DEFENSE ADVISORY COMMITTEE ON MILITARY PERSONNEL TESTING

January 22-23, 2025 Meeting

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

Minutes approved for public release.

Fred Oswald

March 02, 2025

Dr. Fred Oswald, Chair, DAC-MPT DATE

DEFENSE ADVISORY COMMITTEE
ON
MILITARY PERSONNEL TESTING

January 22-23, 2025

The Fiscal Year (FY) 2025 first session of the Defense Advisory Committee on Military Personnel Testing (DAC-MPT) was held at the Hilton Garden Inn El Paso University, El Paso, TX on January 22-23, 2025. The meeting was conducted in person; however, one DAC-MPT committee member participated virtually using the Microsoft® Teams online collaboration tool. The Assistant Director, Office of Accession Policy (AP) opened the meeting by stating that it was being held under the provisions of the Federal Advisory Committee Act (FACA) of 1972 (5 USC, Appendix, as amended), the government in the Sunshine Act of 1976 (5 USC, 552b, as amended), and all other governing Federal statutes and regulations, and open to the public. The Assistant Director (AP), said the meeting agenda was available on the DAC-MPT website[1] and public comments would be received at the end of each day's scheduled sessions.

The Assistant Director thanked the committee members for their participation and the presenters for their support of the committee's activities. Addressing the administrative components of the virtual meeting, a complete record of attendance was collected. The Assistant Director also informed participants that the meeting was *not* being recorded on the Microsoft Teams® system. Teams participants were directed to mute their devices and to click the "raise hand" button when they wanted to speak. Participants then introduced themselves.

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

1. Accession Policy Brief (Tab E)

The Director and Assistant Director of AP presented the briefing.

> The Director (AP) began by comparing FY23 recruiting results to those of FY24. FY23 was the most difficult year since the inception of the All-Volunteer-Force and the first time since 1979 that three DoD active components failed their recruiting goals. Only the Marine Corps and Space Force met their recruiting accession missions. In FY24, however, all DoD components, except for the Navy Active Duty, Army Reserve and Navy Reserve, met their missions. The Director (AP) clarified that the Navy Active Duty actually made its contracting goals yet fell short of shipping all 40,600 due to basic training capacity limitations. The presenter also noted that the Services FY25 Delayed Entry Program (DEP) numbers were 10% higher than during the first quarter of FY24.
>
> The Director (AP) then presented two graphs and two charts to illustrate the current and future recruiting market dynamics. The first graph showed that, due to decreasing birth rates, the 18-24 youth population will begin declining in 2026, from over 31 million to just under 28 million in 2045. A second graph showed

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

an impending decrease in college attendance that corresponds to the youth population decline. A chart depicted the increase in the percentage of non-prior service accessions with waivers from 17% in 2022 to 34% as of the end of the 3rd quarter of FY24. The second chart showed the results of a Joint Advertising Market Research and Studies (JAMRS) Youth Poll conducted in the Fall of 2023 on reasons for not joining the Services. The top ten reasons and percentages who identified those reasons, were (1) possibility of physical injury/death (73%), (2) possibility of post-traumatic stress disorder or other emotions/psychological issues (66%), (3) leaving family and friends (60%), (4) other career interests (47%), (5) possibility of interference with college education (38%), (6) dislike of military lifestyle (37%), (7) too long of a commitment (36%), (8) required to live in places I don't want to live in (35%), (9) don't want to be deployed overseas (33%), and (10) possibility of sexual harassment/assault (32%).

The presenter closed by presenting strategic lines of effort designed to improve the accession pipeline by growing propensity and expanding eligibility. Propensity is grown by increasing awareness, consideration, and motivation to serve. Four initiatives support this objective. The first is the launch of the JAMRS adult influencer media campaign and the youth digital media campaign. TV/streaming commercials aired from 30 September-November 10, 2024. Adult influencers who see at least one JAMRS ad are 47% more likely to recommend military service. The second is to develop a standardized methodology to provide states with military affiliation data to include military readiness into their education accountability plans. This will incentivize school officials to promote benefits of military service. The third is to continue to work legislative proposals that improve quality access. The fourth initiative is to coordinate and collaborate with industry, academia, non-profits, the military, and across government to operationalize permeability and grow interest in public service.

The objective of expanding eligibility is to expand the aperture for those interested in serving, and this is accomplished through four initiatives. The first is to expand the Medical Accessions Records Pilot (MARP) from 38 to 51 conditions. Conditions include asthma in the last 4 years; Attention-Deficit Disorder (ADD) and Attention-Deficit/Hyperactive Disorder (ADHD), for which adjustments were made to time-since-last-diagnosis; and learning disorders, to be added within one year. The second is to explore the feasibility of exploring alternative medical accessions standards frameworks based upon updated information, medical advances, and a range of possible assumptions. The third is to develop a Joint Enlistment Composite for the current non-cognitive personality test (Tailored Adaptive Personality Assessment System [TAPAS]) that can be used to redefine applicant quality and expand the pool of eligible applicants by adding personality into the definition of quality. The fourth initiative is to develop an Armed Services Vocational Aptitude Battery (ASVAB) special purpose test – a new assessment of fluid intelligence called Complex Reasoning (CR) – which is less reliant on traditional academic knowledge and now available to the Services. The CR test is to be evaluated for inclusion into Armed Forces Qualification Test (AFQT).

At the end of the briefing, a committee member asked for more information on the reasons for the populations' lack of trust in the military. The Director (AP) referenced an across-the-board decline in trust in government institutions and cited a similar decline in military favorability metrics. The Director suