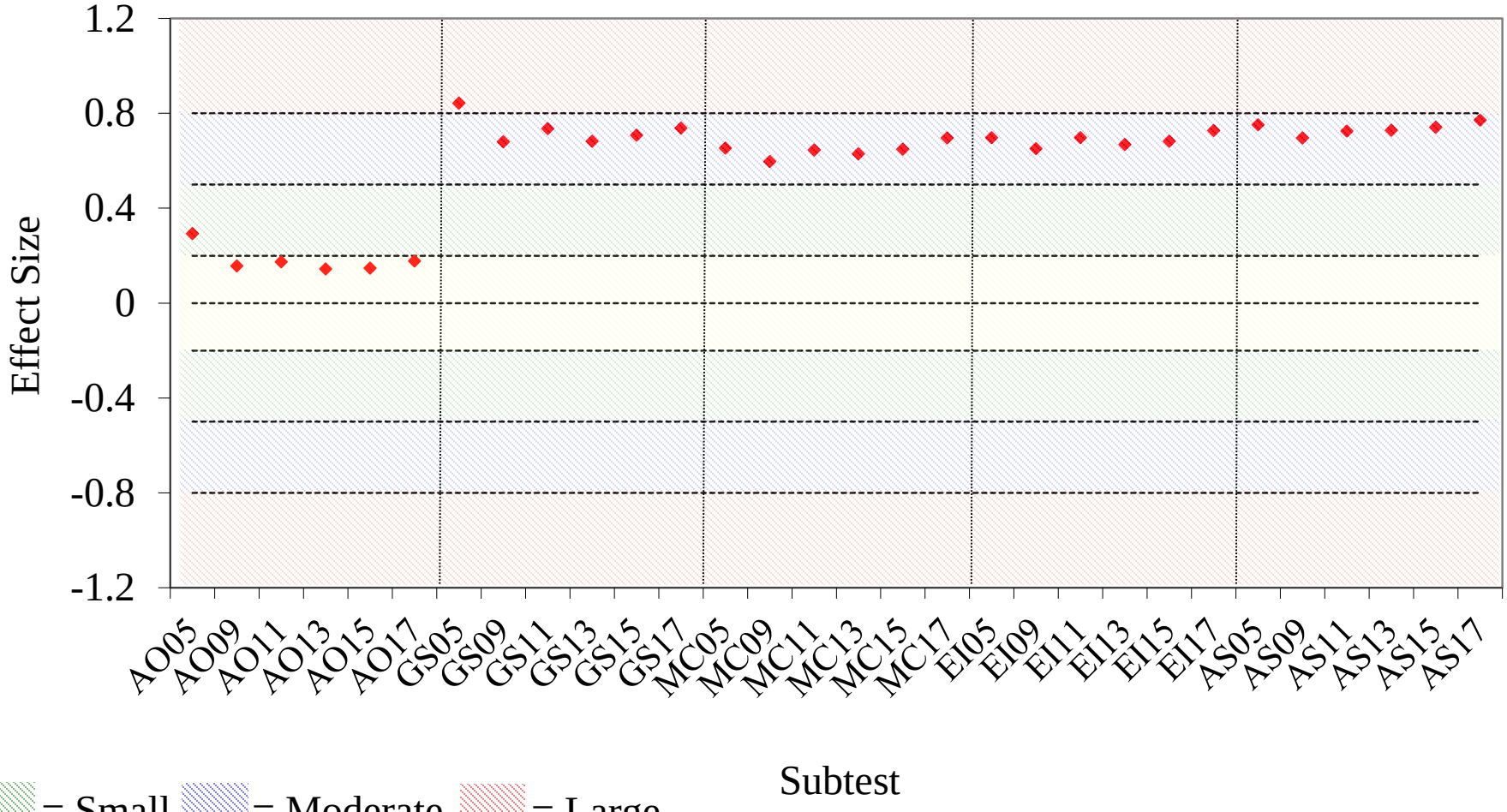
Comparison of Effect Sizes for FY05, FY09, FY11, FY13, FY15 & FY17
Non-Hispanic Whites Versus Hispanic Whites
AFQT Tests/Score



= Small = Moderate = Large

OPA
OFFICE OF PEOPLE ANALYTICS

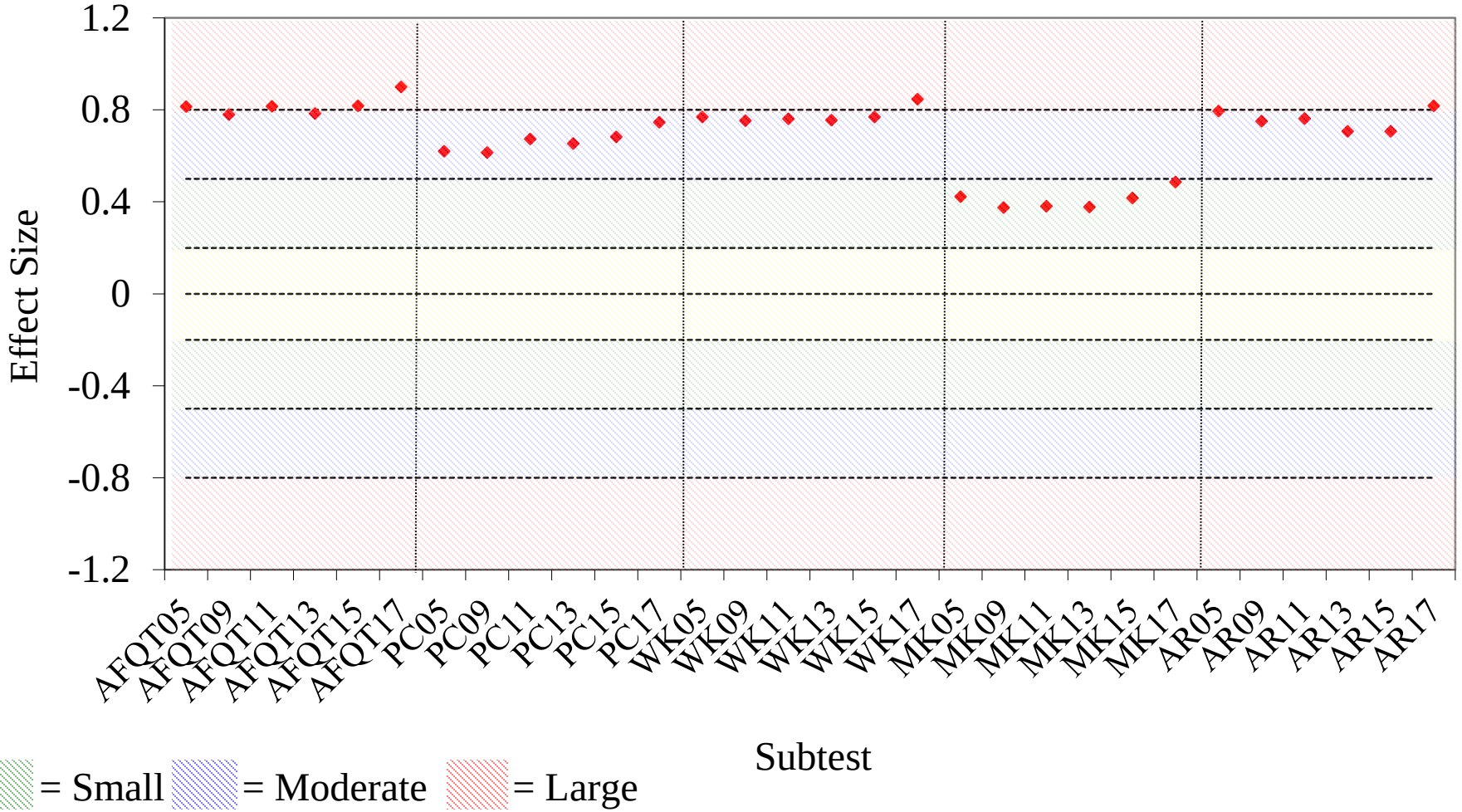
Comparison of Effect Sizes for FY05, FY09, FY11, FY13, FY15 & FY17
Non-Hispanic Whites Versus Hispanic Whites
Non-AFQT Tests



= Small = Moderate = Large

OPA
OFFICE OF PEOPLE ANALYTICS

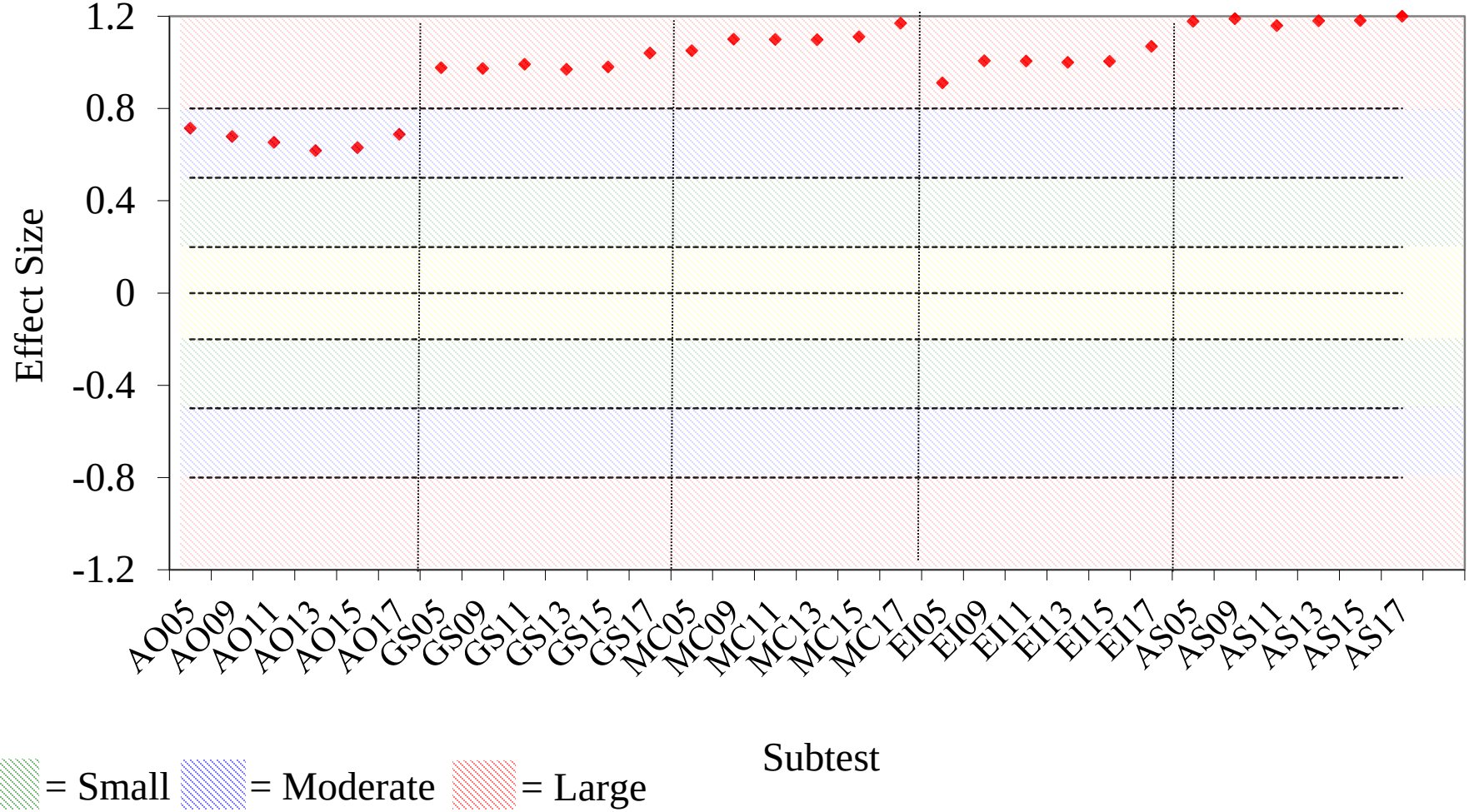
Comparison of Effect Sizes for FY05, FY09, FY11, FY13, FY15 & FY17
Non-Hispanic Whites Versus Non-Hispanic Blacks
AFQT Tests/Scores



= Small = Moderate = Large

OPA
OFFICE OF PEOPLE ANALYTICS

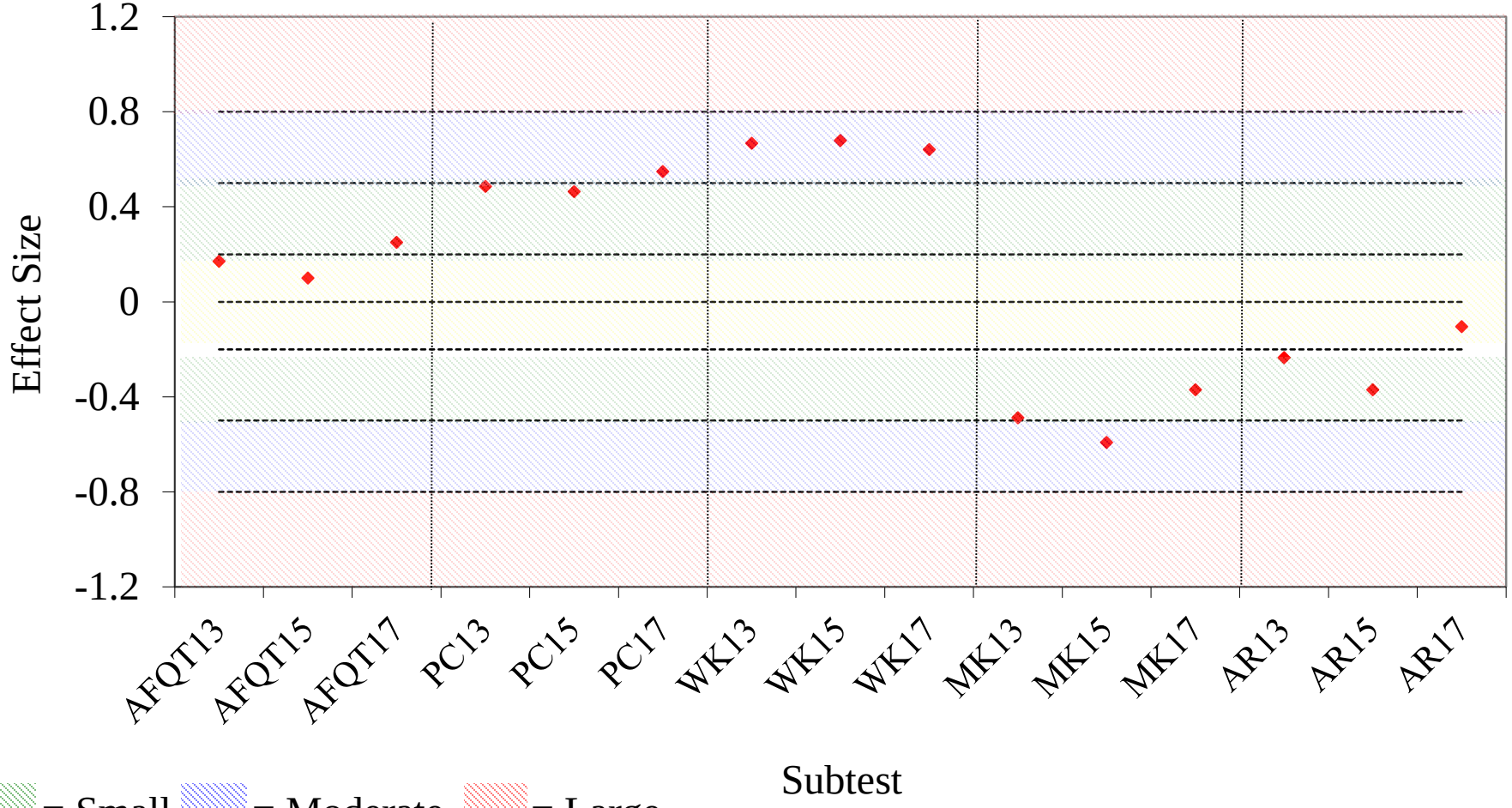
Comparison of Effect Sizes for FY05, FY09, FY11, FY13, FY15 & FY17
Non-Hispanic Whites Versus Non-Hispanic Blacks
Non-AFQT Tests



= Small = Moderate = Large

OPA
OFFICE OF PEOPLE ANALYTICS

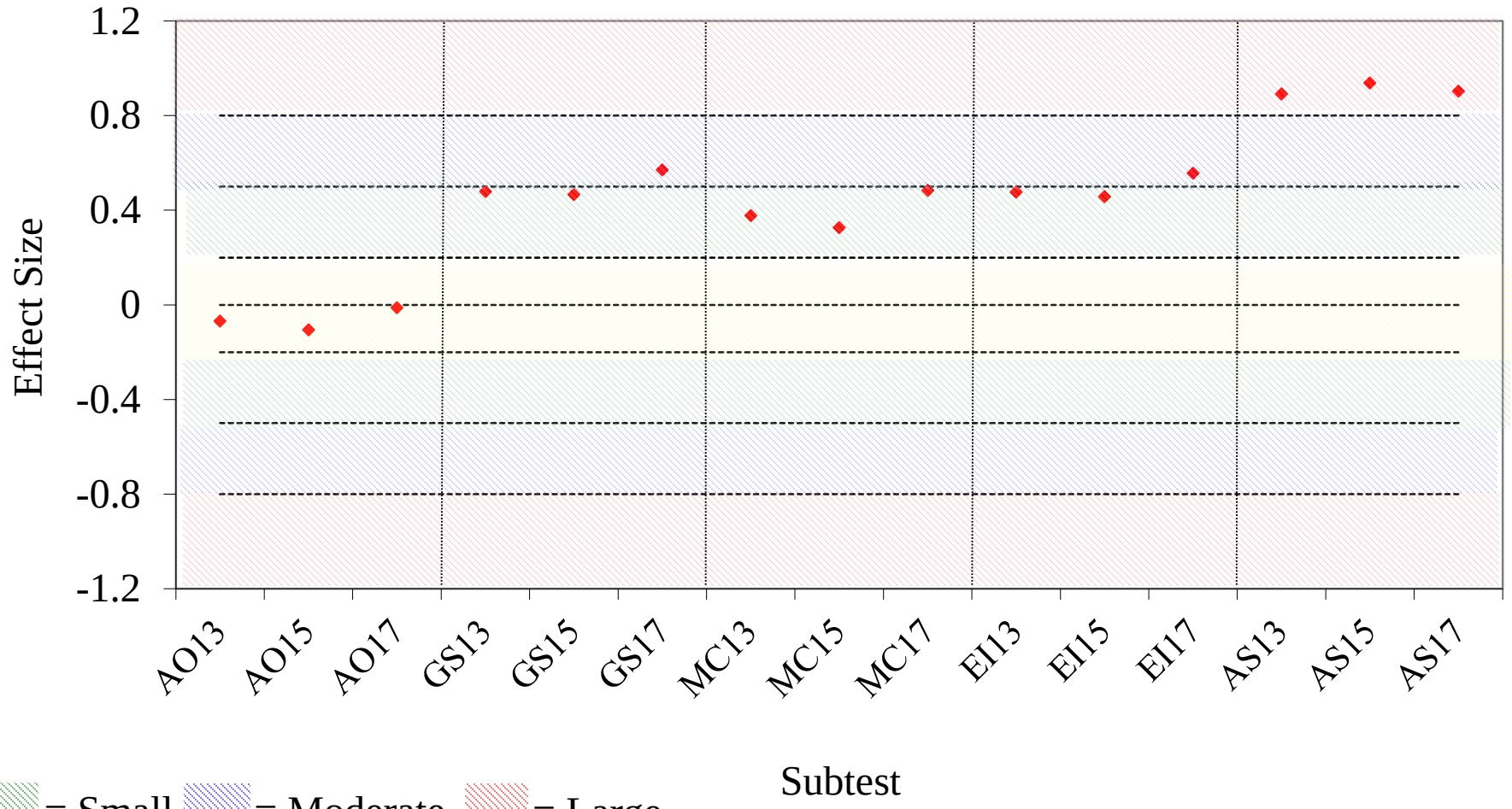
Comparison of Effect Sizes for FY3, FY15. & FY17
Non-Hispanic Whites Versus Non-Hispanic Asians
AFQT Tests/Scores



= Small = Moderate = Large

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes for FY13, FY15 & FY17
Non-Hispanic Whites Versus Non-Hispanic Asians
Non-AFQT Tests



= Small = Moderate = Large

OPA
OFFICE OF PEOPLE ANALYTICS

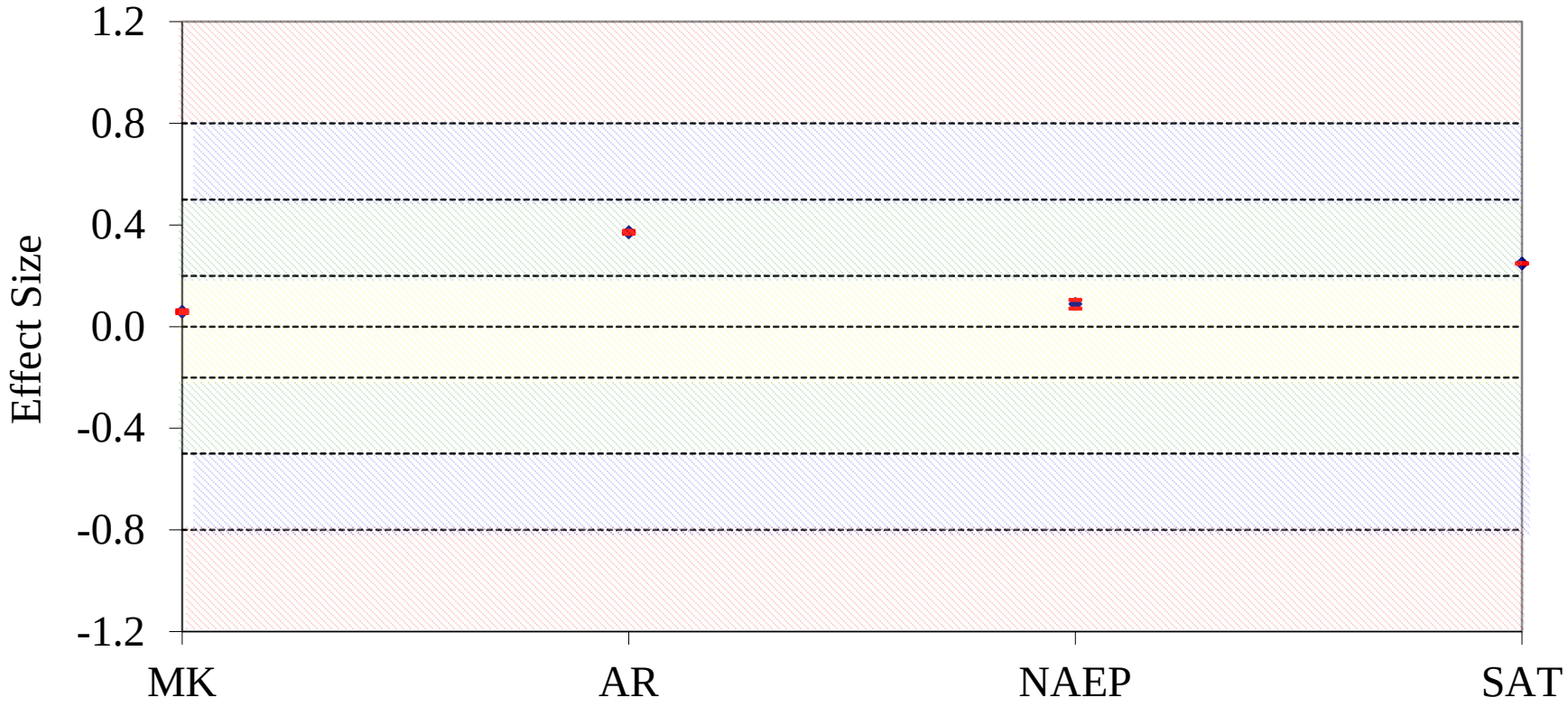
WHAT DOES IT MEAN?

- **The magnitude of impact on the ASVAB has remained fairly constant across fiscal years, but still varies in size from negligible to large across tests and groups.**
- **A comparison of impact across different testing programs gives some indication of whether the observed FY2017 magnitudes are reasonable.**
- **Sufficient information for estimating effect sizes is available online for two other large-scale testing programs:**

1. SAT – 2016 College Bound Seniors (Math and Reading)
2. NAEP – 2015 Grade 12 (Reading, Math, and Science)

OPA
OFFICE OF PEOPLE ANALYTICS

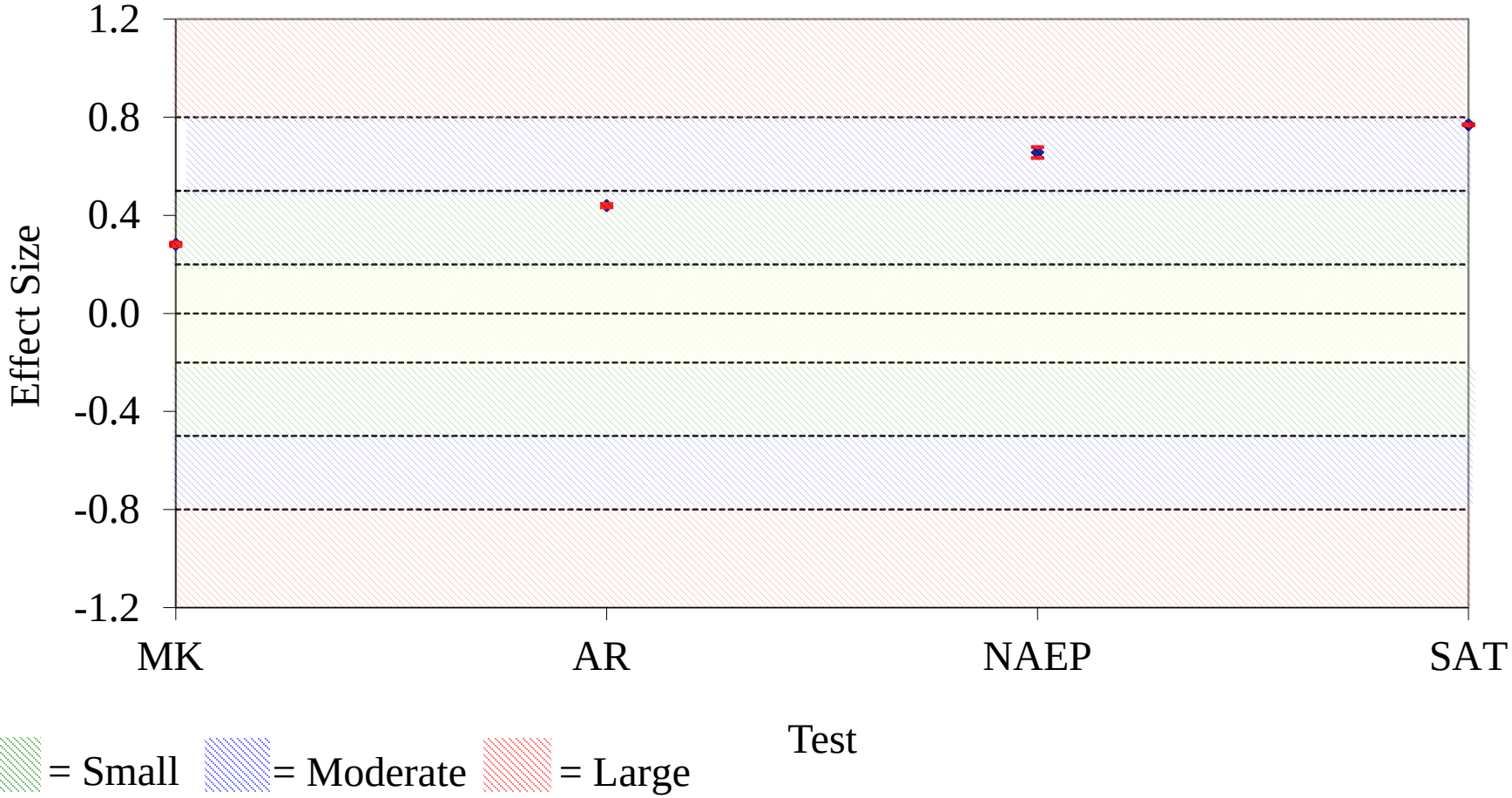
Comparison of Effect Sizes Across Testing Programs
Content Area = Math
Males Versus Females



= Small = Moderate = Large

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs
Content Area = Math
Non-Hispanic Whites Versus Hispanics*

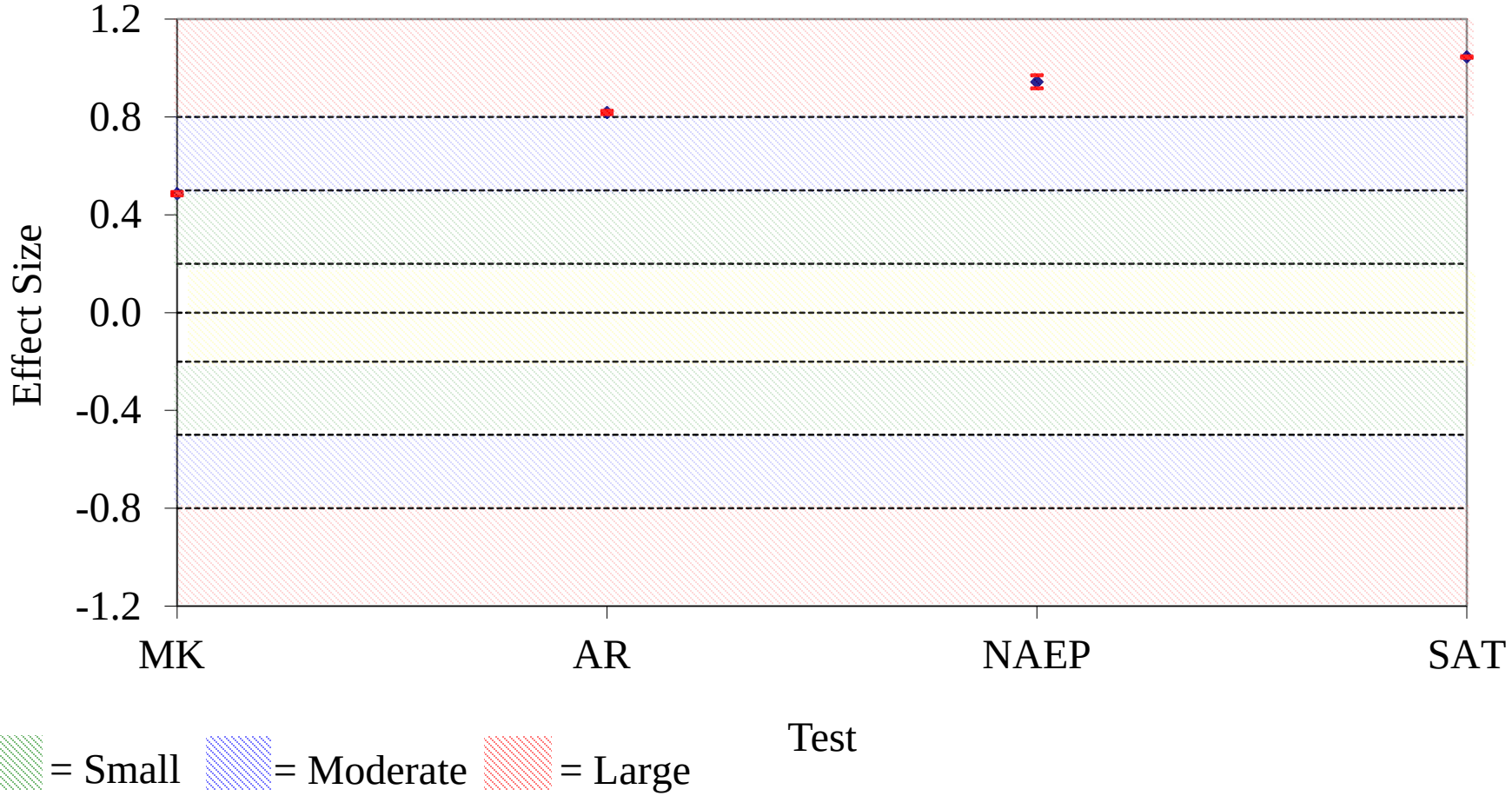


= Small = Moderate = Large

* Hispanics and Whites are mutually exclusive groups for SAT and NAEP.

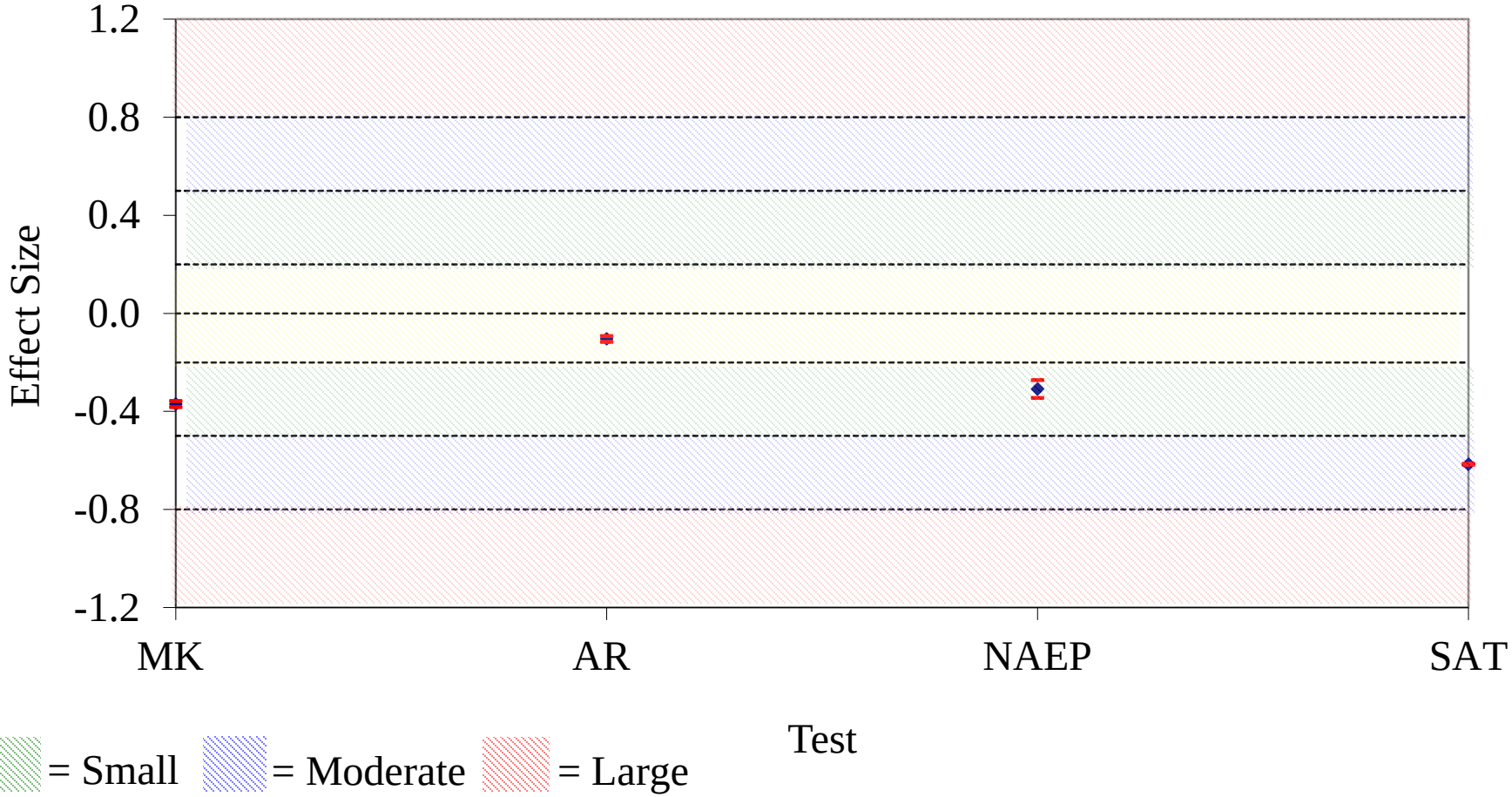
OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs
Content Area = Math
Non-Hispanic Whites Versus Non-Hispanic Blacks



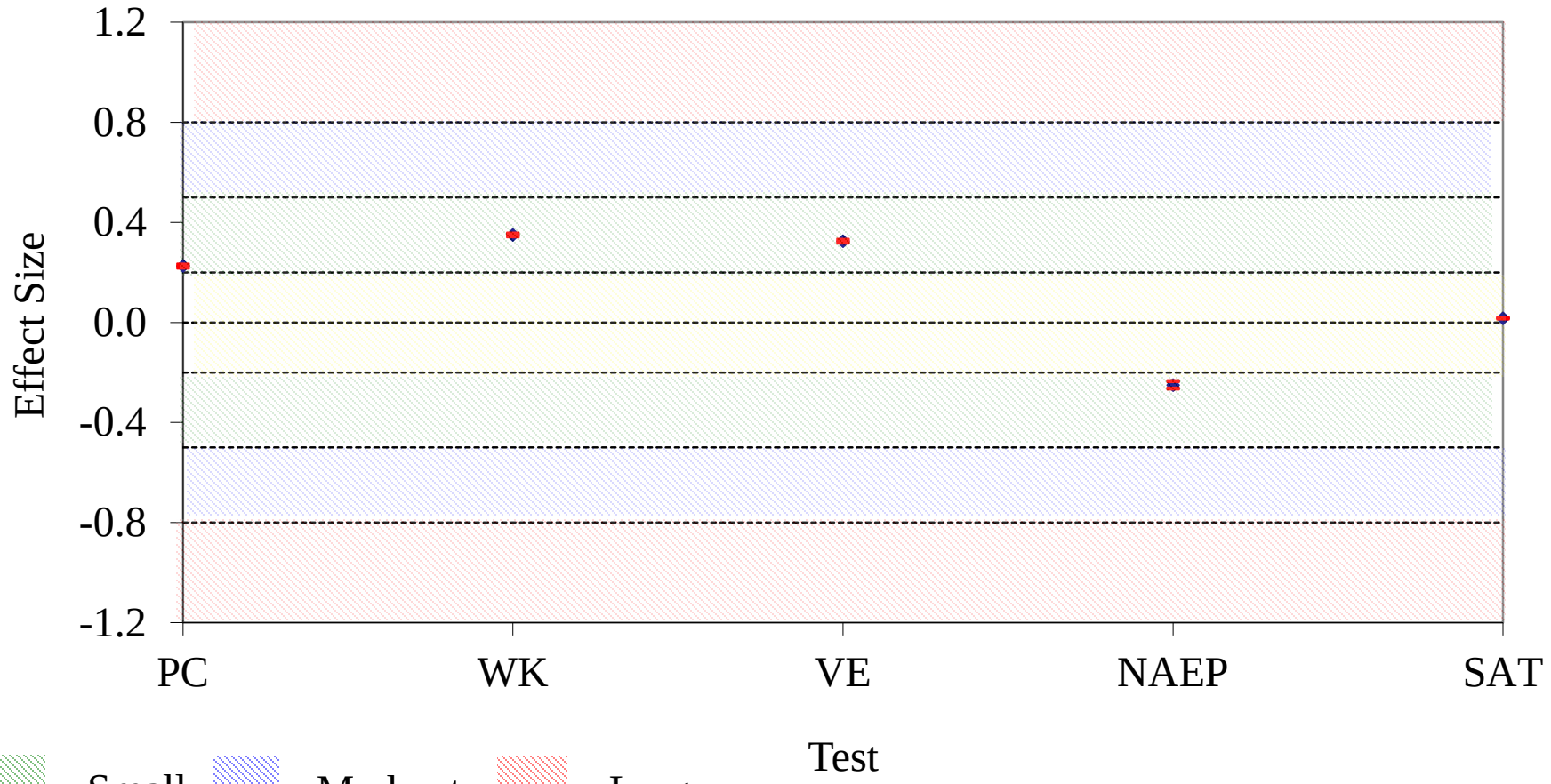
OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs
Content Area = Math
Non-Hispanic Whites Versus Asians



OPA
OFFICE OF PEOPLE ANALYTICS

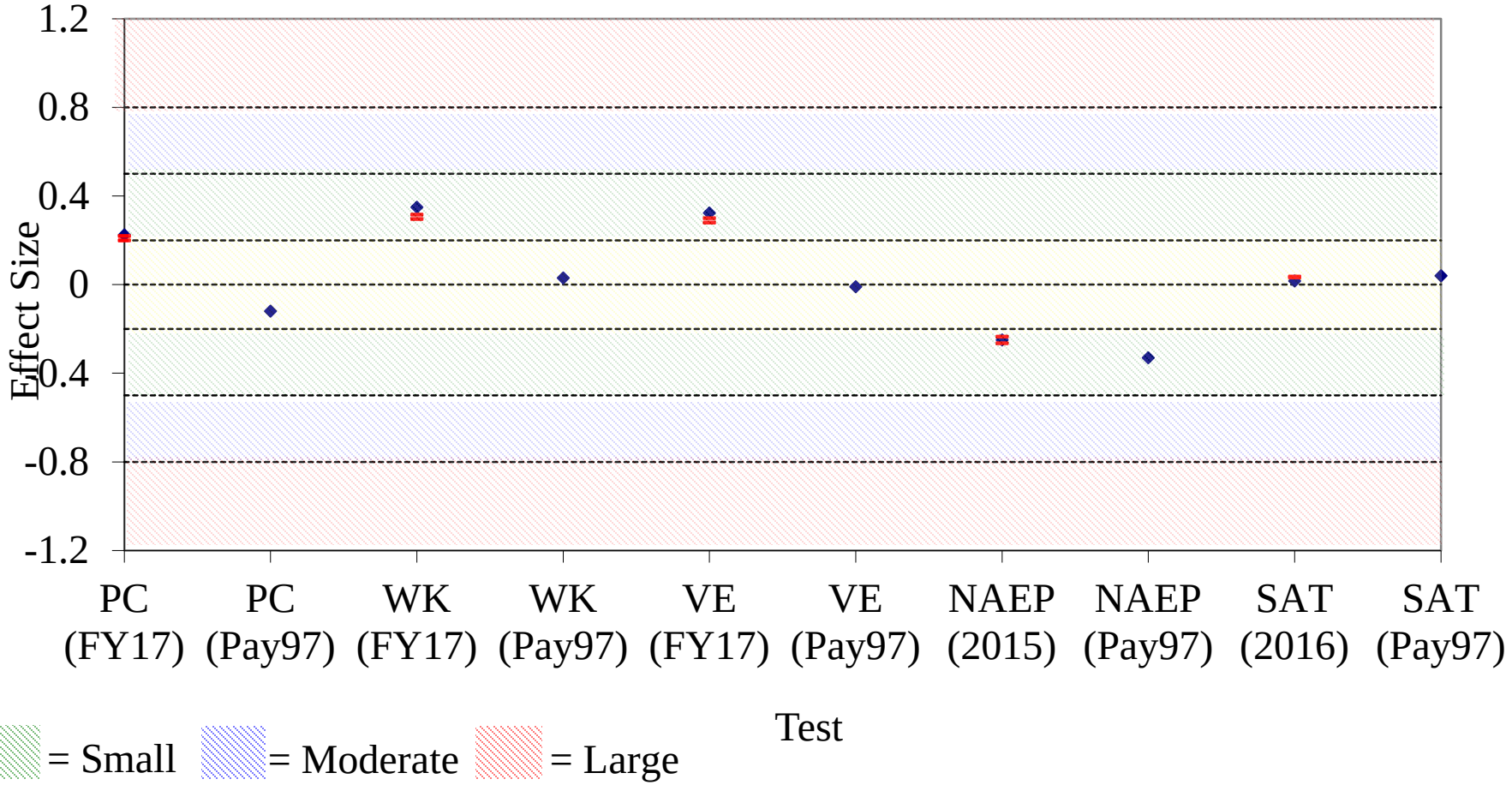
Comparison of Effect Sizes Across Testing Programs
Content Area = Reading/Verbal
Males Versus Females



= Small = Moderate = Large

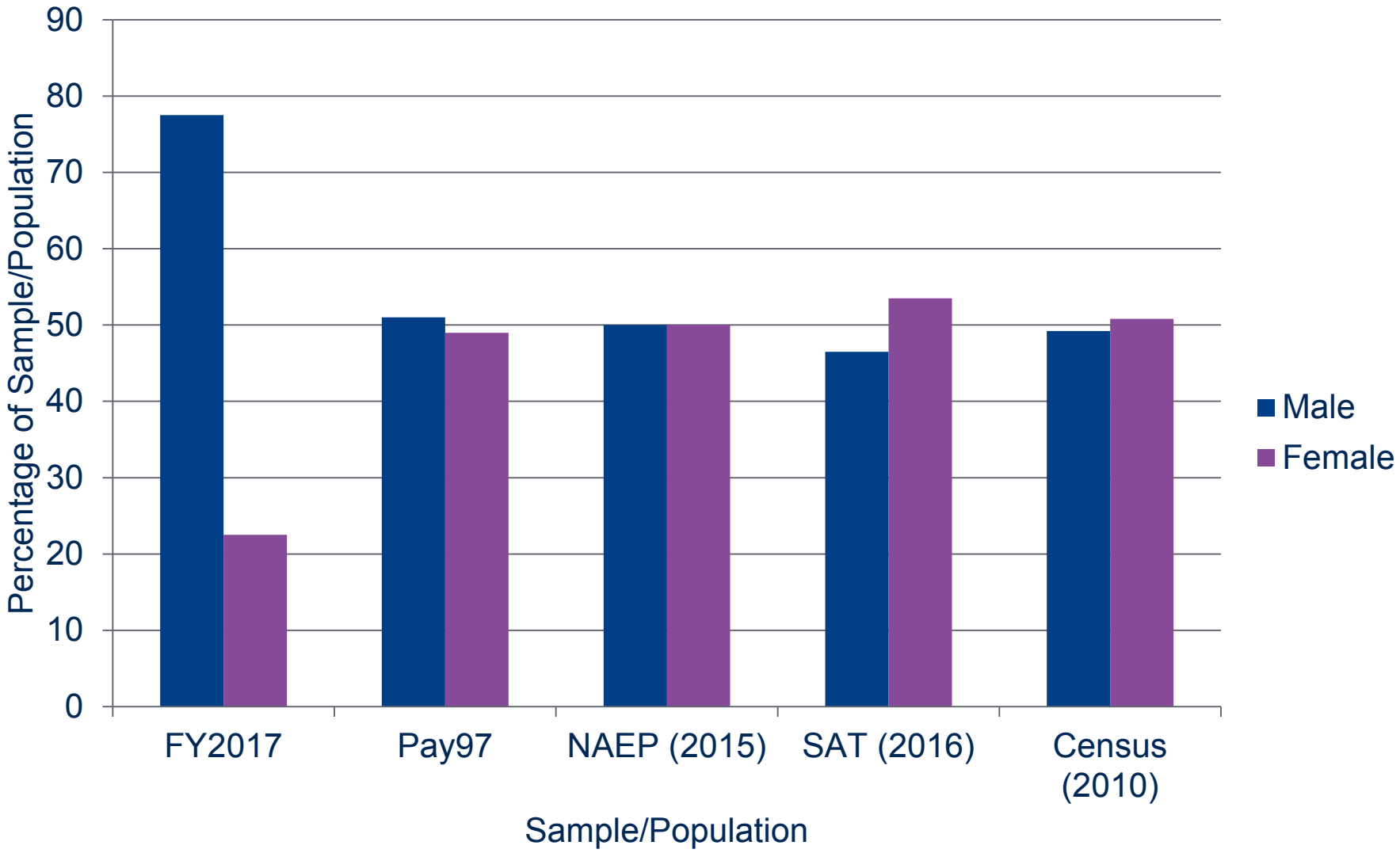
OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs
Content Area = Reading/Verbal
Males Versus Females



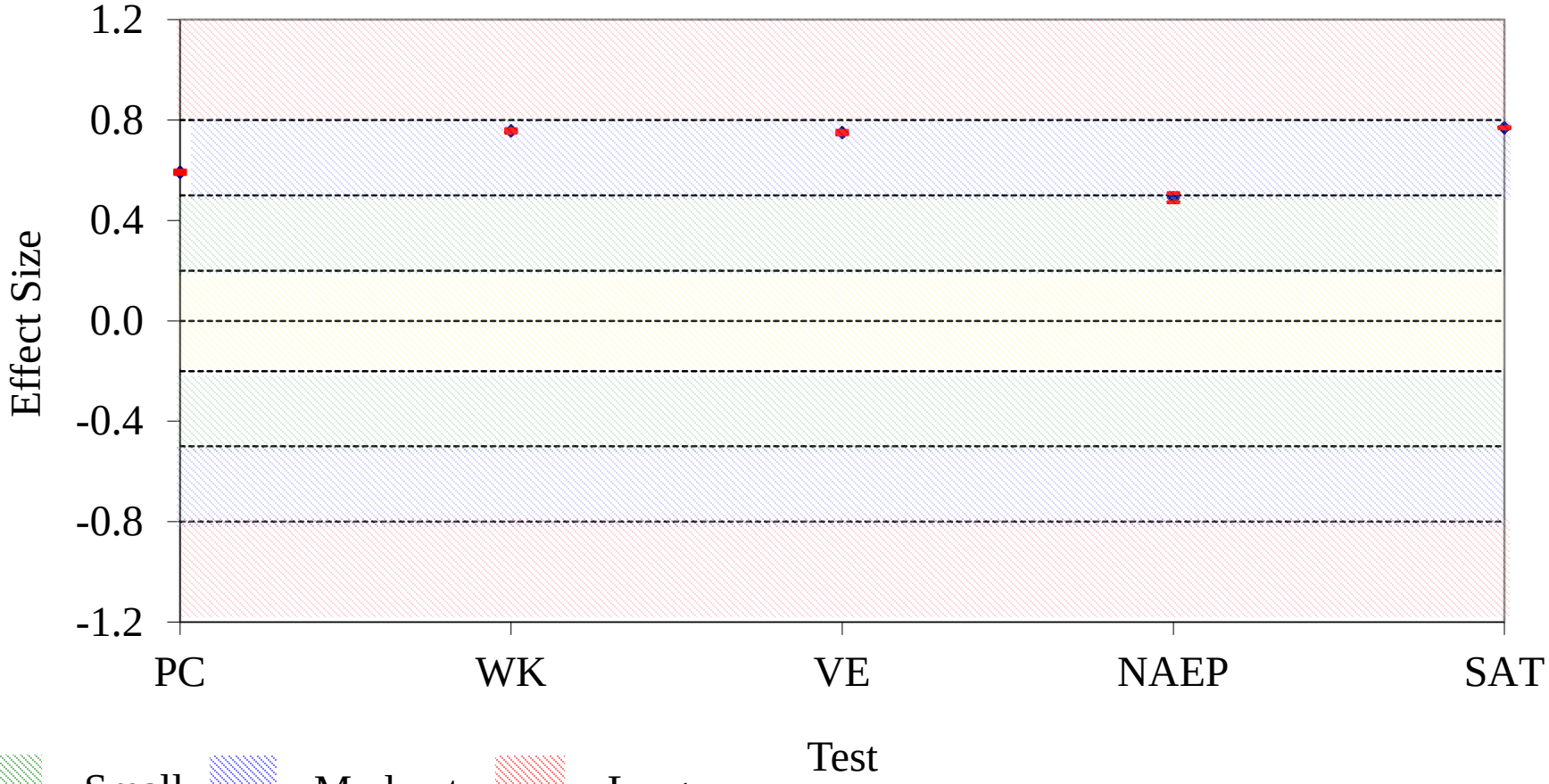
OPA
OFFICE OF PEOPLE ANALYTICS

Gender Representation Across Samples/Populations



OPA
OFFICE OF PEOPLE ANALYTICS

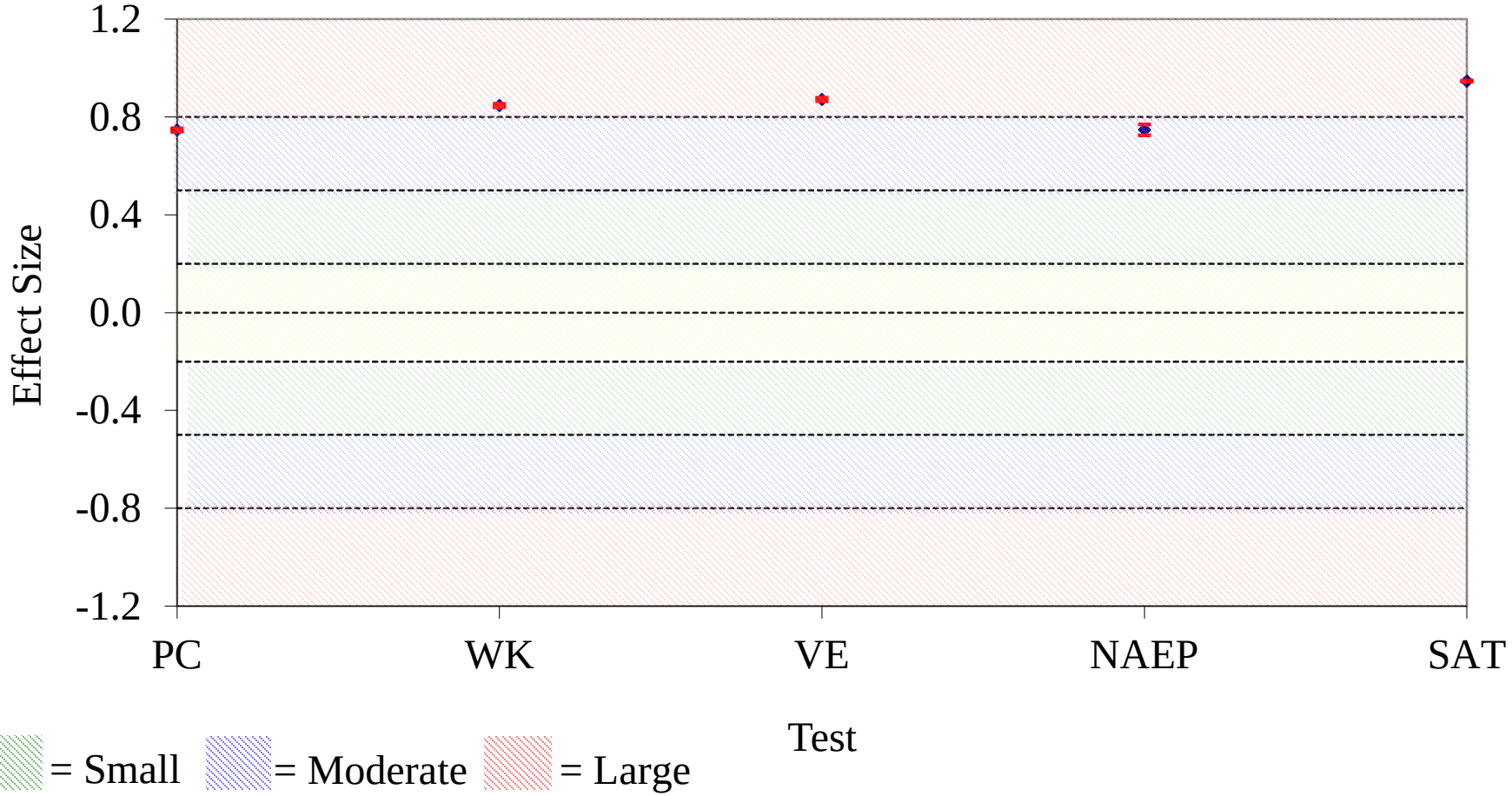
Comparison of Effect Sizes Across Testing Programs
Content Area = Reading/Verbal
Non-Hispanic Whites Versus Hispanics*



OPA
OFFICE OF PEOPLE ANALYTICS

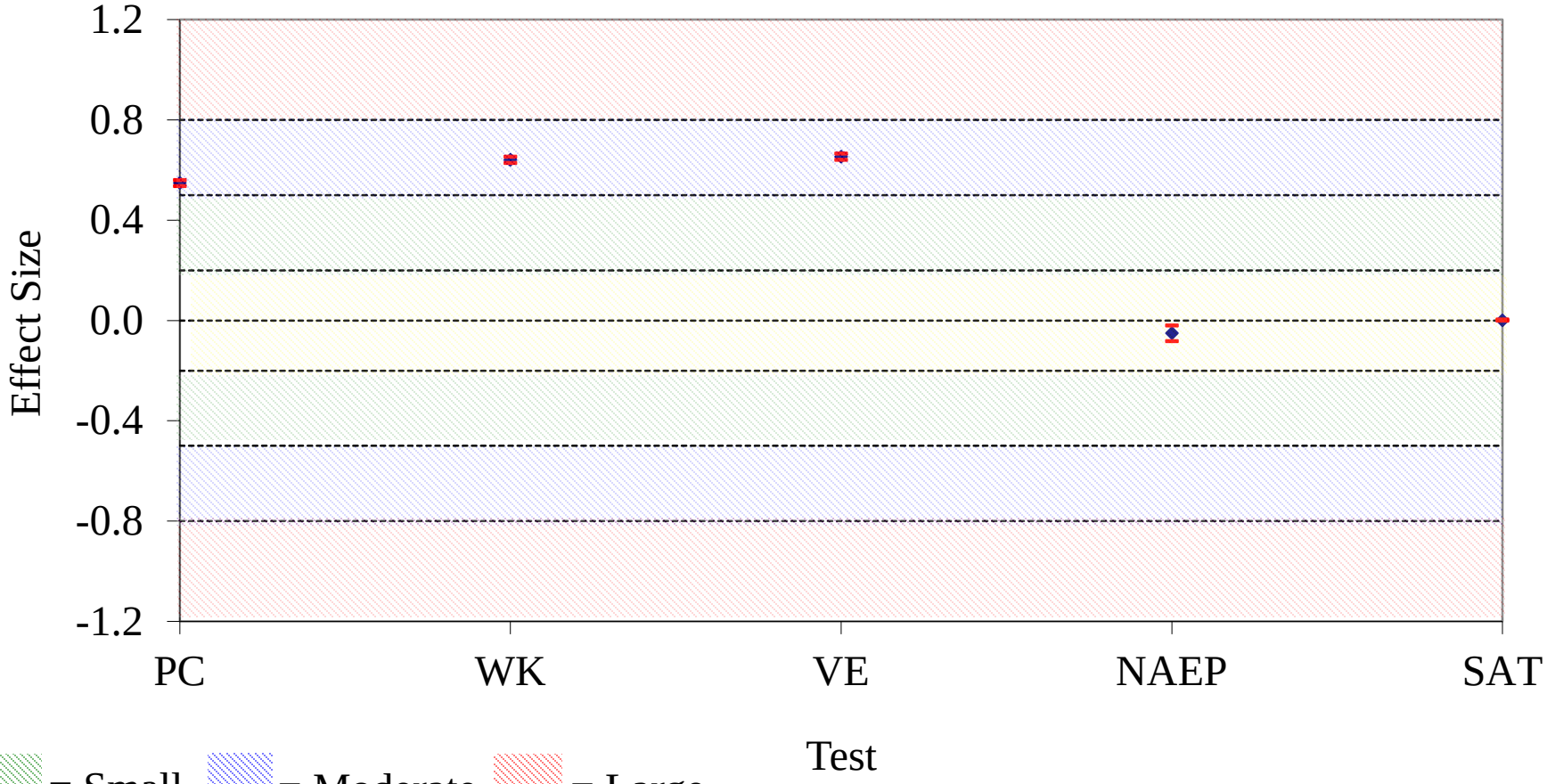
* Hispanics and Whites are mutually exclusive groups for SAT and NAEP.

Comparison of Effect Sizes Across Testing Programs
Content Area = Reading/Verbal
Non-Hispanic Whites Versus Non-Hispanic Blacks



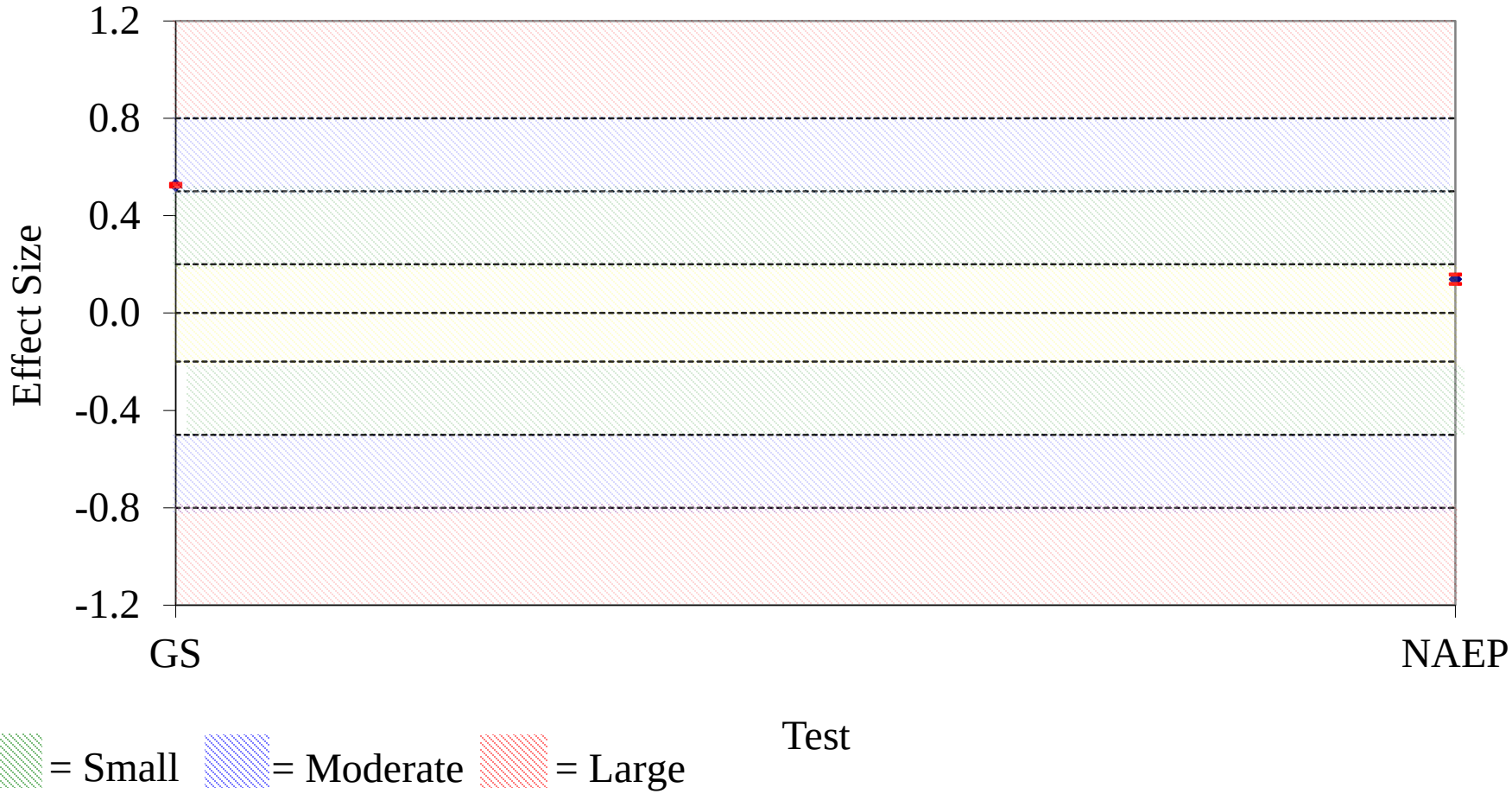
OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs
Content Area = Reading/Verbal
Non-Hispanic Whites Versus Asians



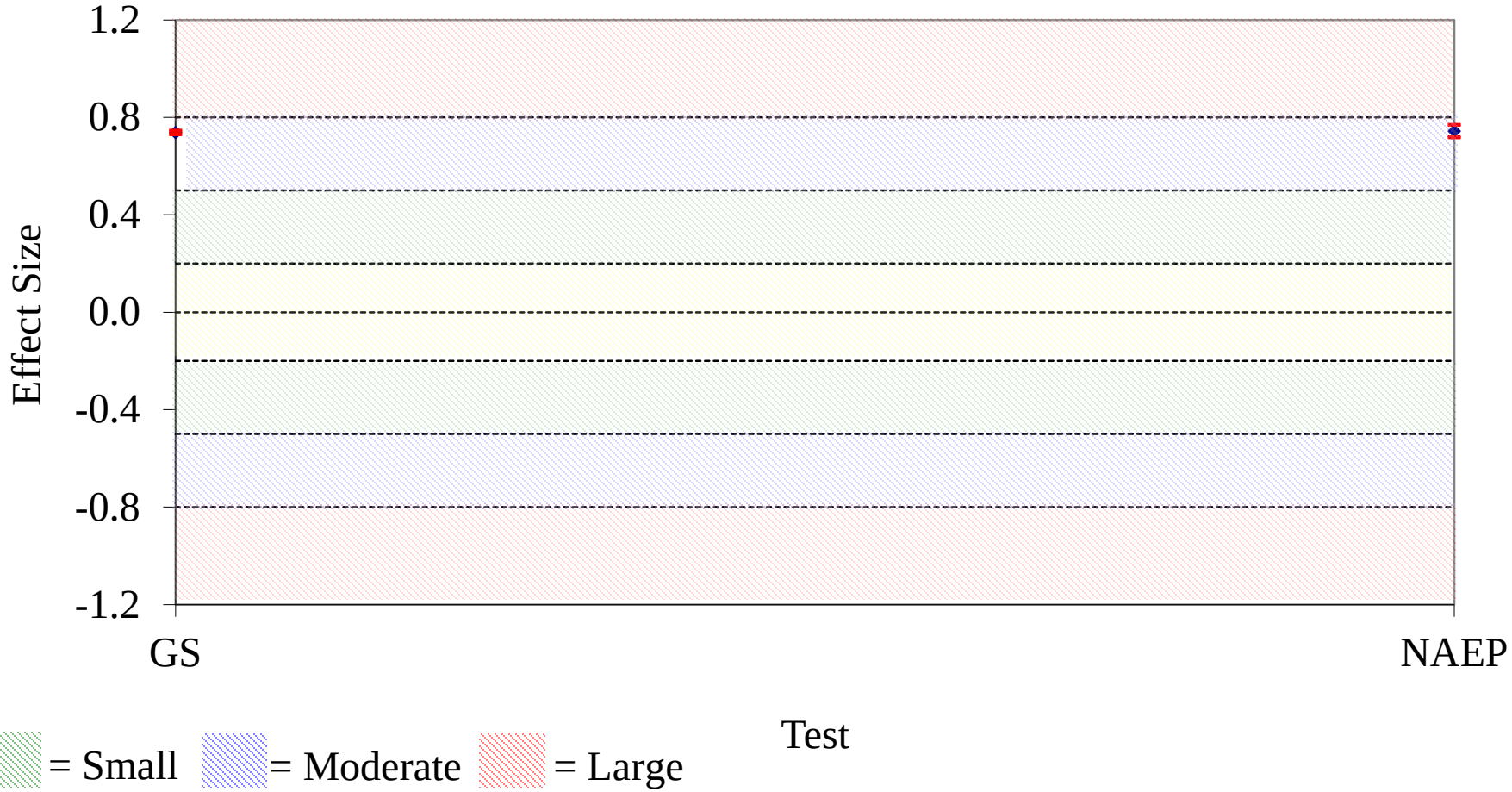
OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs
Content Area = Science
Males Versus Females



OPA
OFFICE OF PEOPLE ANALYTICS

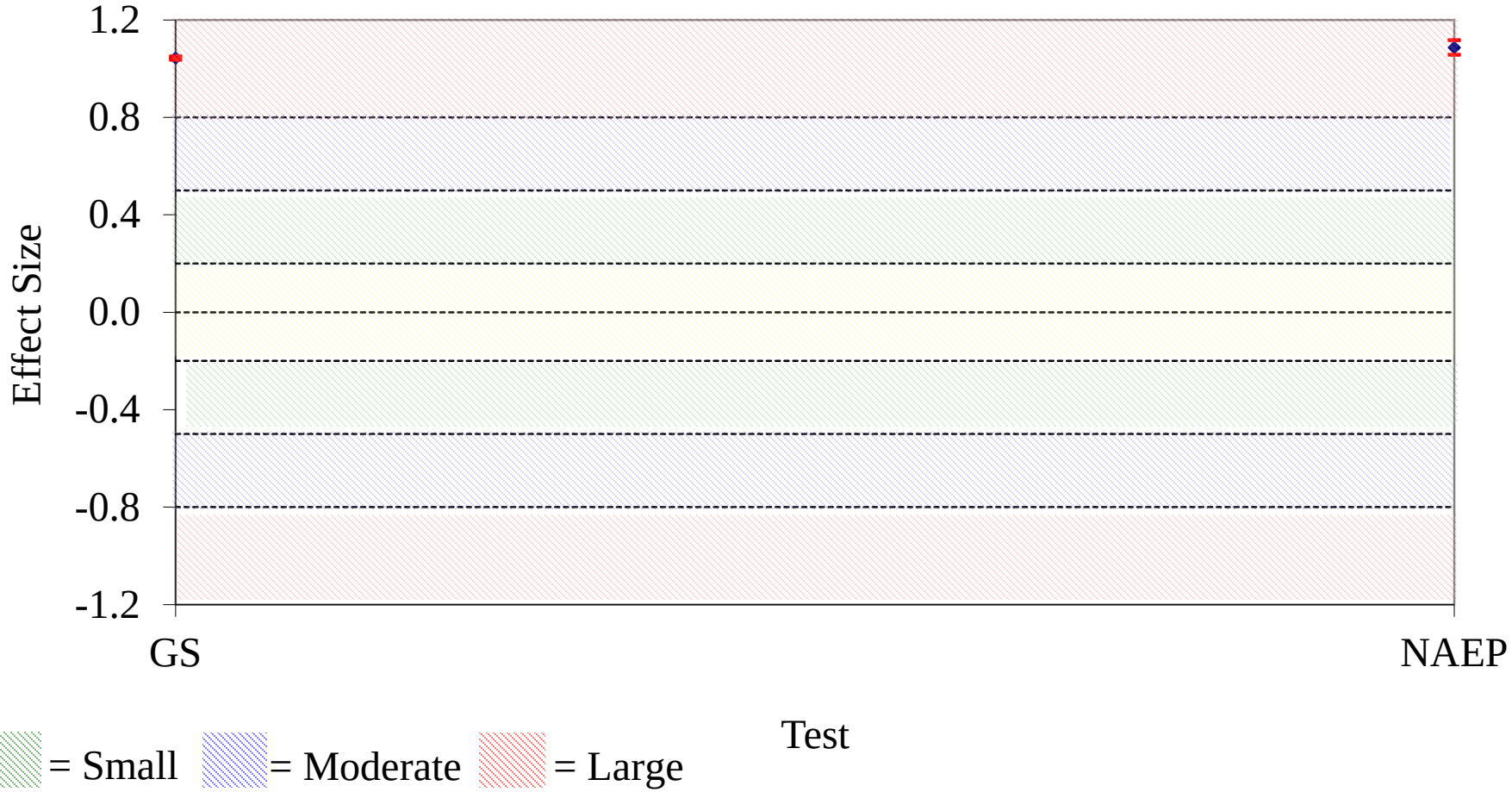
Comparison of Effect Sizes Across Testing Programs
Content Area = Science
Non-Hispanic Whites Versus Hispanics*



OPA
OFFICE OF PEOPLE ANALYTICS

* Hispanics and Whites are mutually exclusive groups for SAT and NAEP.

Comparison of Effect Sizes Across Testing Programs
Content Area = Science
Non-Hispanic Whites Versus Non-Hispanic Blacks



OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs
Content Area = Science
Non-Hispanic Whites Versus Asians

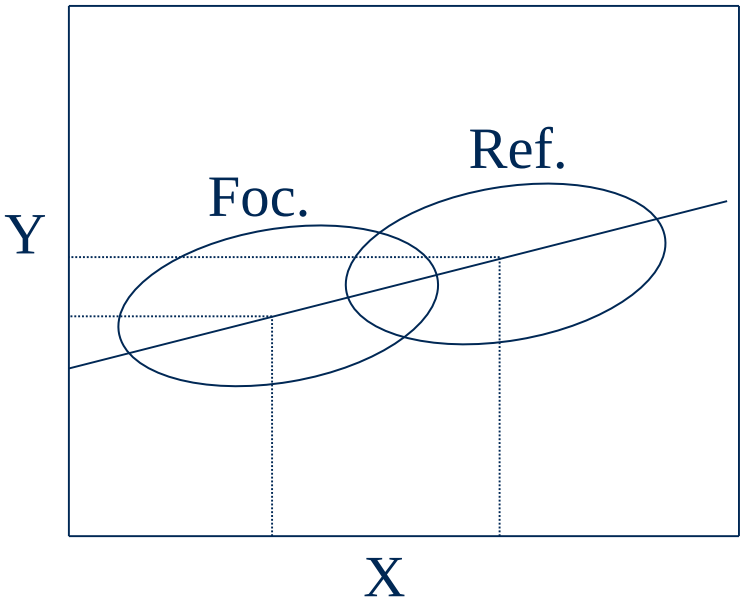
OPA
OFFICE OF PEOPLE ANALYTICS

CONCLUSIONS AND CAVEATS

- **For the AFQT tests (and GS), the direction and magnitude of overall impact is largely consistent with that observed on comparable SAT and NAEP tests, which suggests that impact on ASVAB tests may reflect legitimate differences in the studied groups.**
  - Comparisons across programs may be somewhat restricted due to differences in group definitions, testing populations, test content, etc.
- **"To the extent that members of one group do more poorly on a subtest of items that are a *legitimate part of the content domain*, we would be reluctant to call the discrepancy evidence of *bias*" (Shepard, 1987).**

CONCLUSIONS AND CAVEATS

- **Adverse impact does *not* reflect bias if validity research shows that the test is equally valid for relevant groups.**
  - Historically, a regression-based approach has been advocated to evaluate the existence of bias. Lack of bias is indicated when the regression line relating the test score [X] and a criterion [Y] is the same for each group.



From Ghiselli, Campbell, & Zedeck (1981). Measurement Theory for the Behavioral Sciences.

CONCLUSIONS AND CAVEATS

- **Previous research on the ASVAB technical tests showed similar prediction lines across (1) males and females and (2) blacks and whites (Wise, et al., 1992), suggesting no bias for the tests and groups studied.**
  - DMDC recommended in 2010 that an updated validity study be conducted for relevant tests and groups.
  - Lack of access to criterion data across Services (except Air Force) presents an impediment to updating the study.
  - More recent thinking in the realm of bias detection is that regression-based approaches may not accurately reflect bias.
- **Better to look to the future? Reducing (adverse) impact will be a high priority when considering revisions to the ASVAB and AFQT contents.**

OPA
OFFICE OF PEOPLE ANALYTICS

Tab N

DPAC Device Evaluation for ASVAB

Defense Advisory Committee on Military Personnel Testing

09.21.2018 | Minneapolis, MN

Tia Fechter

DEVICE EVALUATION

- Goals
- Existing Research
- Course of Action Options
- Recommendations
- DAC Role
- Device Evaluation Questions
- Evaluation Design
- Recommendations Discussion

OPA
OFFICE OF PEOPLE ANALYTICS

GOALS

- Facilitate delivery device expansion of the ASVAB iCAT and PiCAT by evaluating examinee performance differences among electronic devices (e.g., tablets, smart phones).
- 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 “Response Design” approach, which automatically formats the test display to alternative devices.

EXISTING RESEARCH

- **Buckland, Becker, & Wiley, 2018**
  - Literature review of studies addressing mode impacts on device usability, item difficulty, and score differences
    - Studies comparing effects of modern electronic devices are sparse
    - Most studies focus on device usability
    - Most studies are found in unpublished literature
  - Considerations that may impact performance*
    - Screen size (minimum of 9.5 inches)
    - Participants' device fluency
    - Item types/features
    - Content visible at one time
    - Device capabilities (e.g., touch screen)
    - Higher test completion times when using mobile devices
  - Simple text-based items tend to not perform differently across devices
  - Consensus on what impacts performance has not been reached

COURSE OF ACTION OPTION 1

- Proceed with implementing operational device expansion for the ASVAB testing platform with no additional research efforts

Strengths

- Significantly reduces evaluation effort costs
- Cuts the time to operationalize expansion by one year

Weaknesses

- Degrades confidence in score interpretation/use
- Raises concerns about score comparability for career counseling and enlistment
- Could hurt the perception of ASVAB testing program if score and measurement invariance is not upheld
- May give the false impression that research was already carried out, and any findings to the contrary may weaken image of ASVAB testing program
- Degrades quality testing experience

COURSE OF ACTION OPTION 2

- Proceed with implementing operational device expansion for the ASVAB testing platform for CEP Grade 10
  - Continue with career exploration
  - Use resulting data to analyze expected impact of device expansion on score comparability for enlistment & classification purposes (post-hoc)

Strengths

- Greater flexibility to CEP for inclusion of students
- May reduce the need to conduct evaluation within the MEPS
- May reduce evaluation costs

Weaknesses

- Not a controlled experimental design
- CEP Grade 10 student outcomes are unlikely to generalize to applicant population
- Outcomes of evaluation may show that CEP scores obtained lack measurement invariance
- May give the false impression that research was already carried out, and any findings to the contrary may weaken image of ASVAB testing program
- Testing on various devices may lead to numerous failed test attempts

COURSE OF ACTION OPTION 3

- Conduct an evaluation of performance differences observed across various devices for select ASVAB subtests before implementing any operational device expansion plans

Strengths

- Allows for a controlled experimental design
- Allows for evaluation to occur with the most representative sample—applicants
- Allows for the evaluation of device familiarity as well as measurement invariance issues across devices and operating systems
- Allows for the opportunity to obtain feedback on the "responsive" interface design

Weaknesses

- Increases the time to operational implementation

RECOMMENDATIONS

- Proceed with the device evaluation and explore findings before operational implementation of alternative electronic devices for ASVAB testing for military entrance and career exploration.
- Concurrent with the device evaluation, begin adapting the ASVAB testing platform and interface (to the extent possible) to be compatible with various web browsers.

DAC ROLE

- Evaluate the recommendation to proceed with device evaluation before operationally delivering ASVAB on alternative electronic devices.
- Provide input on the appropriateness of the current evaluation design.
- Assist DPAC with mitigating technical challenges anticipated for carrying out the device evaluation.
- Provide additional ideas for shortening the duration of evaluation efforts.

DEVICE EVALUATION QUESTIONS

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

EVALUATION DESIGN

- Sampling Plan
- Methods
- Analyses

EVALUATION DESIGN—SAMPLING PLAN

- **Participants**
  - Applicants (MEPS)
    - 10 low-volume sites

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	1750
3	F05/F06						X		30	24	1750
4	F07/F08							X	30	30	1750
5	F09/F10	X			X				30	30	1750
6	F11/F12			X				X	30	40	1750
7	F13/F14					X			30	24	1750
8	F15/F16				X				28	14	1750
TOTALS										186	14000

[a] Forms will be administered using a counterbalancing design.
[b] AI, EI, and SI are not included. GS is intended to represent results for these four subtests.
[c] Test Time and Number of Items is cumulative between the two forms.

EVALUATION DESIGN—SAMPLING PLAN

- **Challenges**
  - Access to evaluation participants
  - Representativeness
    - MEPS with access to WiFi
    - Smaller MEPS
  - Generalize to CEP
  - Generalize to subtests and items not included
  - Motivation
    - Pulling applicants while they wait for next application procedure
  - Time constraints

EVALUATION DESIGN—METHODS

- **Devices**
  - 7 devices selected based on results of GUI usability evaluation
    - 1 Windows-based PC (control condition)
    - 2 Laptops
    - 2 Tablets
    - 2 Smart phones

- **Device Assignment Strategy**
  - Each participant takes two test forms, one on each of two randomly assigned devices. Forms are parallel, consisting of ASVAB items from a selected number of subtests.
  - Application will allow for two different logins for each participant, one per device—each participant serves as his/her own control

Name	Condition Code	Access Code 1	Device 1/ Admin 1	Access Code 2	Device 2/ Admin 2
Doe, Jane	1	**C01**D1**F0** 145199**A1**	Windows 10 Desktop	**C01**D7**F02** 45199**A2**	iPhone X
…	…	…	…	…	…
Smith, Joe	8	**C08**D3**F0** 888206**A1**	Microsoft Surface Pro	**C08**D2**F07** 88206**A2**	MacBook Pro

EVALUATION DESIGN—METHODS

- **Form Administration Order**
  - Each participant is randomly assigned to a condition code 01–16, which allows for counterbalanced administration of forms.
  - Each participant will have two access codes that contain information about group, device, and form assignment.

Name	Condition Code	Access Code 1	Device 1/ Admin 1	Access Code 2	Device 2/ Admin 2
Smith, Joe	8	**C08**D3**F0** 8**8**8206**A1**	Microsoft Surface Pro	**C08**D2**F0** 7**8**8206**A2**	MacBook Pro

EVALUATION DESIGN—METHODS

• Forms Development

Subtest	P&P # of Items
AO	25
AR*	30
GS	25
MC	25
MK*	25
PC	15
WK	35

- Eight pairs of forms are developed for each ASVAB subtest to be parallel to one another
- Contain the same number of items that a P&P form would have*
- Adheres to table of specifications for the P&P forms
- Over samples items with special features (e.g., extended text length) that may result in difficulty differences when administered on different electronic devices

• Form Subsets

Form	Number of Items from Each Subtest	
F01/F02	12 AR	
F03/F04	12 AR	
F05/F06	24 MC	
F07/F08	30 AO	
F09/F10	20 GS	10 PC
F11/F12	20 WK	20 AO
F13/F14	24 MK	
F15/F16	14 PC	

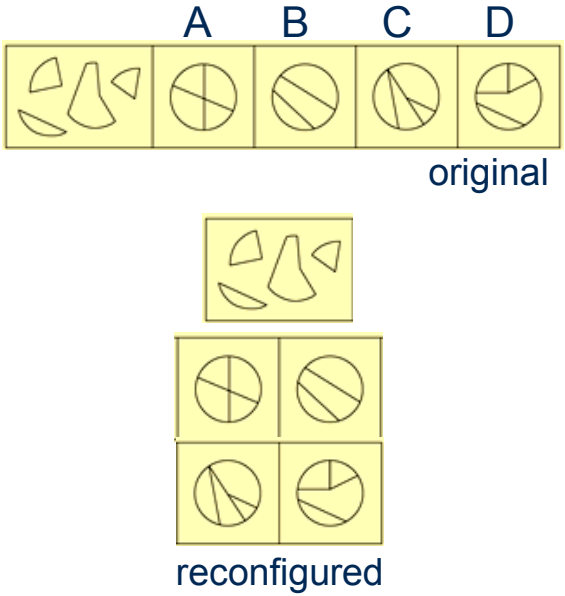
- Eight pairs of forms are developed that contain some items from a selection of ASVAB subtest pairs
- Items are selected based on the proportion of items from each subtest that contains items with special features**
- Over samples items with special features
- Includes some text-only items as controls

EVALUATION DESIGN—METHODS

- **Item Special Features**

Special Feature	Relevant Subtests
Graphic	AO, AR, EI, GS, MC, MK, AI, SI
Reconfigured Graphic	AO
Complex Graphic*	GS, MC, MK
Answer Choice as Graphic	AO, MC, SI
Long Stems/Extended Text	AR, PC
Stacked Fractions	AR, MK
Equation	MK
Square Root	MK
Exponents	MK
Pi	MK
Degree Symbol	MK

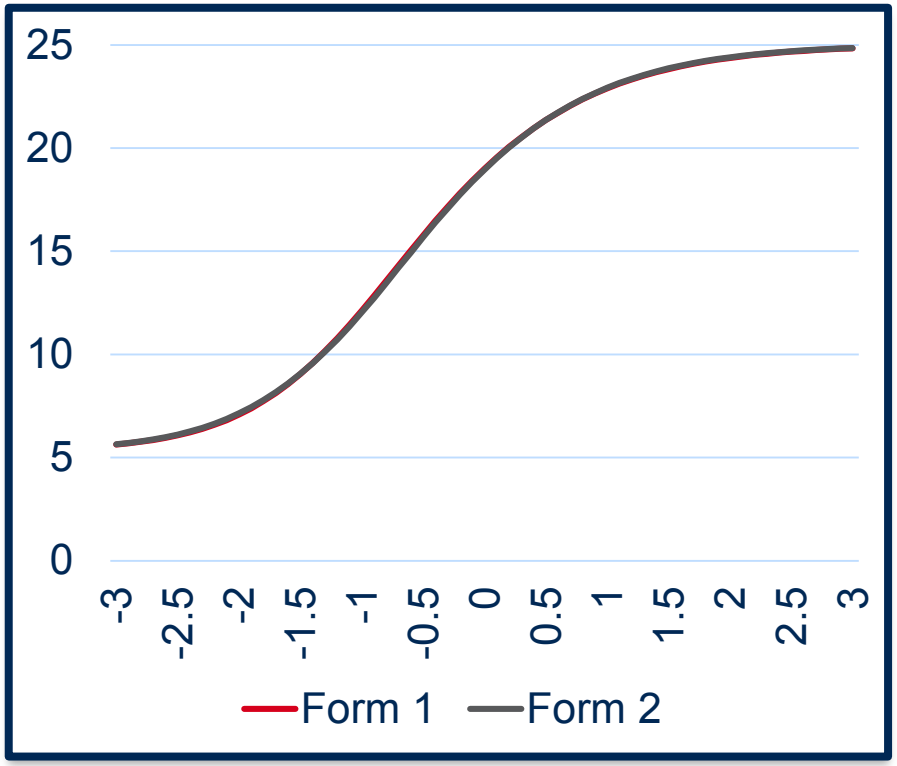
Reconfigured Graphic Example**

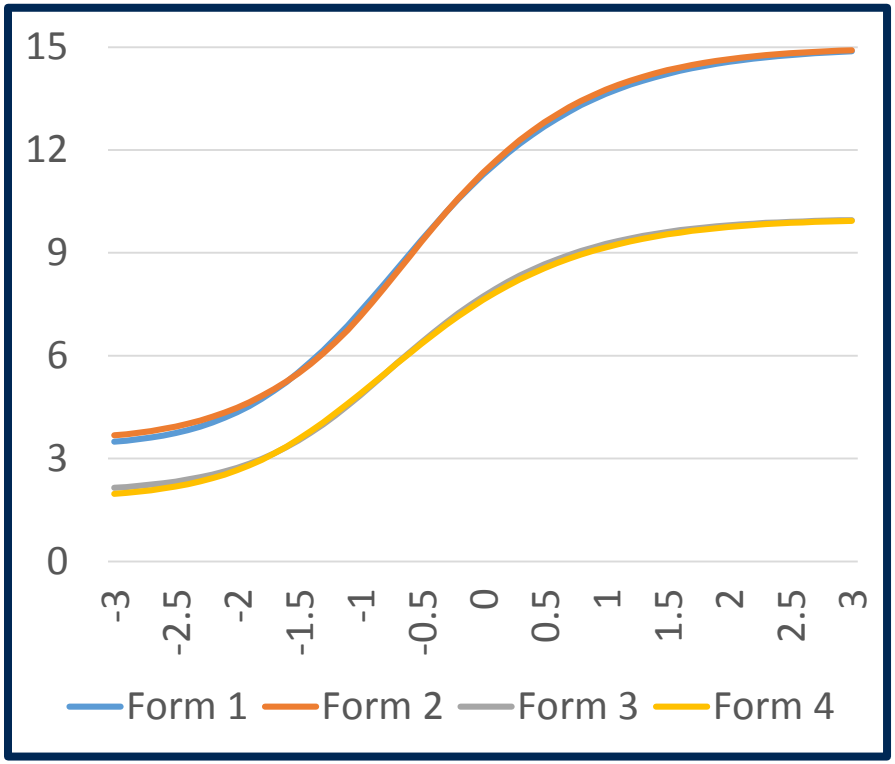


OPA
OFFICE OF PEOPLE ANALYTICS

EVALUATION DESIGN—METHODS

- **Test Characteristic Curves for AO Form Pairs**
  - Average TCC difference* between full forms = 0.04
  - Average TCC difference* between selected items: F1&F2 = 0.10; F3&F4 = 0.07

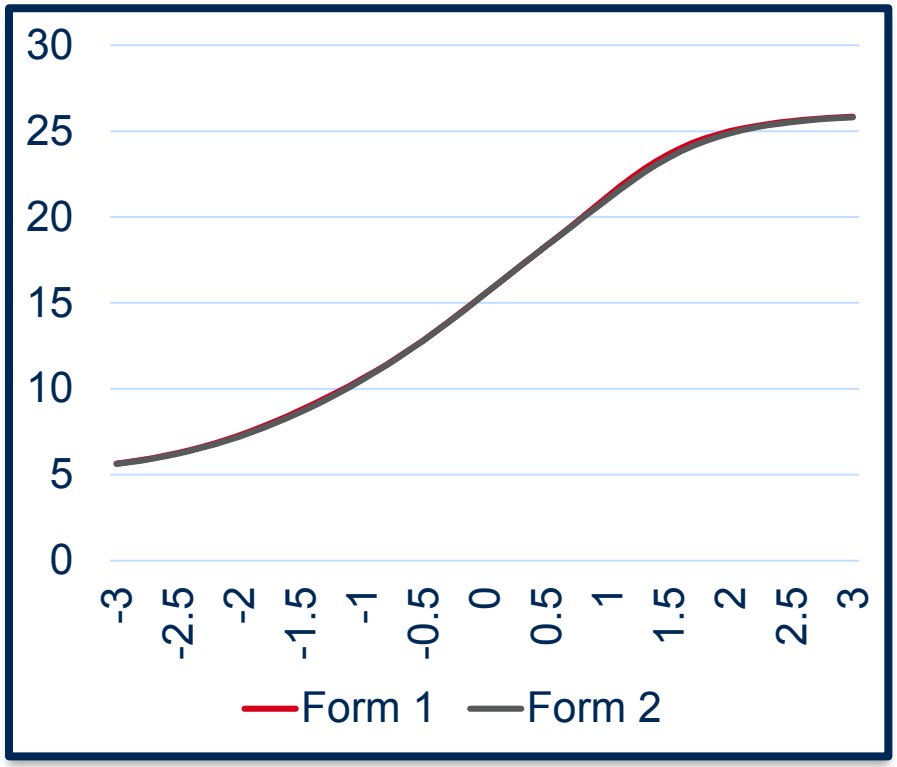


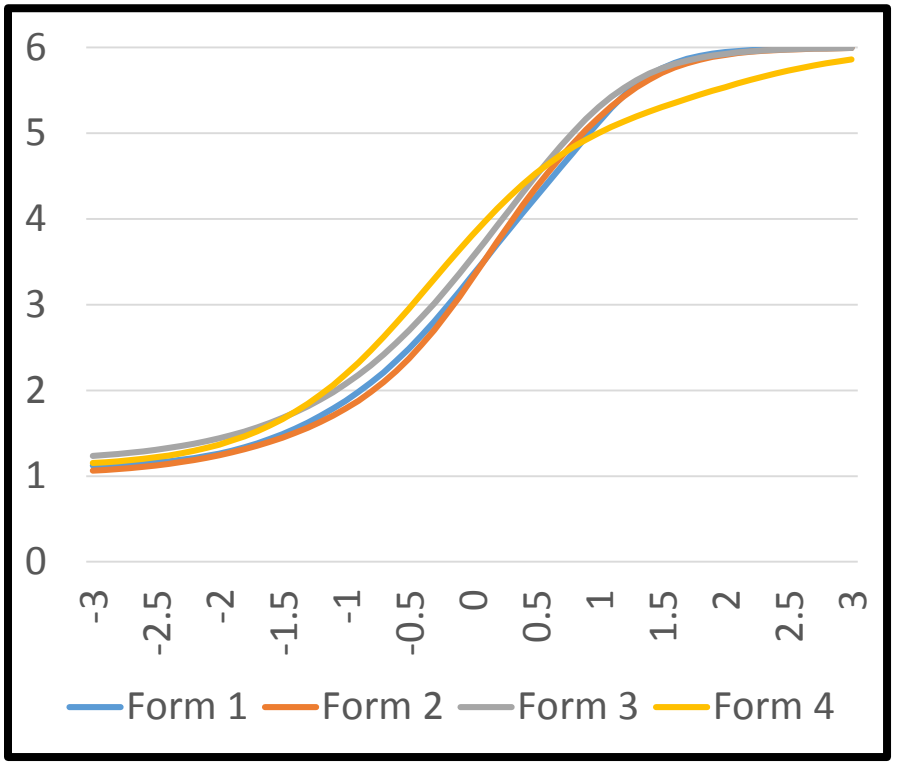


OPA
OFFICE OF PEOPLE ANALYTICS

EVALUATION DESIGN—METHODS

- **Test Characteristic Curves for AR Form Pairs**
  - Average TCC difference* between full forms = 0.07
  - Average TCC difference* between selected items: F1&F2 = 0.05; F3&F4 = 0.20

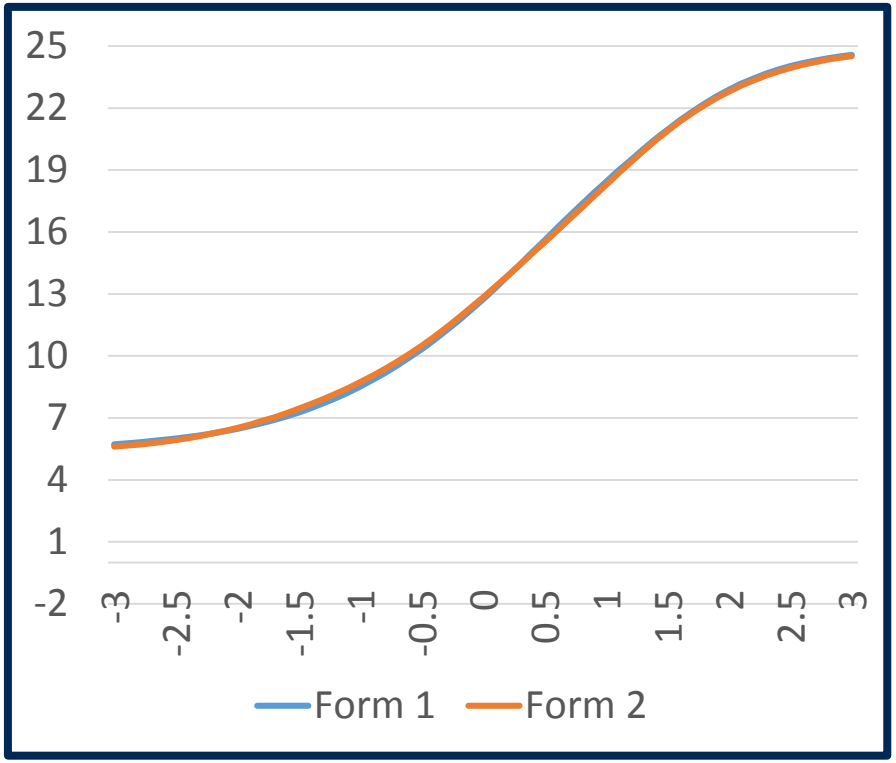


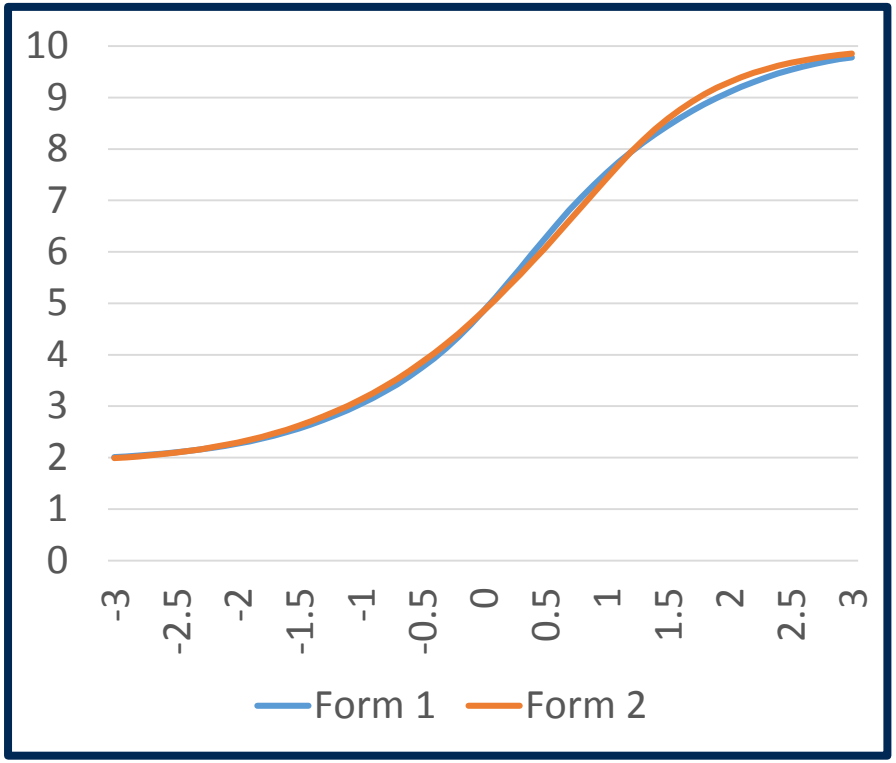


OPA
OFFICE OF PEOPLE ANALYTICS

EVALUATION DESIGN—METHODS

- **Test Characteristic Curves for GS Form Pairs**
  - Average TCC difference* between full forms = 0.11
  - Average TCC difference* between selected items = 0.09

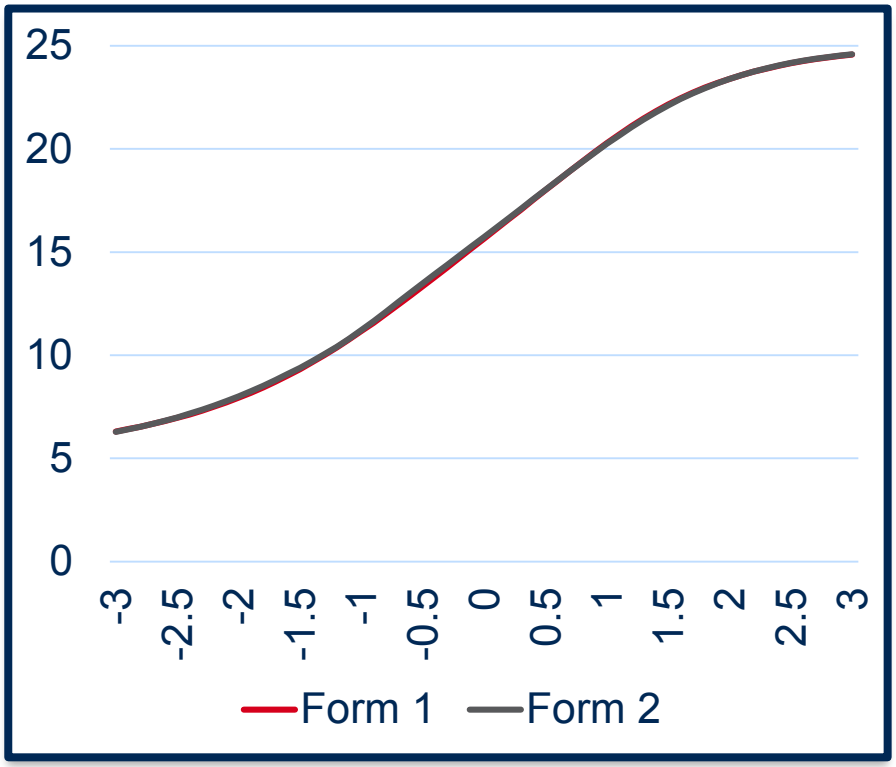


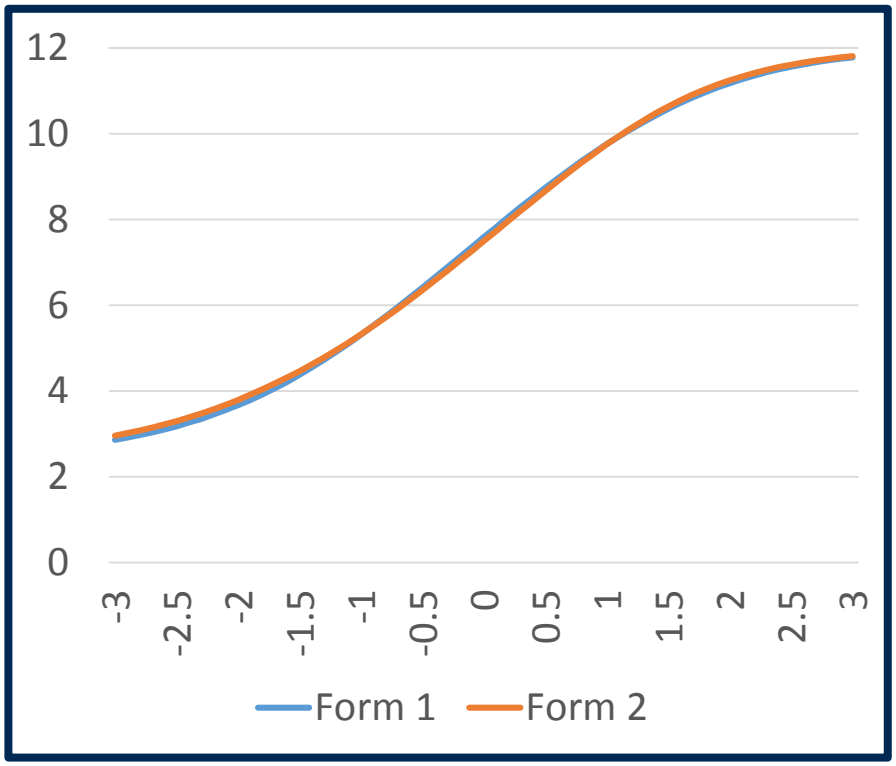


OPA
OFFICE OF PEOPLE ANALYTICS

EVALUATION DESIGN—METHODS

- **Test Characteristic Curves for MC Form Pairs**
  - Average TCC difference* between full forms = 0.05
  - Average TCC difference* between selected items = 0.07

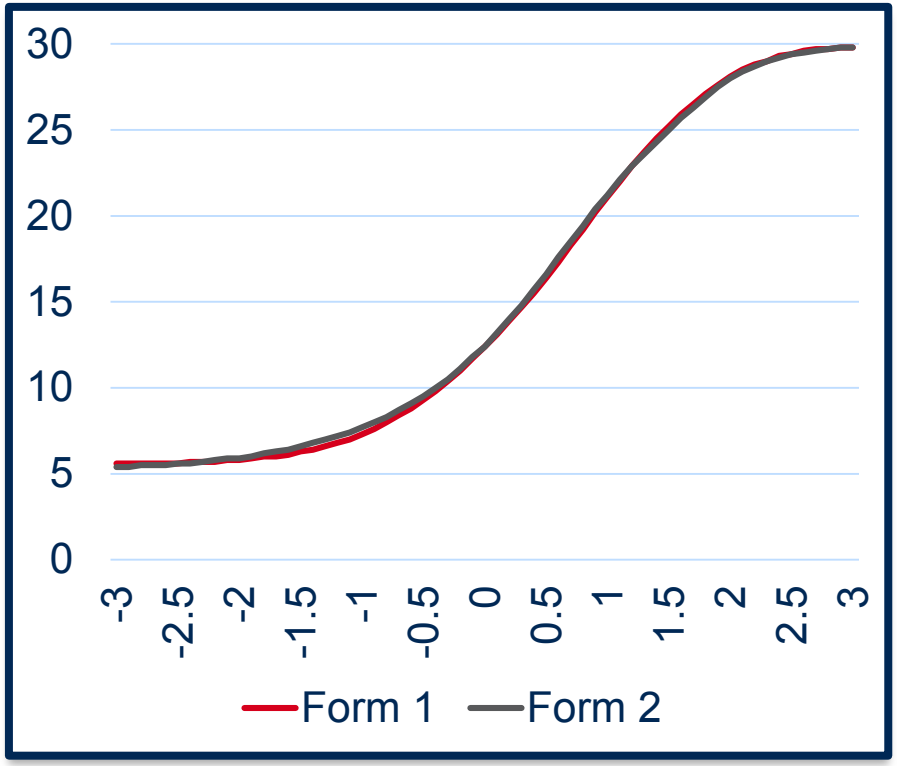


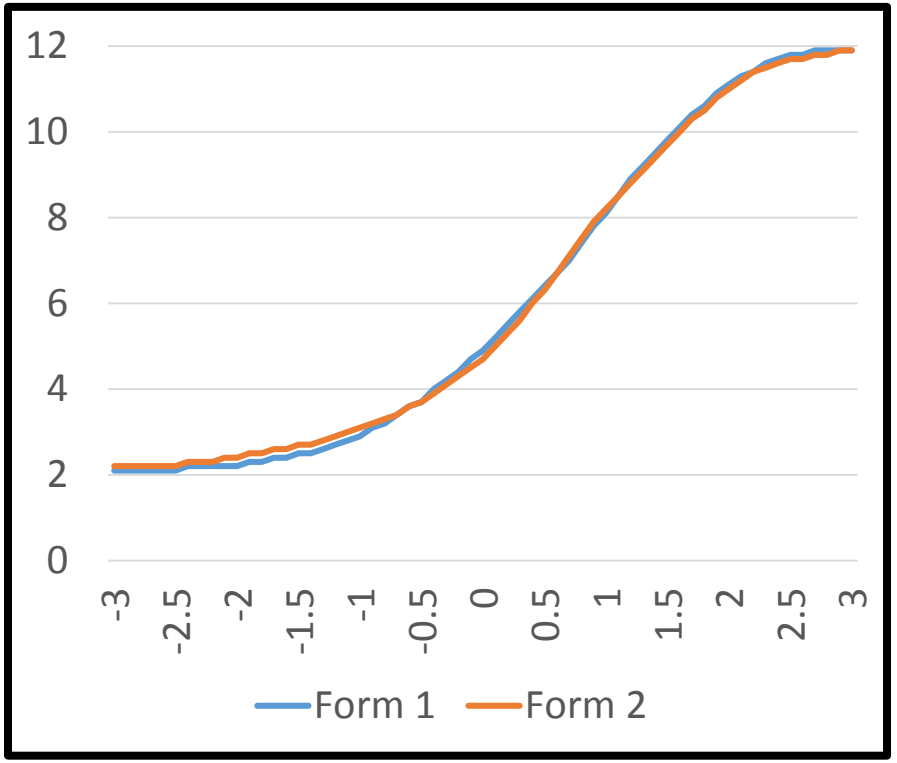


OPA
OFFICE OF PEOPLE ANALYTICS

EVALUATION DESIGN—METHODS

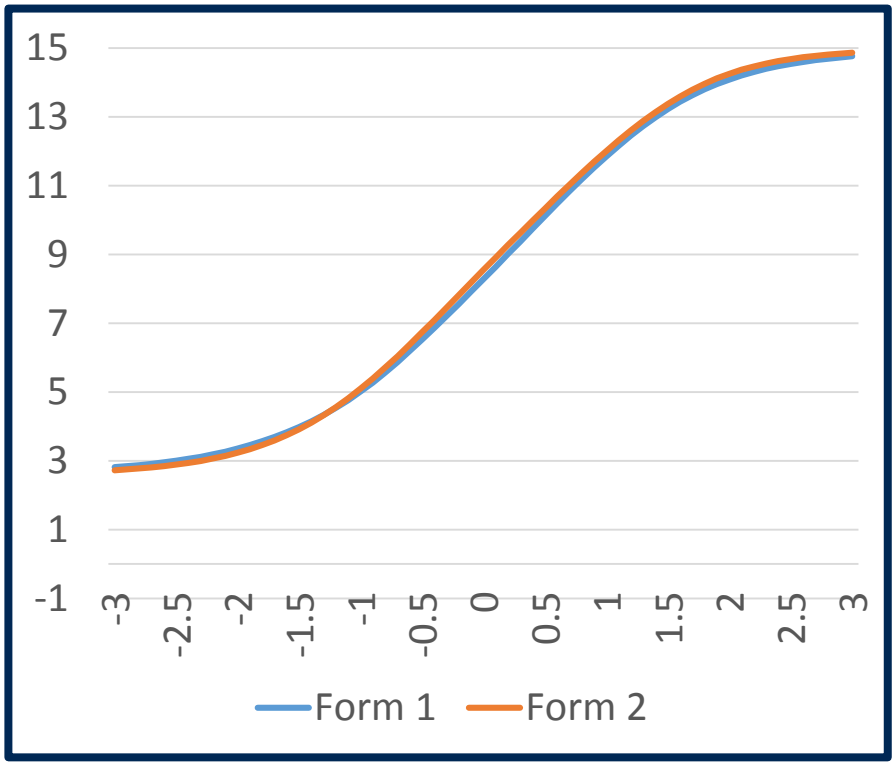
- **Test Characteristic Curves for MK Form Pairs**
  - Average TCC difference* between full forms = 0.16
  - Average TCC difference* between selected items = 0.12

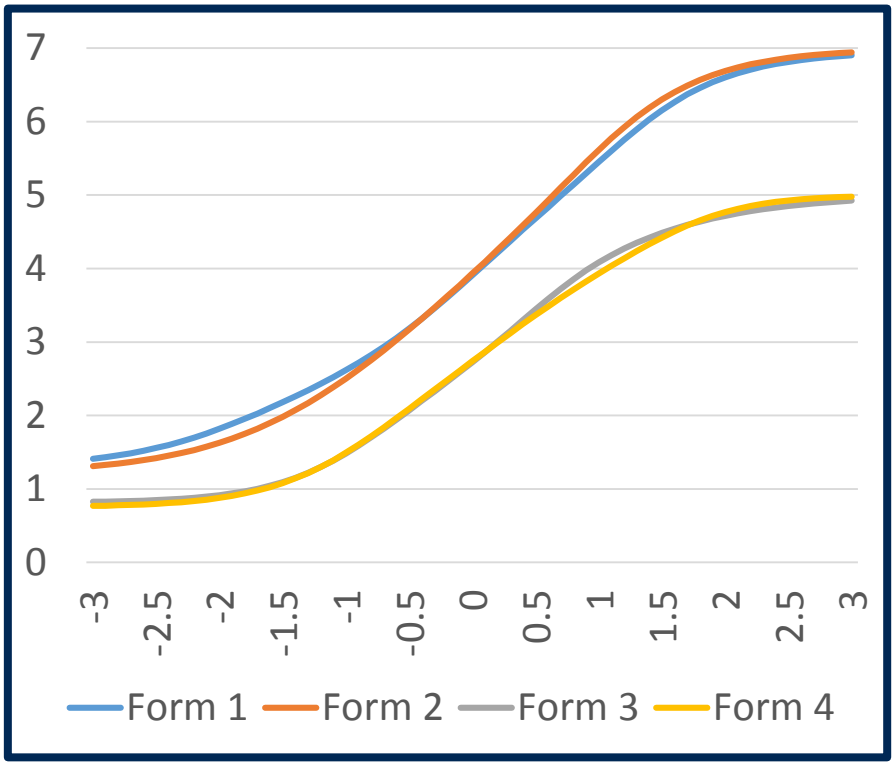




EVALUATION DESIGN—METHODS

- **Test Characteristic Curves for PC Form Pairs**
  - Average TCC difference* between full forms = 0.15
  - Average TCC difference* between selected items: F1&F2 = 0.11; F3&F4 = 0.05

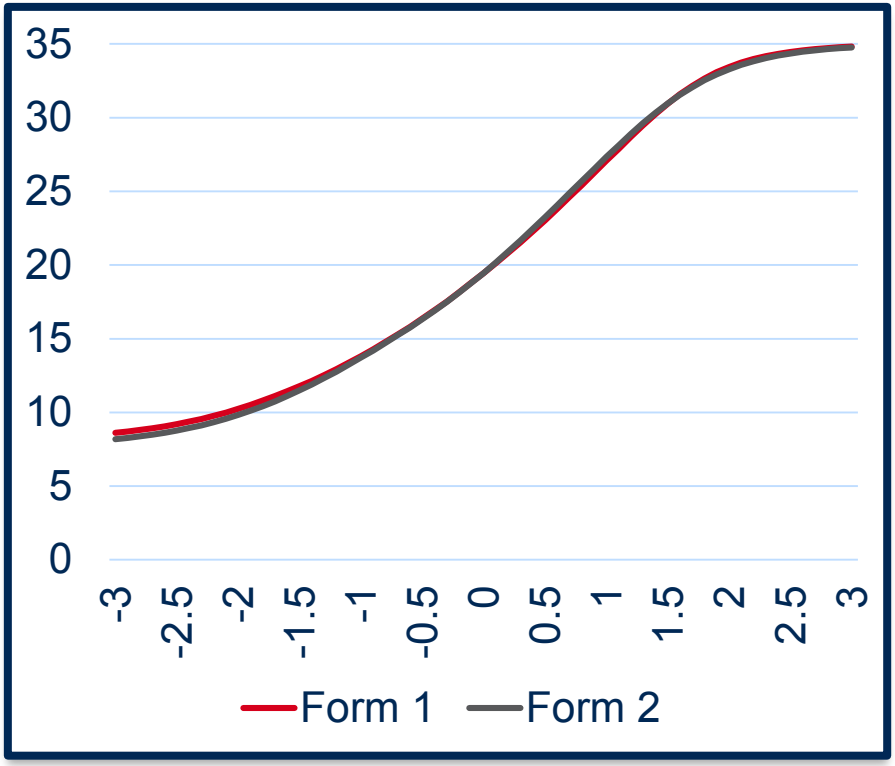


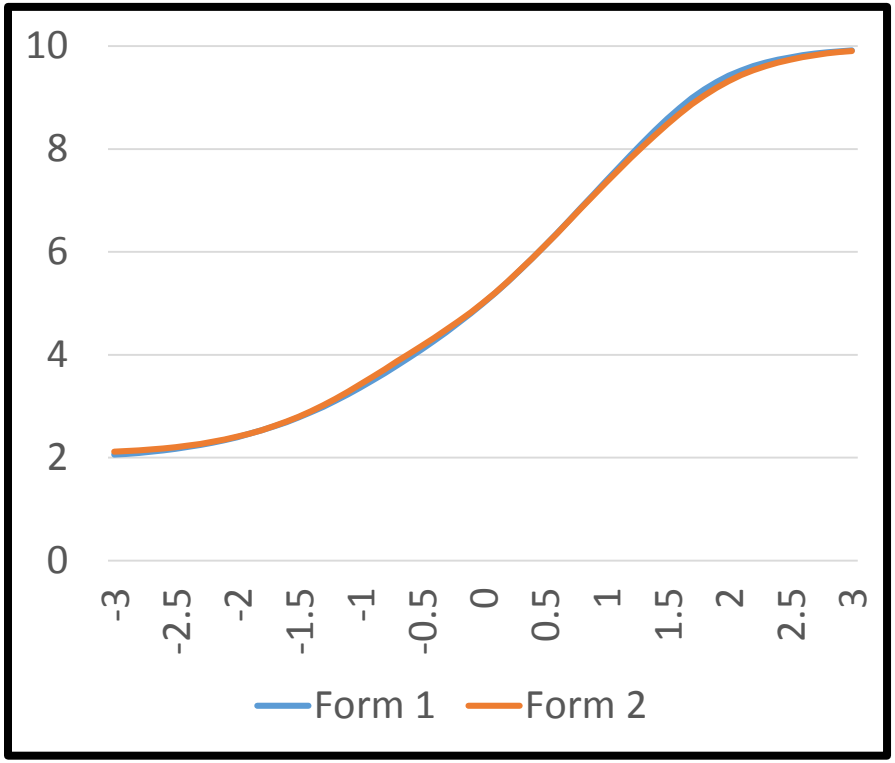


OPA
OFFICE OF PEOPLE ANALYTICS

EVALUATION DESIGN—METHODS

- **Test Characteristic Curves for WK Form Pairs**
  - Average TCC difference* between full forms = 0.20
  - Average TCC difference* between selected items = 0.04





OPA
OFFICE OF PEOPLE ANALYTICS

EVALUATION DESIGN—METHODS

- **Form Delivery Order**

Form	Number of Items from Each Subtest		Order of Priority
F01/F02	12 AR		4
F03/F04	12 AR		4
F05/F06	24 MC		2
F07/F08	30 AO		4
F09/F10	20 GS	10 PC	2
F11/F12	20 WK	20 AO	1
F13/F14	24 MK		3
F15/F16	14 PC		2

- Deliver F11/12 to participants first.
  - WK represents worst-case scenario where our hypothesis is that we should not see any differences in item performance due to device delivery method.
    - If we do, there is little value in moving forward with the entirety of the evaluation.
  - AO represents the best-case scenario where our hypothesis is that AO item types are the most likely to show differences in item performance due to device delivery method (based on past research with these types of items and their configuration).
    - If we don't observe differences between devices for AO, we gain confidence in moving forward with operationalizing the device expansion.

EVALUATION DESIGN—METHODS

- **Post-Test Feedback**
  - Collected electronically following each administration
  - Collected following each device (Device 1 & 2) administration

Motivation

What was your motivation to answer questions correctly while taking this test? Choose the statement you agree with most.

- I answered all questions to the best of my ability.
- I answered most questions to the best of my ability.
- I answered a few of the questions to the best of my ability.
- I did not answer any questions to the best of my ability.

EVALUATION DESIGN—METHODS

- **Post-Test Feedback**
  - Following Device 2 administration

Perception of Device

You took a version of the same test using two different electronic devices. Choose which statement you agree with most.

- I performed better on the test using DEVICE 1.
- I performed better on the test using DEVICE 2.
- My performance on the test was the same using DEVICE 1 and DEVICE 2.
- My performance on the test was different using DEVICE 1 and DEVICE 2, but my performance was **NOT** impacted by the devices I used to take the test.

EVALUATION DESIGN—METHODS

- **Post-Test Feedback**
  - Following Device 2 administration

Perception of Device

You took a version of the same test using two different electronic devices. Choose which statement you agree with most.

- It was easier to use DEVICE 1.
- It was easier to use DEVICE 2.
- Both DEVICE 1 and DEVICE 2 were easy to use.
- Both DEVICE 1 and DEVICE 2 were difficult to use.

EVALUATION DESIGN—METHODS

- **Post-Test Feedback**
  - Following Device 2 administration

Perception of Device

I am comfortable using a **TABLET** to take tests.

- Agree
- Disagree

EVALUATION DESIGN—METHODS

- **Post-Test Feedback**
  - Following Device 2 administration

Familiarity with Device

- Check the boxes next to devices that you use on a regular basis. Check as many as apply.
  - ❑ Phone 1
  - ❑ Phone 2
  - ❑ Tablet 1
  - ❑ Tablet 2
  - ❑ Notebook 1
  - ❑ Notebook 2
  - ❑ Desktop PC

EVALUATION DESIGN—METHODS

- **Post-Test Feedback**
  - Following Device 2 administration

Background Questions

Do you consider yourself to be fluent or near-fluent in the English language?

- Yes
- No

What is the highest level of education attained by at least one of your parents?

- Less than high school
- Some high school, no diploma
- High school graduate
- Some college, no degree
- Associate's degree (for example: AA, AS)
- Bachelor's degree (for example: BA, AB, BS)
- Master's degree (for example: MA, MS, MEng, MEd, MSW, MBA)
- Professional degree (for example: MD, DDS, DVM, LLB, JD)
- Doctorate degree (for example: PhD, EdD)

EVALUATION DESIGN—METHODS

- **Next Steps & Considerations**
  - Training of TAs
  - Develop procedure for collecting appropriate information for ASVAB score matching
  - Results using newer devices for the evaluation may not generalize for those who use older devices that are more commonly available to applicants and students in lower SES groups
  - Pulling one or two applicants at a time for participation in evaluation may result in inefficient use of TA time

EVALUATION DESIGN—ANALYSES

- **Item-Level Comparisons**
  - Item difficulty
  - Item information
  - Area between ICCs
  - Response time
  - DIF
    - Sex
    - Race/Ethnicity
    - Parent education level*
    - EL status

- **Score-Level Comparisons**
  - Differences within participant and between devices
  - Factor analysis: measurement invariance between devices
  - ASVAB subtest score correlations
  - Device familiarity
  - Motivation

- **Feature-Level Analysis**
  - Statistical models to estimate impact of systematic difficulty differences due to item features for groups of items

EVALUATION DESIGN—ANALYSES

- **Challenges & Next Steps**
  - Ensuring participants accurately report their level of motivation
  - Estimating score differences using minimal items administered
    - Relies on the accuracy of the assumption that items without special features will perform similarly across all electronic devices
    - Relies on the assumption that participants are equally motivated during the evaluation as they are when taking the ASVAB for a score of record
  - DIF analyses may not be feasible if sample sizes do not permit

RECOMMENDATIONS DISCUSSION

- Proceed with the device evaluation and explore findings before operational implementation of alternative electronic devices for ASVAB testing for military entrance and career exploration.
- Concurrent with the device evaluation, begin adapting the ASVAB testing platform and interface (to the extent possible) to be compatible with various web browsers.

OFFICE OF PEOPLE ANALYTICS

Appendix

09.21.2018

PRIVACY ACT STATEMENT

PRIVACY ACT STATEMENT

AUTHORITY: 10 U.S.C 136, Under Secretary of Defense for Personnel & Readiness; DoD Instruction 1304.12E, DoD Military Personnel Accession Testing Programs; and E.O. 9397 (SSN), as amended.

PURPOSE: To collect information to inform policy decisions regarding changes to standardized test administration practices. Your social security number is used to link past performance to performance during administration evaluation conditions for comparative purposes.

ROUTINE USE(S): Disclosure of records to Federally Funded Research and Development Centers for the purpose of statistical research and reporting in which individuals will not be identified. Additional routine uses are listed in the applicable system of records notice DMDC-15-DOD, Armed Services Military Accession Testing, and is located at: http://dpcld.defense.gov/Privacy/SORNsIndex/DOD-wide-SORN-Article-View/Article/570568/dmdc-15-dod/

DISCLOSURE: Participating in this study is voluntary. There is no penalty if you choose not to participate. However, maximum participation is encouraged so that data will be complete and representative. If you do not authorize disclosure for the purposes described above you will not be allowed to take the test(s).

If you agree, click the Agree button. Otherwise, click the Help button.

EVALUATION DESIGN—METHODS

- **Form Administration Order**
  - Each participant is randomly assigned to a condition code 01–16. Codes indicate
    - which pair of forms (e.g., F07 and F08) and
    - the order the forms are administered
  - Allows for counterbalanced administration of forms
  - Example: C08
    - Forms F07 and F08 are administered
    - F08 is administered first
    - F07 is administered second

- **Access Code Composition**
  - The first three characters indicate randomly assigned condition code values (C01–C16).
  - The next two characters indicate which device the participant should receive (D1–D7).
  - The next three characters indicate which form the participant should receive (F01–F08).
  - Characters 7–11 are a unique numerical identifier that is randomly generated and held constant for the participant
  - The last two characters indicate which administration the code is used for (A1 or A2). The first administration is coded A1, and second administration is coded A2. This information will signal the system as to which exit feedback questions should be asked following the administration.

Name	Condition Code	Access Code 1	Device 1/ Admin 1	Access Code 2	Device 2/ Admin 2
Smith, Joe	08	**C08**D3**F0 8**88206**A1**	Microsoft Surface Pro	**C08**D2**F0 7**88206**A2**	MacBook Pro

EVALUATION DESIGN—SAMPLING PLAN

- **Sample Size Justification**
  - Item calibrations will require at least 500 examinees per condition, implementing a 3-PLM
  - Comparing means: subset number correct scores
    - Desktop PC (control) vs. 6 devices (treatments)
    - E.g., $\alpha=0.05,\ 1-\beta=0.85,\ \text{Effect Size} = 0.10$
      - n = 450 per group
  - Comparing proportions: item difficulty (b-values converted to p-values)**
    - E.g., $\alpha=0.05,\ 1-\beta=0.85, p_A = 0.50, p_B = 0.60$
      - n = 443 per group

Tab O

ASVAB WK AIG

DAC-MPT Progress Report

Minneapolis, September 20, 2018

Presented by
Isaac Bejar
ETS

WK Report
Co-PI: Michael Flor (NLP)

Linguistics: Paul Deane

Project Management: James Bruno

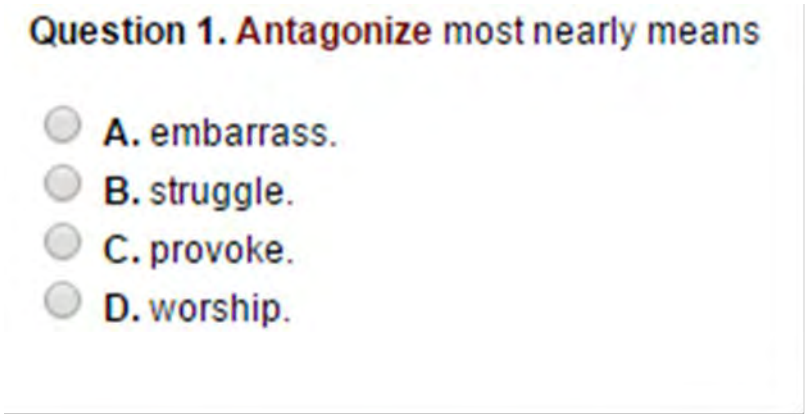
Psychometrics:
Dan McCaffrey, Jonathan Weeks

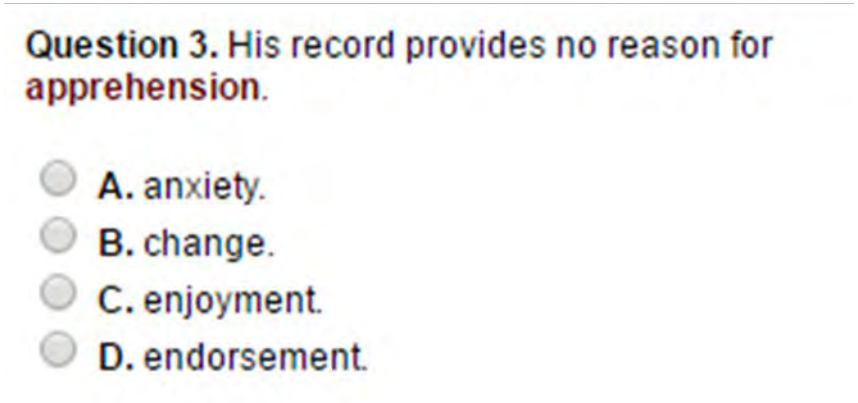
Data Analyst: Steven Holtzman

Test Development: Adam Banta, Serguei Denissov

Project goals

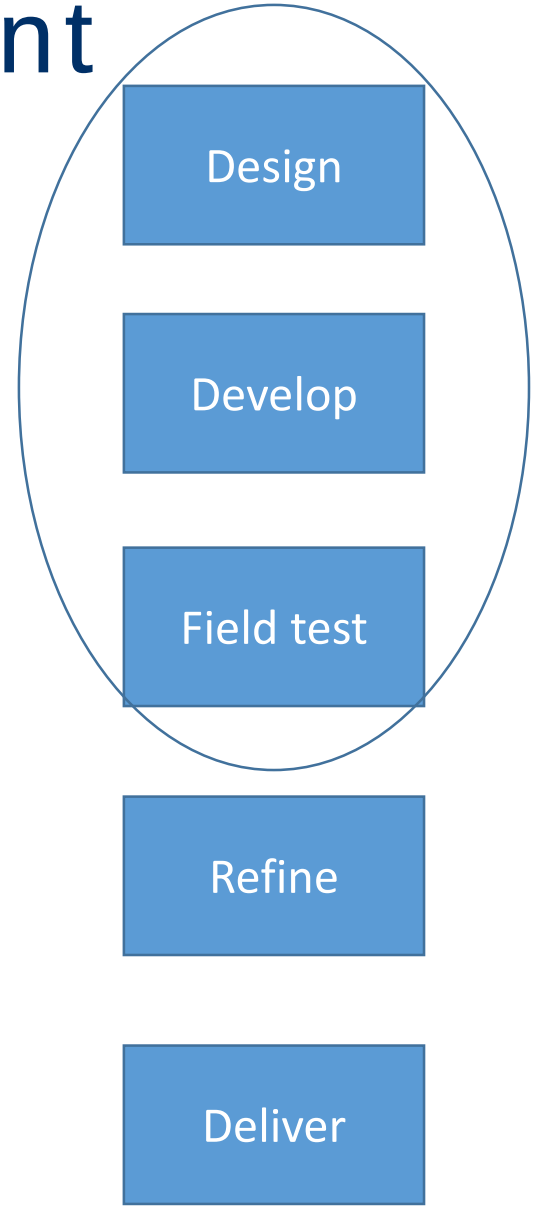
- Automate the production of 4-option ASVAB-WK items (without MWEs)
  - Definitional
  - Contextual
- The expectation is that the generated items will have fairly well estimated difficulty and be *above* a certain level of discrimination
- High level approach
  1. Generate
  2. Predict difficulty
  3. (System limited to items without MWE)





Overview of development process

- Conceptualize approach
- System design
- Field test
- CAT simulations



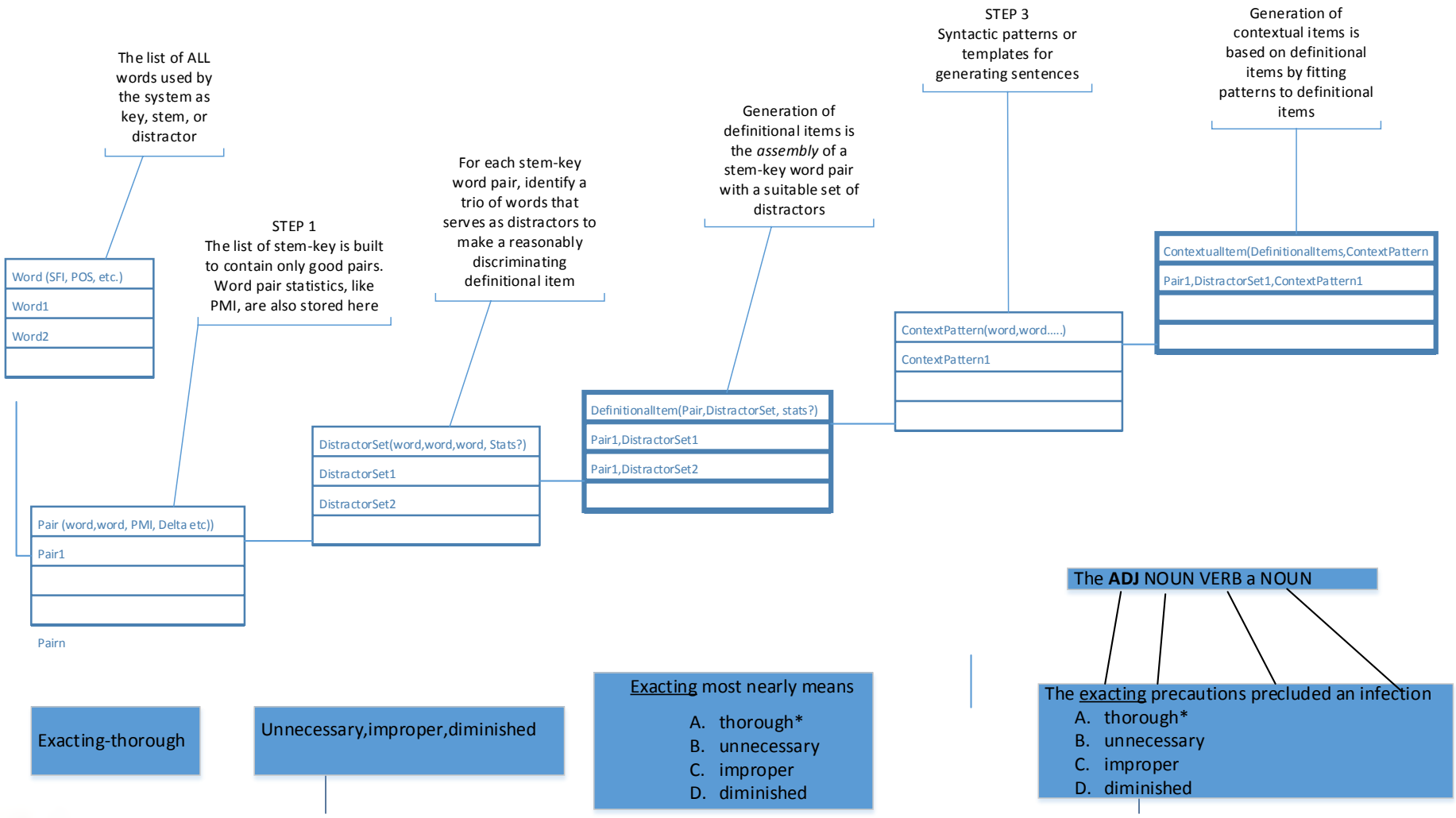
One-year extension of WK work

- 4.2.1 Evaluate WK generated items ☺
- 4.2.2 Refine difficulty model ☺
  - Addition of discrimination model 😐
  - Improved predictors ☺
  - Possible addition of SME estimate ☺
- 4.2.3 Expand templates for contextual items 😐
- 4.2.4 Refine WK generator ☺
  - Definitional generator ☺
  - Contextual generator 😐
  - Evaluation of refined generator (without field testing) ??
- 4.2.5 Generate and review 3000 WK items ☺
  - Securing the items
- System packaging

Approach and considerations

- Automated item Generation (AIG) enhances validity
  - Enhances efficiency and validity
  - Efficiency:
    - Many items produced
    - Items' difficulty variation sufficiently understood to reduce pre-testing
  - Validity
    - Construct *representation*: Accounting of difficulty is grounded on relevant science
      - Difficulty is a function of word familiarity *and* depth of word familiarity
    - Construct *preservation*: Having many more items to work with improves test security
    -

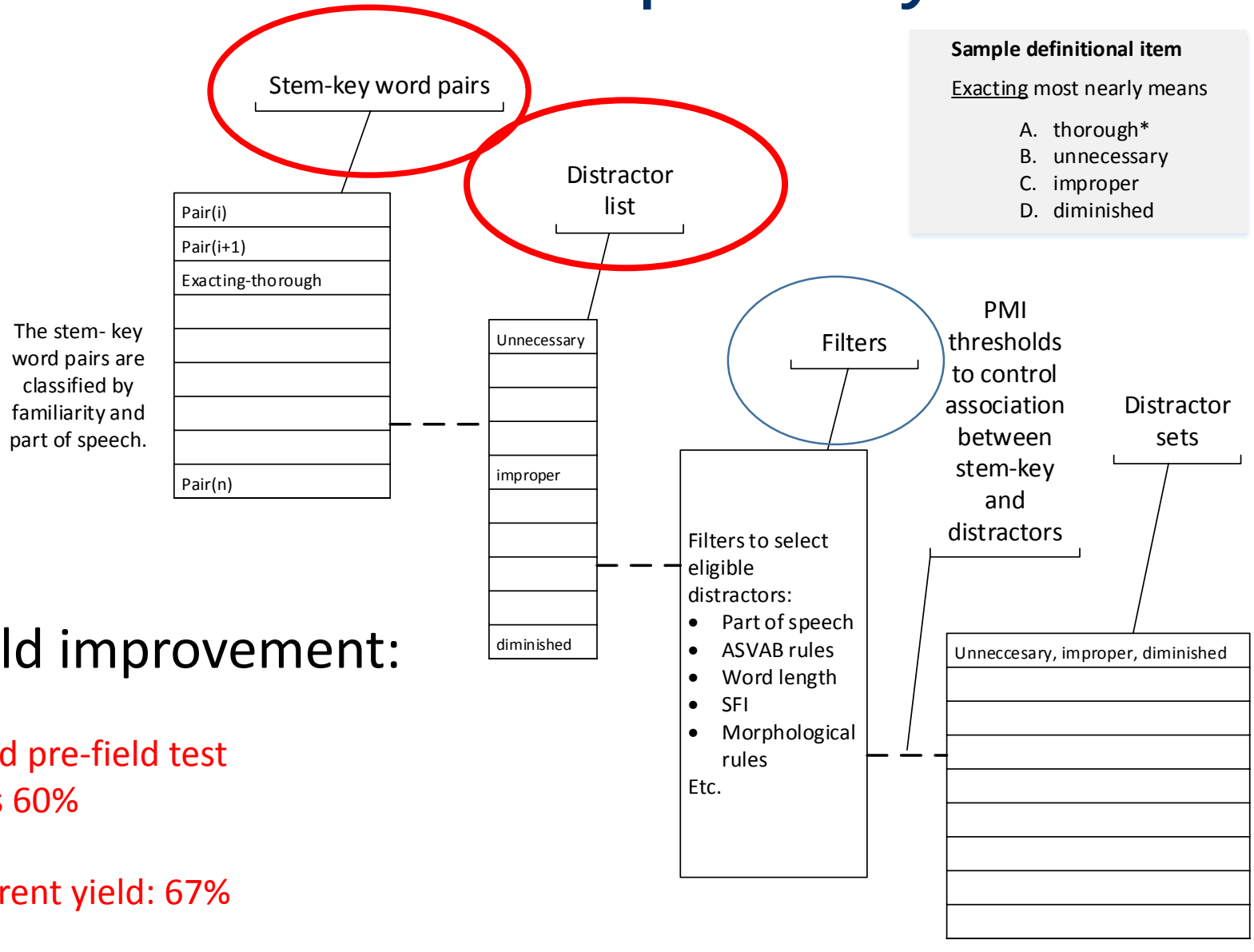
Approach and considerations



Field test results summary

- Generated 1,000 items
- Yield: 60% accepted items by SME
- 10 items with negative biserials
- Difficulty prediction r-squared of 0.26 using GBM
- Discrimination: 75% above of 0.80 (vs. 80% in HumRRO analysis)

Revisions to improve yield



Yield improvement:

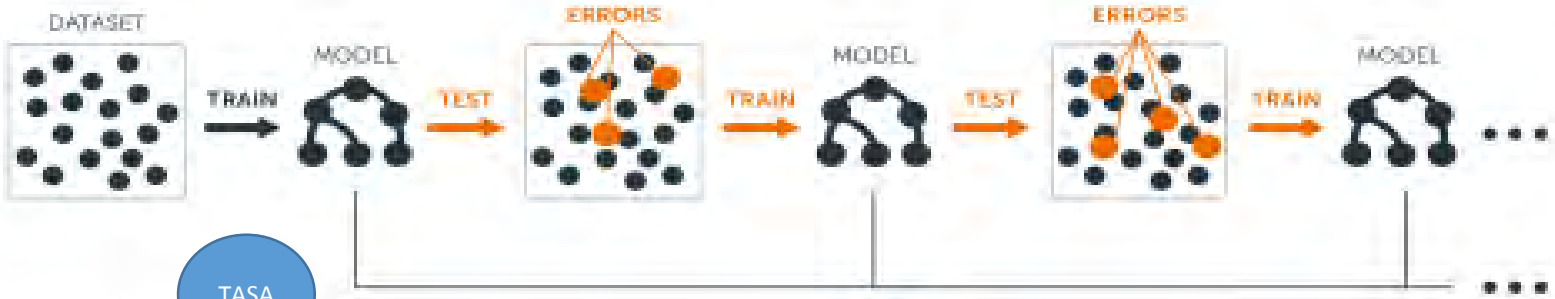
Yield pre-field test was 60%

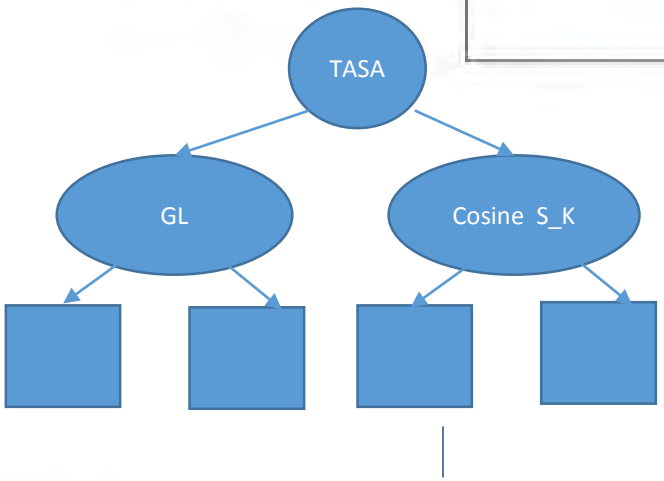
Current yield: 67%

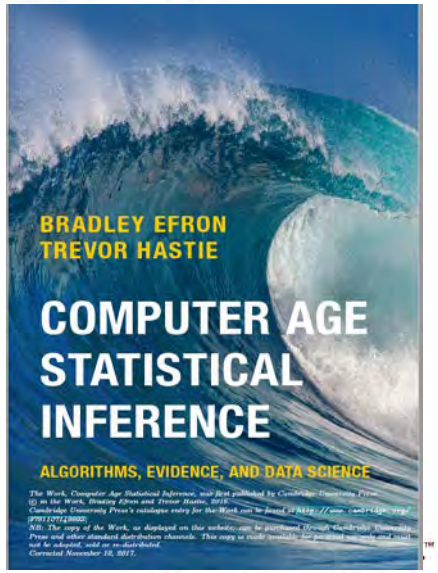
Difficulty modeling

Approach: Gradient Boosting Machines (GBM)

- Long history of relying on linear regression
- Here we use GBM







Gradient Boosting Machines

Revisions to improve difficulty prediction

- Additional and improved features
- Version 2
  - Predictors
    - Word level (familiarity)
    - Inter-word level (depth of word familiarity)
    - Variable importance
  - R-squared LOO
    - GBM (gradient boosting machine)
  - Cross validation
    - Word level
    - Inter-word level
    - Variable importance
- Version 2 + SME

Difficulty modeling (r-squared predicted and estimated)

Version	Obtained from WK items	Cross validated on 5-option WK items	Applied to generated items	Obtained from field tested items	Version 2 applied to WK
1 (pre-field test)	0.26	0.34	0.25		
2 (post field test)				0.34	0.35
N (items)	183	77	337	252 (subset of 337, a > 0.80)	76 (subset of 183 a > 0.80)

Version 1: Field test difficulty model based on WK operational items

Version 2: Final test difficulty model (based on generated field-tested items)

Importance of difficulty predictors, based on field test data

Most important features for predicting difficulty (all predictors)		Relative importance
TASA_adjustedSFI_Target	(Corpus-based familiarity)	13.207750926
PGLB_Target	(K-12 student-based tested familiarity)	6.518500716
Cosine_W2V_GN_Target_Key	(Corpus-based depth of familiarity)	6.072907650
Cosine_Glove6B300D_Target_Key	(Corpus-based depth of familiarity)	5.675464925
PMI_Target_Key	(Corpus-based depth of familiarity)	3.686734148
LnFreqCriterion_Target	(K-12 student-based familiarity expressed in writing)	3.002441342
Cosine_WordFit_Target_Key	(Corpus-based (Depth of familiarity)	2.850595830
LWVGL_Target	(K-12 student-based tested familiarity)	2.629443102
njPMI_Target_Key	(Depth of familiarity)	2.404460026
SLL_Target_Key	(Depth of familiarity)	2.375988728
Cosine_Glove6B300D_Target_D_Min	(Depth of familiarity)	2.114946966
Wiki12_adjustedSFI_Target	(Familiarity)	2.110742934
LnFreqCriterion_	(K-12 student-based familiarity expressed in writing)	1.678484246
CP_Target_Key	(Corpus-based depth of familiarity)	1.383976236
Cosine_W2V_GN_Target_D_Min	(Corpus-based depth of familiarity)	1.311576391
Cosine_Glove6B300D_Target_D_Max	(Corpus-based depth of familiarity)	1.247878802
NumberBaseFormWordFamilyMembers_Key		1.113784596

Contrast between items generated and produced by SME

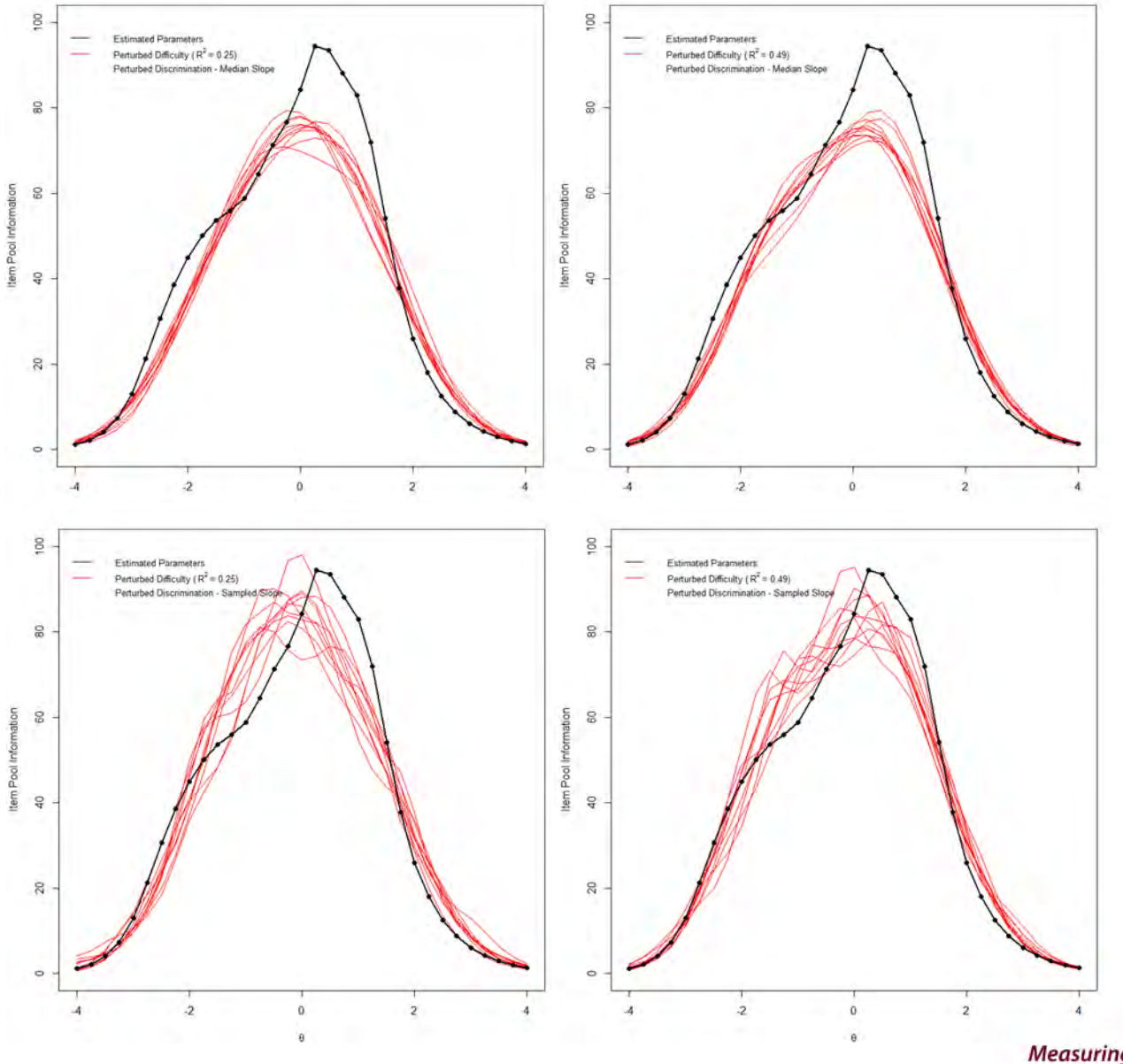
Version	Obtained from WK items	Cross validated on 5-option WK items	Applied to generated items	Obtained from field tested items	Version 2 applied to WK
Familiarity				0.23	0.39
Depth of familiarity				0.22	0.06
N (items)	183	77	337	252 (subset of 337, *a* > 0.80)	76 (subset of 183 *a* > 0.80)

SME contribution to difficulty prediction

Version	Obtained from WK items	Cross validated on 5-option WK items	Applied to generated items	Obtained from field tested items	Version 2 applied to WK	Obtained from field tested items + 2 SMEs
1 (pre-field test)	0.26	0.34	0.25			
2 (post field test)				0.34	0.35	0.50
N (items)	183	77	337	252 (subset of 337, *a* > 0.80)	76 (subset of 183 *a* > 0.80)	252 (subset of 337, *a* > 0.80)

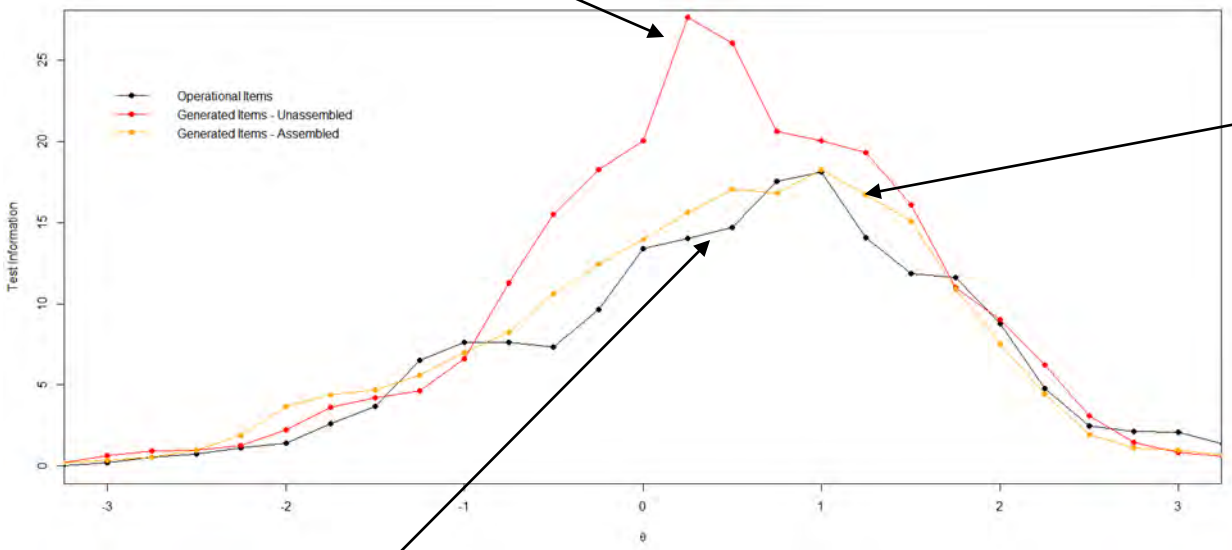
CAT simulations

Evaluation: A CAT simulation



$LI(\theta, \hat{\theta}|a, b, c)_1$, the item pool consists of 339

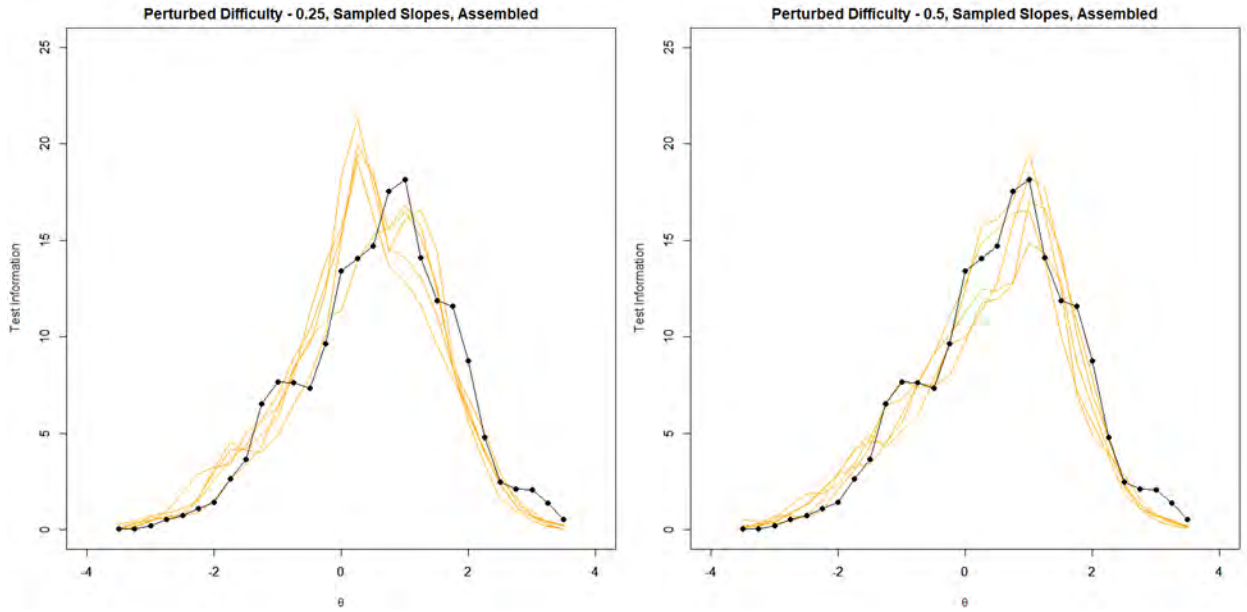
$LI(\theta, \hat{\theta}|a, b. c))_3$, the item pool consists of 75 field tested items, assembled from the 339 field tested items, to match $LI(\theta, \hat{\theta})_2$



$LI(\theta, \hat{\theta}|a, b, c)_2$, the item pool consists of the 75 items in WK 7 calibrated

CAT with degraded parameter estimates and item pool of 75 items

Current r-squared: 0.34, 0.50 with two SMEs



Information functions $LI(\theta,\hat{\theta})_2$ compared to multiple $LI(\theta,\hat{\theta})_3$ with degraded difficult estimates and sampling of discrimination: $LI(\theta,\hat{\theta})_{3\,(a^*,b^*,c^*,r^2_{.25})}$ (left panel), and $LI(\theta,\hat{\theta})_{3\,(a^*,b^*,c^*,r^2_{.50})}$ (right panel).

Conclusions regarding CAT simulations

- For estimating ability, an approach to compensate imprecise parameter estimates is to lengthen the pool (and not the test!)
- When assembling 75 item pools from the degraded 339 field tested items, we did not take the imprecise parameters into account, although there is methodology for that purpose
- For WK, it seems plausible to avoid or greatly reduce pretesting
- Caveat: Need to avoid or detect items with negative biserials!

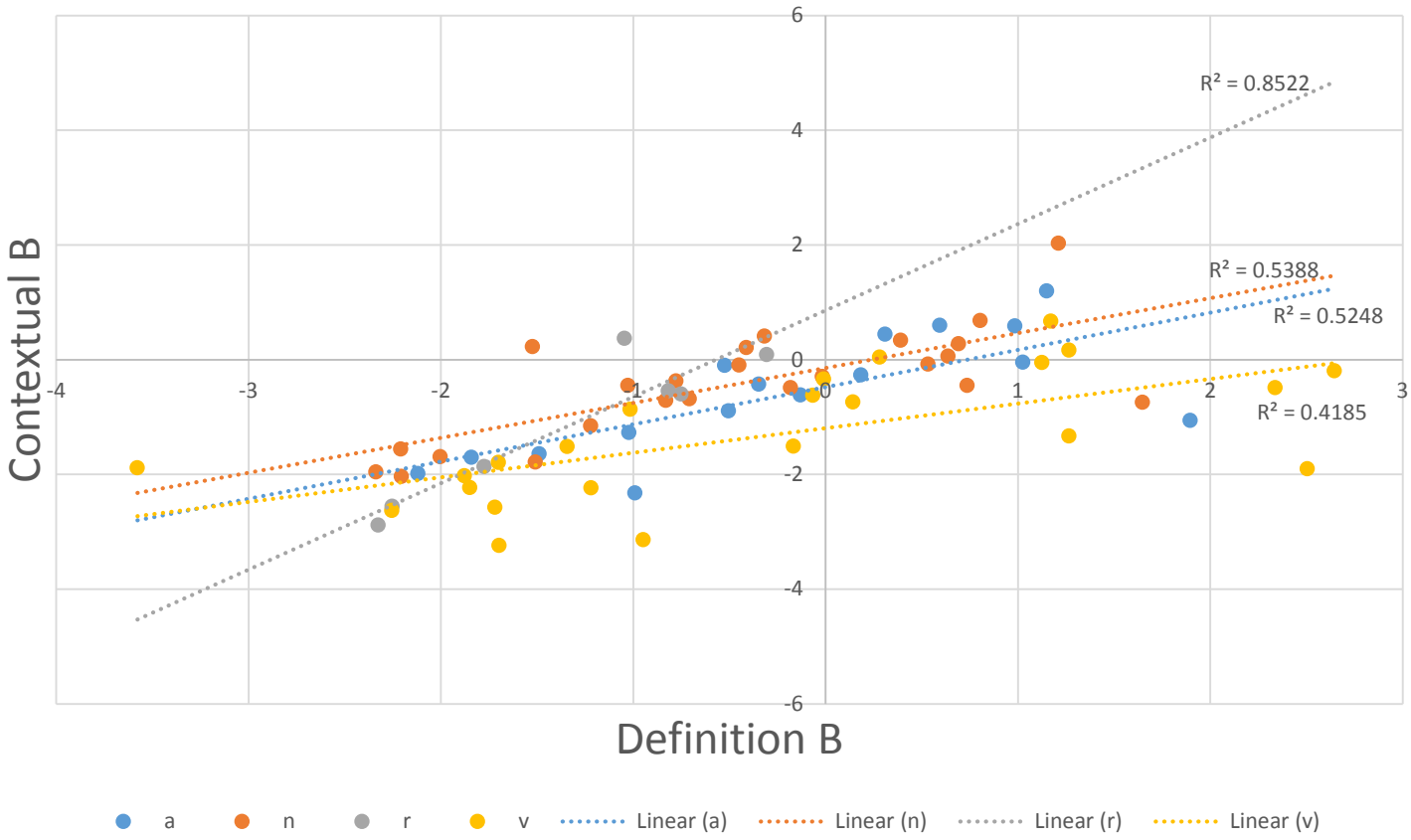
Veldkamp, B. P., Matteucci, M., & de Jong, M. G. (2013). Uncertainties in the item parameter estimates and robust automated test assembly. *Applied Psychological Measurement, 37(2), 123-139. doi:10.1177/0146621612469825*

Contextual items

Status of contextual generator

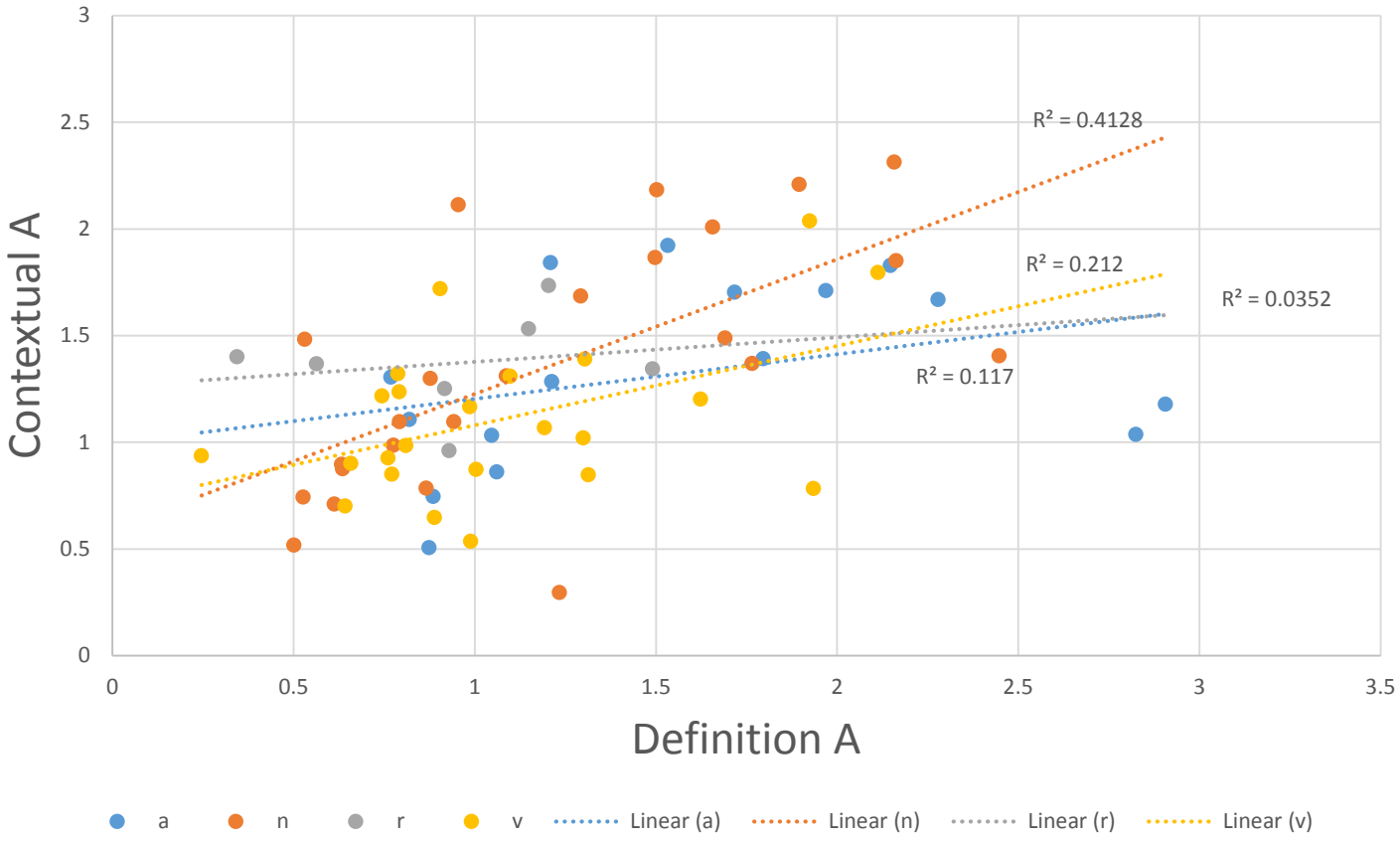
- Has been on hold to finish the other system improvements
- We are taking a approach:
  - The stem, key and distractors are provided by a definitional item
  - Templates inspired by successful WK 7 and WK 3 vocabulary items were created
  - We first place the stem in the template
  - We then fill the template by fitting n-grams
    - (For the field test SMEs filled the templates)
  - Challenge of NLG: Although the resulting sentence are grammatically correct, often they do not make sense
  - Yield expected to be low but the items will likely perform well

Relationship of difficulty estimates for definitional and corresponding contextual items



Correlations	
All items	0.63
A Only	0.72
N Only	0.73
R Only	0.92
V Only	0.65

Relationship of discrimination estimates for definitional and corresponding contextual items



Correlations	
All items	0.48
A Only	0.34
N Only	0.64
R Only	0.19
V Only	0.46

N-gram analysis of WK contextual items

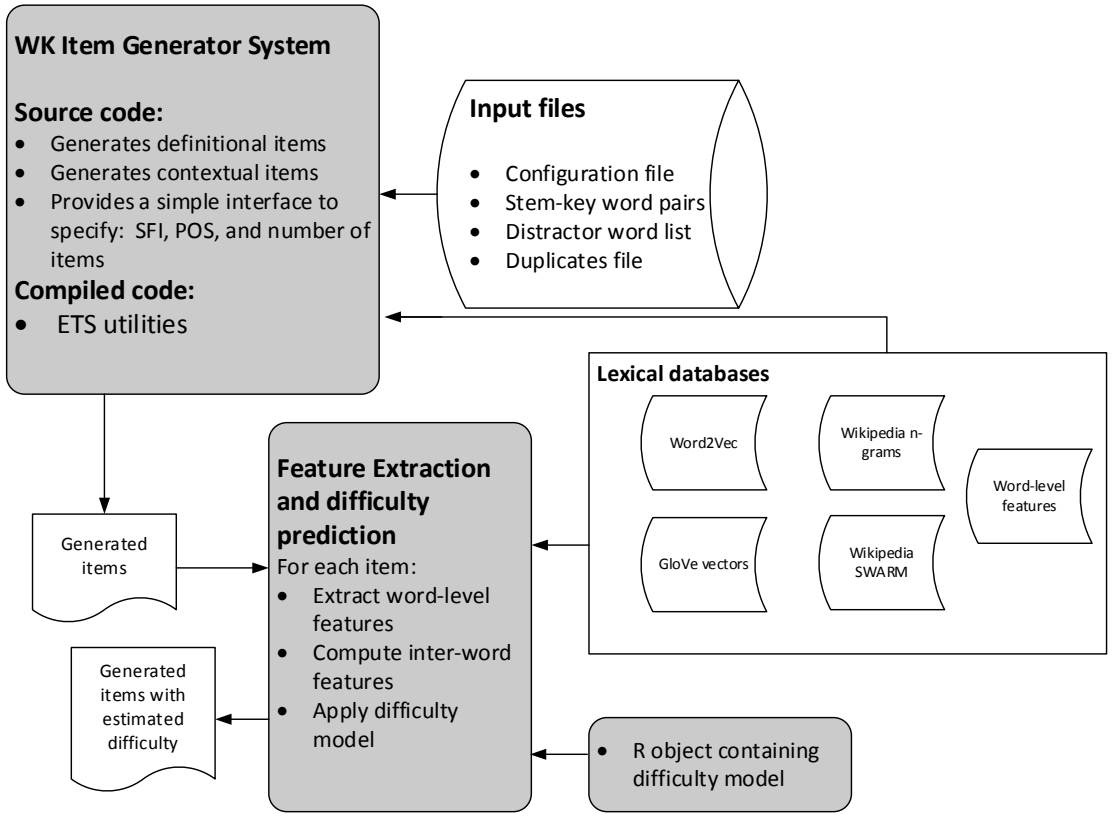
Number of PhraseFinder Hits by ngrams												
id	**n=2**	**n=3**	**n=4**	**n=5**	**n=6**	**n=7**	**n=8**	**n=9**	**n=10**	**n=11**	**word_count**	**max_n_match**
id55	125,918,547	3,446,322	714,348	2,342	0	0	0	0			9	5
id144	3,575,410,111	106,327,169	10,381,981	980,904	0	0	0	0	0	0	11	5
id154	1,432,944	0	0	0	0						6	2
id169	80,483,193	320,369	37,495	160	0						6	5
id212	158,137,985	2,381,763	47,293	582	0	0	0	0	0	0	11	5
id58	73,293,033	5,379,755	184,201	0	0	0	0	0			9	4
id48	42,146,092	65,474	57	0							5	4
id24	435,987,947	24,625,971	53,125	2,116	0	0					7	5
id5	11,607,074	679,524	1,393								4	4
id18	54,920,696	1,865,074	4,901	216	0	0	0				8	5
id143	15,607,380	9,797	0	0	0	0					7	3
id39	222,015	11,463	0								4	3
id164	97,756,658	206,665	8,220	0	0	0					7	4
id176	20,414,124	12,746	550	0	0	0					7	4
id177	613,163,797	449,684	84,112	0	0	0	0				8	4
id41	6,535,203	167,781	215	0	0						6	4
id28	5,876,377	0	0	0	0						6	2
id30	1,590,813,452	1,533,054	57,525	42	0	0	0				8	5
id8	318,182,572	6,978,588	119,901	491	0	0	0				8	5
id57	4,093,363	333	0								4	3

Deliverables

Deliverables

- A final report: Development of the ASVAB Word Knowledge Automated Item Generation System
- A research report
- The item generation system as depicted in Figure 1.
- 349 generated field tested items
- 253 additional items accepted by SME but not field tested
- Templates for the generation of contextual items
- 3,000 generated definitional items produced by the post field test system
- Excel interface to elicit SME estimate of difficulty

WK Item Generation System



Next steps

- System has configuration file with several “dials”
- Experimentation and evaluation of system by OPA

Q&A

Tab P

OPA

OFFICE OF PEOPLE ANALYTICS

ASVAB Career Exploration Program

September 21, 2018

ASVAB CEP Numbers and Metrics

Year**	Number of Students Tested
2012	672,311
2013	670,836
2014	690,950
2015	687,900
2016	706,200
2017	684,223 (includes CEP-iCAT)
2018	**713,777** (includes CEP-iCAT)

Year**	Number of Schools Tested	Percentage of Schools Tested
2012	12,540	56.2%
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%**

**School year runs from July 1- June 30. Final numbers (as of June 30, 2018).

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Year-to-Date*

Paper and Pencil Numbers

	Examinees 16-17	Examinees 17-18
TOTAL	670,886	**662,564**

CEP iCAT Numbers

	Examinees 16-17	Examinees 17-18
TOTAL	14,011	**51,213**

*Total students as of 30 June, 2018.

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

Year	ARMY	NAVY	AIR FORCE	MARINE CORPS	COAST GUARD	TOTAL
2014	14,579	4,444	3,588	5,272	232	28,115
2015	15,281	4,964	3,822	5,793	302	29,860
2016	14,805	4,106	4,718	5,659	340	29,628
2017	14,496	4,922	4,371	5,848	376	30,013
2018	**14,384**	**4,647**	**4,219**	**5,362**	**405**	**29,017**

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Leads and Options

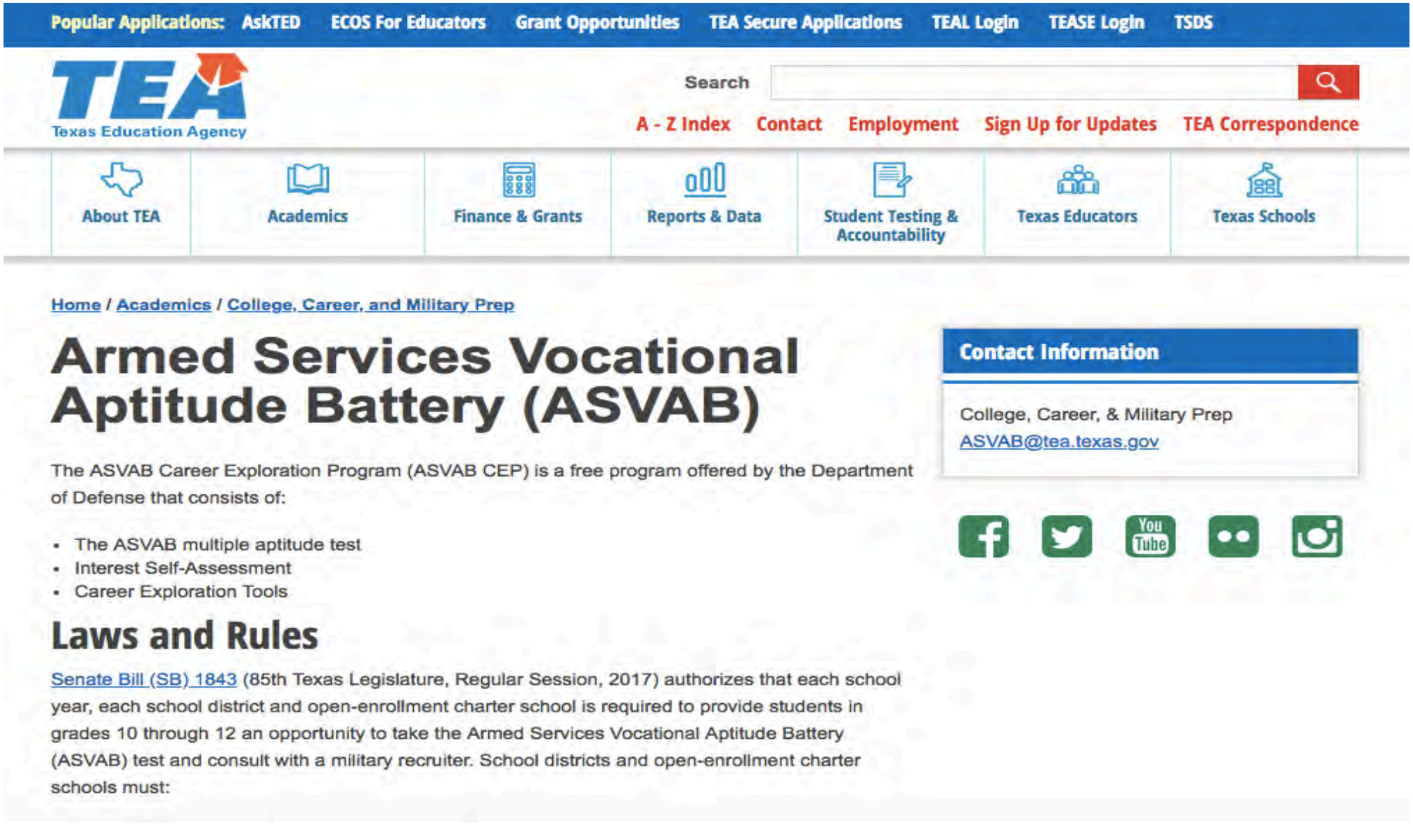
Year**	Leads Provided to Military Services
2014	492,419
2015	470,229
2016	478,196
2017	440,542
2018	**433,317**

Option	Results to Recruiting Services
1	7 days after test scores are mailed
2	60 days after test scores are mailed. No contact prior to that time.
3	90 days after test scores are mailed. No contact prior to that time.
4	120 days after test scores are mailed. No contact prior to that time.
5	End of school year. No contact prior to that time.
6	7 days after test scores are mailed. No telephone solicitations by recruiters.
7	Administrative option used by USMEPCOM ONLY for test administration issues for individual or group tests (test abandoned, cheating, insufficient proctors, fire drills, etc.). Not valid for enlistment purposes. Results not released to Recruiting Services.
8	Not released to Recruiting Services *While student results are not released to military services, scores are valid for enlistment for two years after test date.

**School year runs from July 1- June 30. Preliminary numbers (as of June 30, 2018).

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM



OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Impact of TX Legislation on testing numbers and options chosen by schools

MEPS SY 16-17	1	2	3	4	5	6	7	8
Amarillo	2,046	644	269	61	173	678	11	1,166
Dallas	7,977	1,923	675	77	572	1,459	122	5,747
El Paso	2,381	5	0	28	1089	288	16	1,498
Houston	6,800	2,384	993	220	1259	604	316	4,165
San Antonio	8,202	1,495	160	0	472	951	107	3,620
9th BN	27,406	6,451	2,097	386	3565	3,980	572	16,196

MEPS SY 17-18	1	2	3	4	5	6	7	8
Amarillo	1,858	520	297	125	48	810	18	1,401
Dallas	9,385	1,538	256	273	2295	1,319	168	8,523
El Paso	2,759	26	850	23	853	173	42	2,179
Houston	5,659	1,268	991	216	1762	1,109	331	6,123
San Antonio	7,686	2,104	102	0	193	1,261	602	4,229
9th BN	27,347	5,456	2,496	637	5151	4,672	1,161	22,455

Does Not Include 10th Grade Students: 60,653

Does Not Include 10th Grade Students: 69,375

**School year runs from July 1- June 30. Final numbers (as of June 30, 2018).

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

ESSA Engagement

ENGAGED STATES	ENGAGED EASTERN MEPS	ENGAGED WESTERN MEPS
Arizona	Albany	Amarillo
Arkansas	Atlanta	Dallas
California	Baltimore	Denver
Colorado	Beckley	El Paso
Georgia	Buffalo	Houston
Kansas	Ft. Dix	Kansas City
Indiana	Ft. Jackson	Little Rock
Maryland	Indianapolis	Los Angeles
Missouri	Jackson	Phoenix
New Jersey	Jacksonville	Sacramento
New York	Knoxville	Salt Lake City
Nevada	Louisville	Phoenix
Oregon	Memphis	Portland, OR
San Juan	Nashville	San Antonio
Tennessee	New York	San Diego
Texas	Pittsburgh	San Jose
Utah	San Juan	Seattle
Vermont	Springfield	Shreveport
Washington	Syracuse	Spokane
West Virginia		St. Louis

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Program Initiatives

- Expert Panel
- Needs Assessment
- Test Security
- CEP iCAT updates
- ESSA and ASVAB CEP engagement
- Website analytics and updates
- New and updated program materials
- Teacher engagement campaign
- Marketing efforts
- Automated processes
- UniFORM

Expert Panel

- FYI analyses
  - Analyses of factors and items
  - Gender differences still present in population
- States and current legislation
- Training for ESS Community
  - NCDA certification inclusion for ESS community

Needs Assessment

- MEPS visits
- School observations of testing and posttest workshops
- Business practices, efficiencies, and model of program delivery with milestones

Test Security: Caveon

- Monitor websites and discussion groups for test compromise
- Report material in question for remediation

CEP iCAT Updates

- Updated functionality to include eliminating case sensitivity for failure recoveries
- Made modifications requested by field: DOB sequence, references to applicant/candidate replaced with student, acronyms spelled out, etc
- 13” monitors allowed for testing
- QA process for scores

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

	16-17	17-18
Unique Visitors	332,408	440,882
Returning Visitors	120,277	203,357
Page Views	5,420,195	6,747,160
Bounce Rate	25.45%	31.41%
Average Time Per Session	13:46	12:55
Number of Pages Per Session	11.98	10.49
Tablet/Mobile Visitors	136,968	212,870

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

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

	16-17	17-18
Unique Visitors	166,638	72,230*
Returning Visitors	16,016	39,515
Page Views	442,215	2,003,165
Bounce Rate	70.63%	31.90%
Average Time Per Session	1:41	4:34
Number of Pages Per Session	2.30	17.93
Tablet/Mobile Visitors	72,097	26,330*

Visitors who land on the site, return 55% of the time!

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

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Access Code Utilization
(July 1, 2017 - June 30, 2018)

Code Type	Visitors	Repeat Visitors
Marketing	3,983	1,137
Counselor	884	349
Student	226,998	53,790
Reserve	1,629	558
All	233,494	55,834
Total Number of Logins	353,317	56,046 came back more than twice

Website Utilization and Access Code Utilization show great ROI for conferences and marketing efforts.

Both of these metrics indicate an opportunity for training in the field regarding PTIs.

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Contact Us: www.asvabprogram.com
(July 1, 2017 - June 30, 2018)

- **Inquiries: 1,080**
- **Bring ASVAB CEP to Your School: 519**
  - **Student or Parent: 211**
  - **Counselor: 308**
- **Score Requests: 1,415**

Total: 3,533

This represents the amount of work that comes through the website requiring manual labor from DPAC and USMEPCOM.

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Contact Us: www.careersinthemilitary.com
(July 1, 2017 - June 30, 2018)

- **Army: 78**
  - **Army National Guard: 5**
- **Marine Corps: 2**
- **Navy: 89**
- **Air Force: 124**
  - **Air National Guard: 2**
- **Coast Guard: 53**

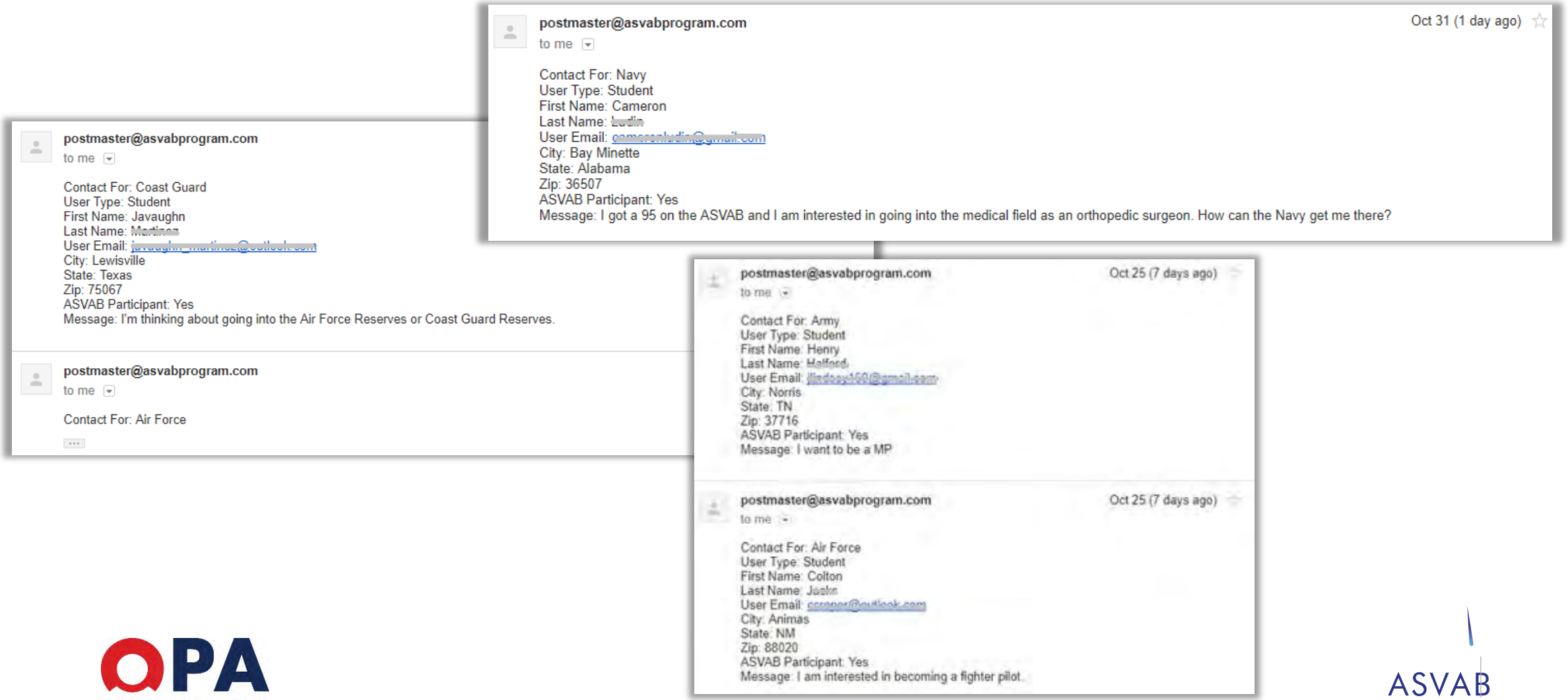
Total: 353

This represents the number of students reaching out to the Military services via Careers in the Military. These inquiries also require man-hours from the services.

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Sample Inquiries: www.CareersintheMilitary.com



postmaster@asvabprogram.com Oct 31 (1 day ago)
to me

Contact For: Navy
User Type: Student
First Name: Cameron
Last Name: [redacted]
User Email: [redacted]
City: Bay Minette
State: Alabama
Zip: 36507
ASVAB Participant: Yes
Message: I got a 95 on the ASVAB and I am interested in going into the medical field as an orthopedic surgeon. How can the Navy get me there?

postmaster@asvabprogram.com
to me

Contact For: Coast Guard
User Type: Student
First Name: Javaughn
Last Name: [redacted]
User Email: [redacted]
City: Lewisville
State: Texas
Zip: 75067
ASVAB Participant: Yes
Message: I'm thinking about going into the Air Force Reserves or Coast Guard Reserves.

postmaster@asvabprogram.com
to me

Contact For: Air Force

postmaster@asvabprogram.com Oct 25 (7 days ago)
to me

Contact For: Army
User Type: Student
First Name: Henry
Last Name: [redacted]
User Email: [redacted]
City: Norris
State: TN
Zip: 37716
ASVAB Participant: Yes
Message: I want to be a MP

postmaster@asvabprogram.com Oct 25 (7 days ago)
to me

Contact For: Air Force
User Type: Student
First Name: Colton
Last Name: [redacted]
User Email: [redacted]
City: Animas
State: NM
Zip: 88020
ASVAB Participant: Yes
Message: I am interested in becoming a fighter pilot.

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

New Features!

www.ASVABprogram.com

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Institution Details

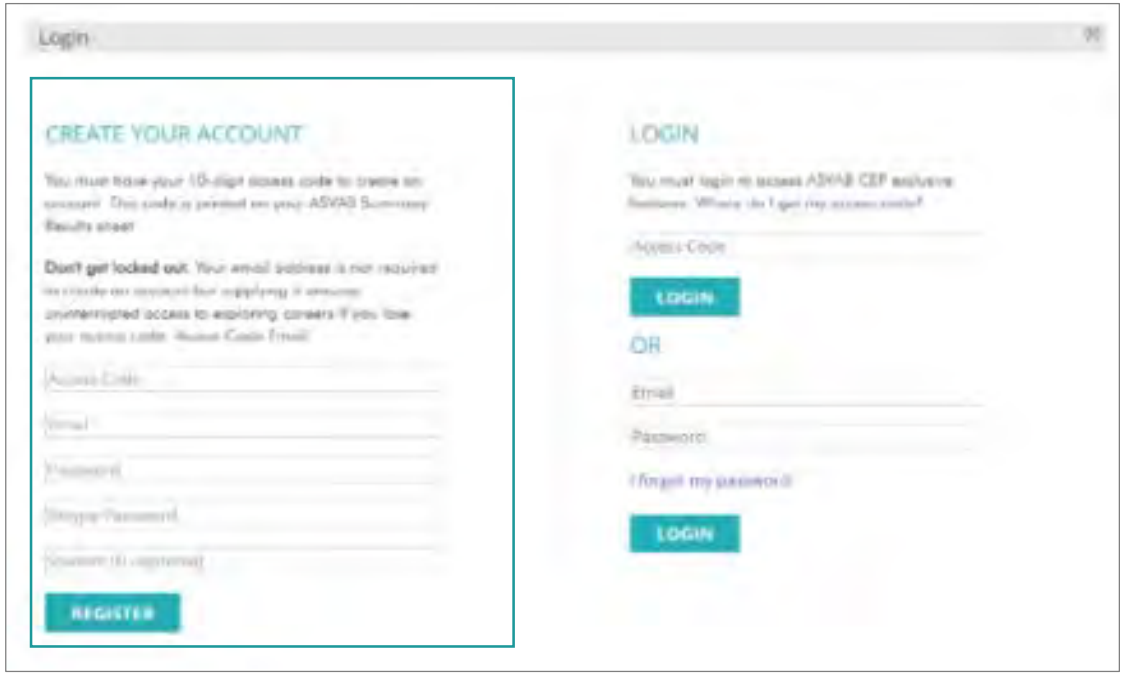
- Ability to make notes and save colleges
- College details include acceptance rates, retention rates, average test scores, and ROTC programs offered by service

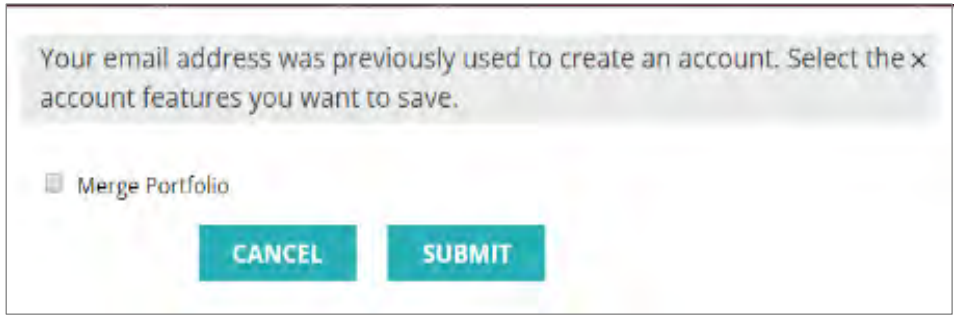
OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Merge Accounts:

- Portfolio
- FYI Resuults
- Saved Occupations and Notes
- Favorites

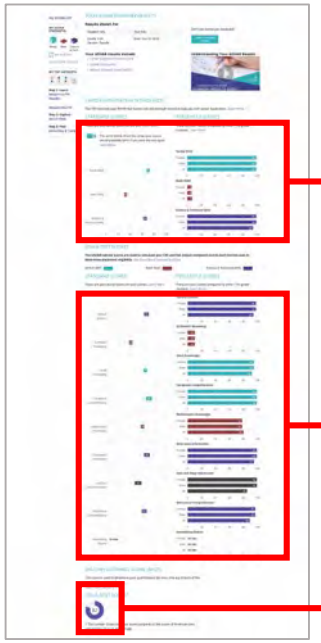


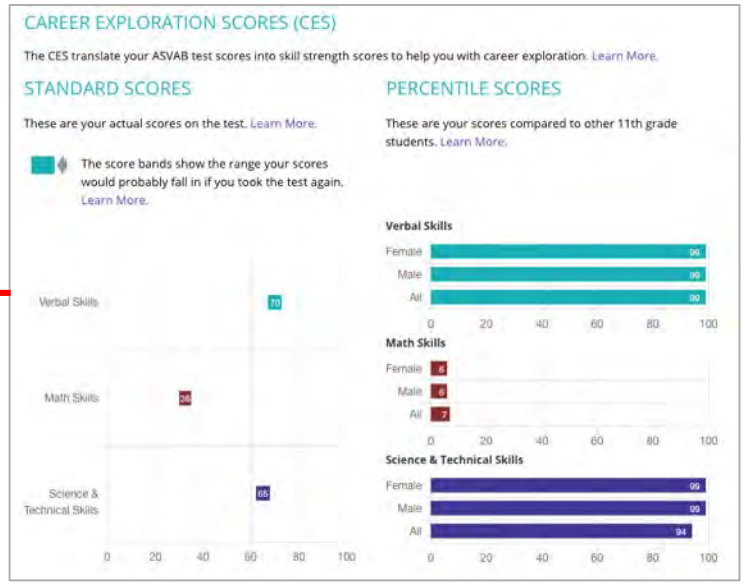


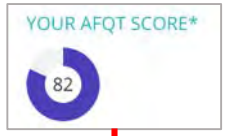
OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

ASVAB Summary Results







OPA
OFFICE OF PEOPLE ANALYTICS

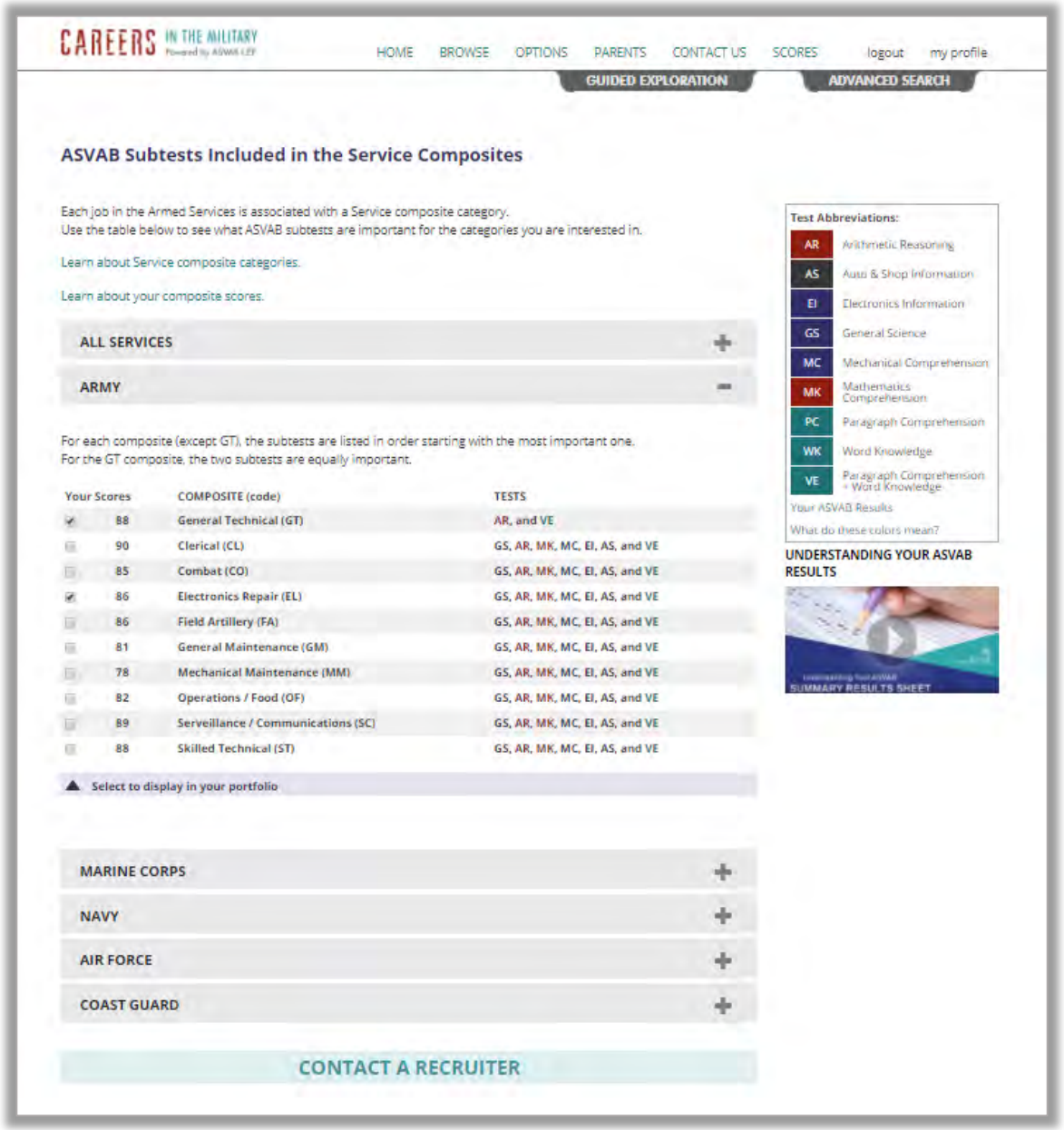
Service Line Scores

- Included so students can have meaningful conversations with recruiters

-Students are able to see that they are employable

-For schools who choose option 8, students have the power to reach out to recruiters and discuss options easily, and in the context of their line scores

-Does not require a pull, so does not count against student as the first test

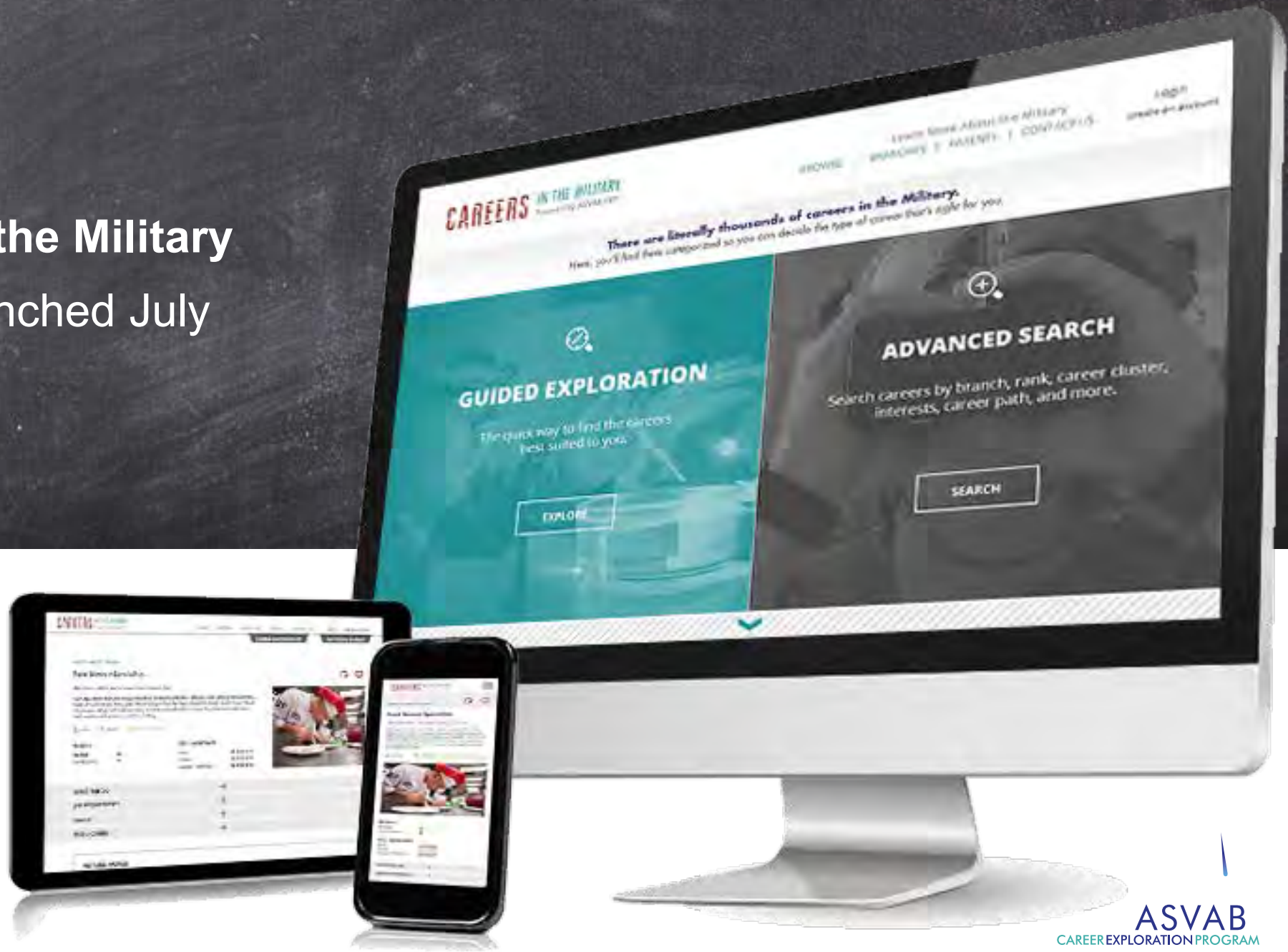


OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Careers in the Military

Update Launched July 2018



OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Navigation Stats (Page views)

- Composite Scores: 304,713
- Parents: 527
- Contact Us: 365
- Options: 8,956
  - Military: 4,004
  - Enlistment requirements: 3,842
  - Types of Service: 3,548
  - Boot camp by service: 3,313
  - Reasons to consider: 3,341
  - Becoming an officer: 3,147
  - Enlisted vs officer: 5,735
  - Enlistment process: 2,239
  - Pay: 2,429
  - Benefits: 2,293
  - Contact a Recruiter: 1,458

Analytics Evaluation

SEARCH

- Guided Exploration 38,635
- Advanced Search 38,127
- 2,504 users sorted by hot jobs
- Advanced Search Sorted by Service
  - Army: 2,163
  - Navy: 1,609
  - Air Force: 1,746
  - Marine Corps: 1,728
  - Coast Guard: 699

REFERRALS & SOCIAL

- Referrals from CEP: 25,851
- Top five shared URLs:
  - 205 Homepage
  - 10 Air Traffic Controllers
  - 8 Air Force 1C131 (Air Traffic Controllers)
  - 6 Advanced Search
  - 5 Accountants and Auditors
  - 5 Marine 3531 Motor Vehicle Operators

Service Representatives… Thank you!

Still needed:
-Military Hot Jobs
-Scores
-Career Paths

Data Point	Army	Marine corps	Navy	Air Force	Coast Guard	Army National Guard
MOS/AFSC/NEC	X	X	X	X	X	X
Career Field/Title	X	X	X	X	X	
Level (Officer/Enlisted)	\	\	\	X	\	
URL to Career Details on Service Site	X		X	X	X	X
Description	X	X	X	X	X	
Tasks		\		X	X	
ASVAB Requirements (Min for entry; job-specific ASVAB composite)	\	\	\	\		\
Physical Demands			X	\		
Basic Training (length)			X	X		
Technical Training (length)			X	X	X	
Technical School Location		X	X	X	X	
Desired High School Courses			\	X	X	
Prerequisites		X				
Videos/Profiles/Images			X	X	X	
Related Jobs/Career Path					X	
Related Civilian Careers		X			X	
Active Duty/Reserve	X				X	
Hot Jobs (critical fill/in demand)						
POC for Analytics						
POC for Contact Us Button	X	X	X	X	X	x

X = data provided for all occupations

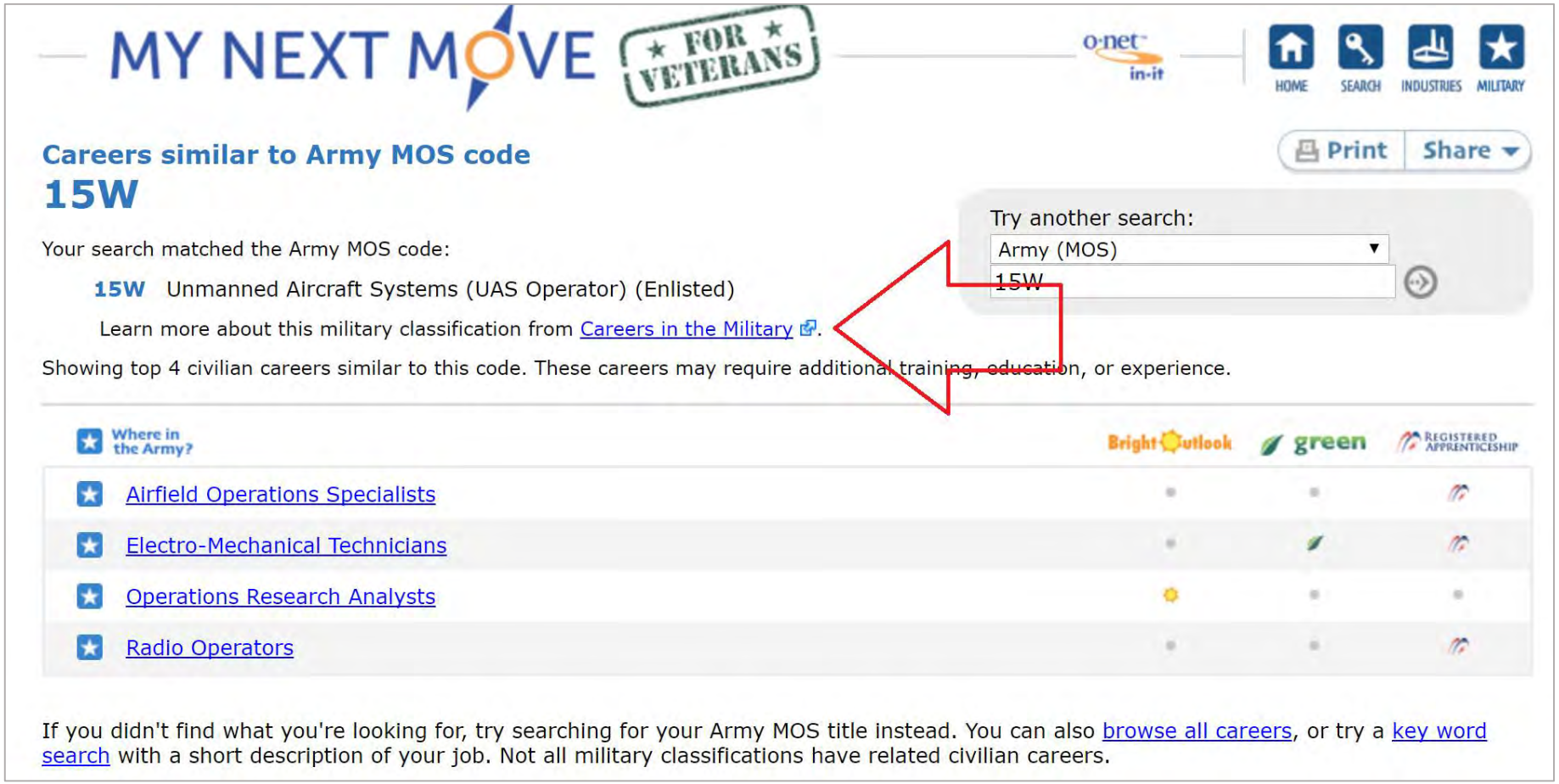
\ = partial data provided for some occupations (not all)

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

New Referrals

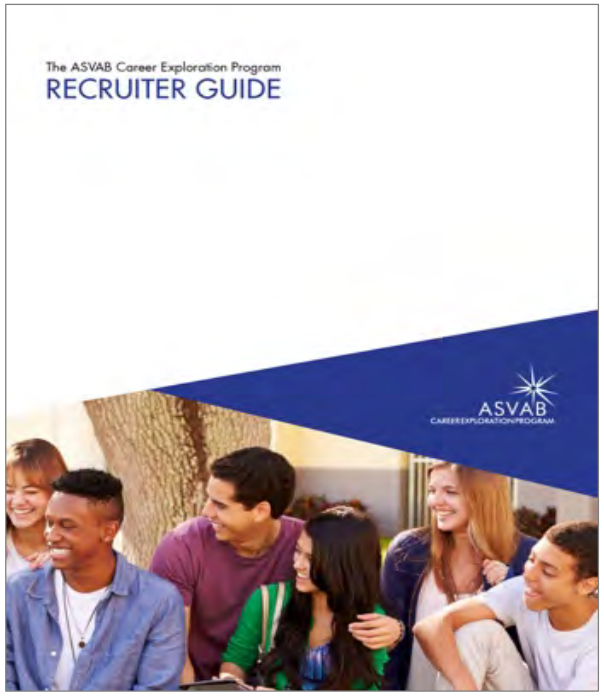
2,044 sessions

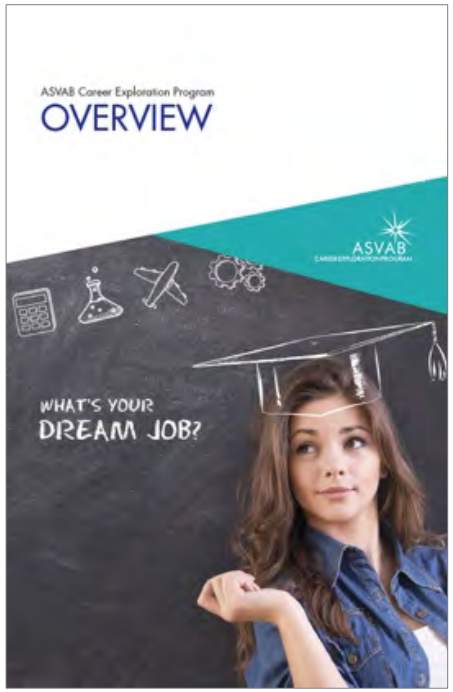


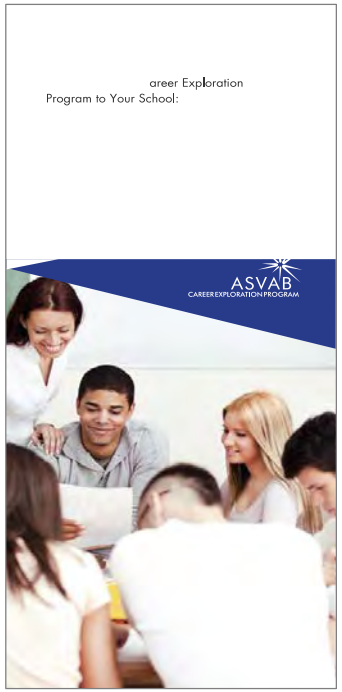
OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Communication Efforts: Program Materials







OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Communication Efforts: Teacher Engagement Campaign

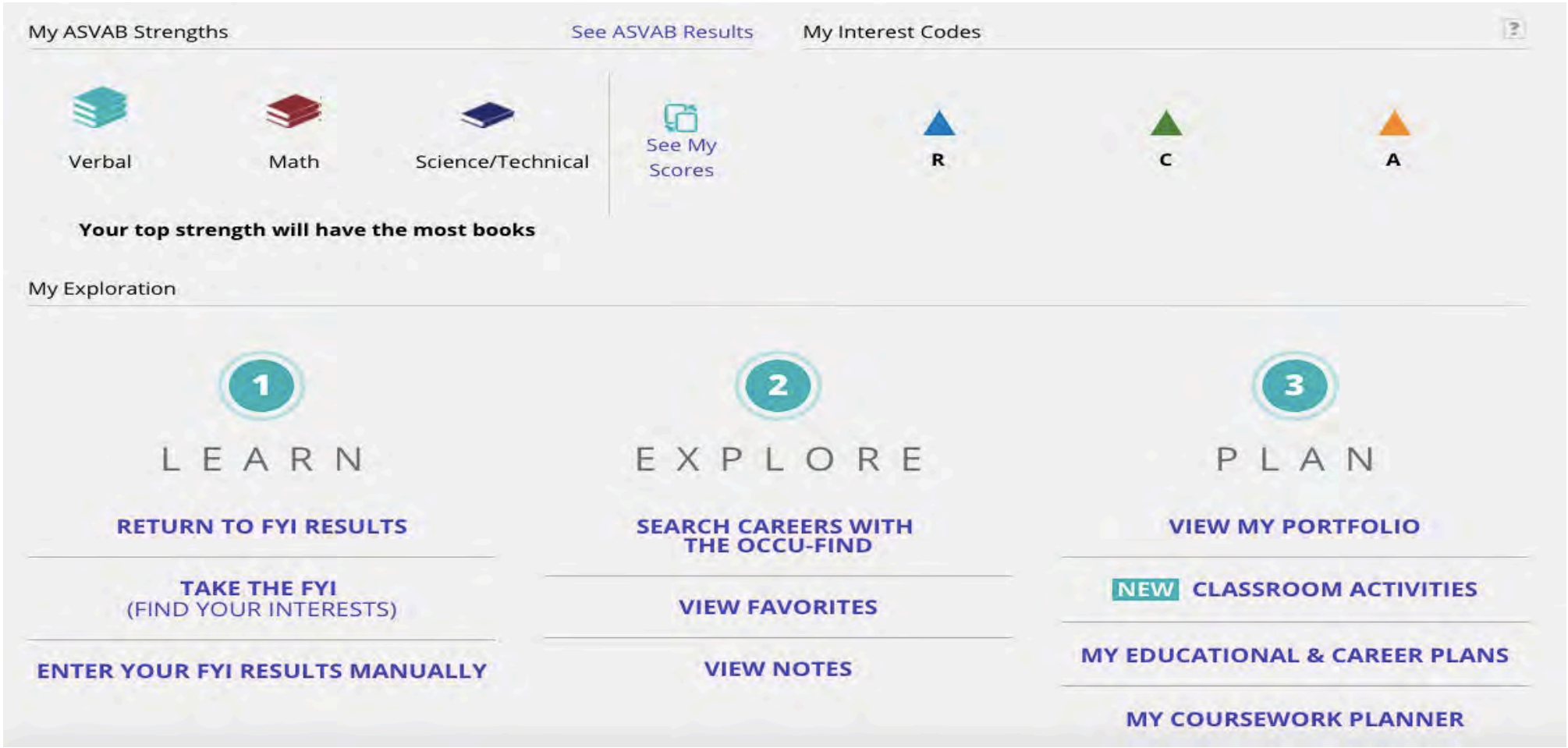
Audience	Totals	Time	Performance # of Accounts Created
CTE, Tech Ed	60k, email	February 2018	CEP123: 626
English, Social Science, Business	230k, post card mail	March 2018	CEP456: 110
Math	95k, email	May 2018	CEP789: 76
Science	85k, email	May 2018	CEP987: 41
School Counselors	84k, email	May 2018	CEP654: 62
Back to School	All of the above and JROTC, Principals, Parents	August 2018	

Per ASCA recommendation
Student to Counselor Ratio should be
250:1
See by State ratio in back-up slides

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Career Exploration Opportunities



OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

CLASSROOM ACTIVITIES

Integrate career exploration into any curriculum. Connect your classroom to the real world.

Math

- Cost of a Career
- Math Skills Analysis
- Share to Inspire

Science

- Scientific Leaps and Bounds
- Which STEM Field is for Me?
- Share to Inspire

Career Tech Ed (CTE)

- Listen Before You Launch
- CTE Reel Review
- Share to Inspire

Social Studies

- What do real jobs look like?
- Interests and Needs
- Giving Back

Activity

COST OF A CAREER

Choose a career using the OCCU-Find. Identify two paths to outline how you could reach this career goal. Make a list of costs associated with each path. Include cost of living, location, the amount of time that it will take to reach your goal, plus any other factors that will affect the cost. Create a comparison chart to show the total amount of cost associated with each path. *Tip: Use the living wage calculator to develop your chart: http://livingwage.mit.edu/

PRINT ACTIVITY CLOSE

Communication Efforts: National Events

Marketing Events	Education/Research Industry	Stakeholder Engagement

• **National Career Pathways Network, October 25-27**
• **Association for Career and Technical Education, December 6-9**
• **American Counseling Association, April 26-29**
  • **Booth** 116 leads
• **National Alliance for Public Charter Schools, June 11-14**
  • **Booth** 121 leads
• **National Career Development Association, June 21-23**
  • **Booth** 47 leads
  • **Presentation:** The Integrative Approaches to Augmenting Vocational Interests in Career Exploration
• **National Parent Teacher Association, June 21-24**
• **American School Counselors Association, July 8-11**
  • **Booth** 542 leads (the most ever!)
	• **Council of Chief State School Officers, Nov 17**
	  • **Presentation (webinar):** The ASVAB Career Exploration Program fulfills readiness and preparedness requirements in states' ESSA plans for accountability
	• **Texas Education Agency, Dec 13**
	  • **Presentation:** Texas Senate Bill 1843 Armed Services Vocational Aptitude Battery (ASVAB) Overview and Resources
	• **RCOE School Counselor Leadership Networking Conference, Feb 6-8**
	  • **Presentation:** Importance of Exploring all the Options
	• **CTE 2.0 Shaping Our Future – CACTE, March 4-6**
	  • **Presentation:** Debunking the myths surrounding the ASVAB and showing the new tools available through ASVAB CEP
	• **Society for Vocational Psychology, June 18-20**
	  • **Presentation:** The development of the Find Your Interests inventory
	• **National Conference on Student Assessment, June 27-29**
	  • **Presentation:** Implications of ESSA on Career Readiness: Expanded Opportunities for Improving Student Achievement Outcomes
		• **New ESS Training, August**
		• **Seminar on utilizing resources effectively, Feb 26-March 1**
		  • **Presentation**: ESS training & TEA Legislation Impact
		• **OPA Research Forum, March**
		  • **Presentation**: Are All Careers Created Equal? A Comprehensive Approach to Military and Civilian Career Exploration
		• **Needs Assessment, March - June**
		• **US Army National Educator Tour, May**
		  • **Presentation**: The ASVAB Career Exploration Program: It's not just a military test
		• **JAMINAR, May**
		• **USMC Recruiter School House, June**
		• **ESS Chalk Talk, July**
		  • **Presentation**: Annual Update
		• **Army ESS Training, August**
		• **JRCC, August**

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

Communication Efforts: ASVAB CEP Marketing

- Potential Magazine (Alabama students and parents)
  - Dedicated email 10/26
  - Military Guide; editorial and ad (Winter)
  - College and Career Guide; activity spread (Summer)
- Indiana School Counselor Association (banner at www.indianaschoolcounselor.org)
- New Jersey School Counselor Association (retargeting for users leaving www.njsca.org)
- Association for Career And Technical Education (banner at www.acteonline.org)

Communication Efforts: MEPS Social Media

Goal: Manage local Facebook Group for each of 65 MEPS

Purpose: Share local, region-specific information regarding ASVAB CEP within the appropriate communities

- Improve reputation
- Build awareness
- Increase participation

Stakeholder Input: 36/65 MEPS POCs have actively participated within their groups

Action Steps: Launch monthly ASVAB CEP Social Media Toolkit containing social media tips and tricks, best practices and sample posts. Generate content to share through local media mentions, event coverage, and test administration. Continue school/guidance office channel following.

Program Initiative: Automating Processes

Test Score Look Up

- Allows DPAC personnel to instantaneously generate a score report using minimal information. The brand new interface is optimized to allow the authorized user to run a dynamic query using as many or as few filters as they have and print a score report identical to the one generated by US MEPCOM. An additional function of this application allows us to create customizable reports about demographics. Upon request, we can now provide states and schools with reports about their populations' ASVAB Test scores and FYI results data.

Access Codes Generation

- Allows an authorized user to select the number of access codes he or she would like to create, assign the number of times the new access code(s) will be able to take the FYI, designate a prefix (streamlining lead tracking), and whether the new code(s) should include sample test scores. The application then generates a report for distribution to field.

Program Initiative: Military to Civilian Crosswalk

Goal: To maximize career exploration, the ASVAB CEP cross-links military and civilian occupations to expose young people to all of their options. The linkages on CEP and CITM are currently based on DMDC's crosswalk, and supplemented with additional information provided by Department of Labor (VOW study)

Purpose: update links using a task analysis approach

- Predictive analytical software to identify potential matches
- Two-rater teams make evaluations and reach consensus on strength of match

Stakeholder Input: OPA is collaborating with other key stakeholders involved in similar efforts, including military COOL, Department of Labor (VOW), and Transitioning Veterans Program (TVP)

- Past meetings: 17 NOV, 7 APR
- Next in person meeting proposed OCT to kick off next phase (Navy & Marine Corps)

In progress: Analysis of all Air Force Officer and Enlisted occupations is complete. Midway through Army.

- DMDC occupational expert and ASVAB CEP subject matter expert occupational analyst are reviewing and validating Air Force links.
- Technical report and completion of Army analysis expected end of current task order period of performance.

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

UNIFORM

A Comprehensive Military Occupational Database

Program Initiative: UNIFORM

Goal: Develop a web-based application to house all Service-provided occupational information and streamline data collection, manipulation, and distribution in a unified format, seamlessly producing a comprehensive representation of the military career information accessible to all government and civilian entities.

Purpose: Currently, each Service submits unique data for over 8,000 active MOCs (MOS, AFSC, NEC collectively) to DMDC who then manually performs occupational analysis to link each code to other occupational structures based on commonality of skills, duties, and training. This process inhibits the timely analysis and dissemination of the military occupation information for career planning.

Status: The application is populated with data provided by DMDC. Reporting functionality is develop. Team is working with service IT reps to facilitate automation.

UNIFORM

UNIFORM

Log In

submit | update | discover
military occupational data

UNIFORM

UNIFORM is a database that contains comprehensive military occupational data in a unified format to help you understand today's world of work in the United States Military.

UNIFORM contains the latest military career information.

DOWNLOADS

ABOUT US

LIBRARY

DATA DICTIONARY

UNIFORM

CEP IPR

- August 16, 2018
- Discussions around MEPCOM Provided suggestions for CEP iCAT expansion
- Interim reports from contractors on existing initiatives to include Test Security, Needs Analysis, Expert Panel Recommendations, etc.

Shannon Salyer, Ph.D.

Shannon.d.salyer.civ@mail.mil

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB
CAREER EXPLORATION PROGRAM

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, 19-20 September 2018

Future Topics

- ASVAB Resources
- ASVAB Development
  - Pool Development
  - Evaluating/Refining Item & Test Development Procedures
  - Item writing guidelines and tools
- Adverse Impact
- PiCAT/Vtest Updates
- 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
- Automatic Item Generation
  - Arithmetic Reasoning and Math Knowledge
  - Other subtests
- Web and Cloud efforts

OPA
OFFICE OF PEOPLE ANALYTICS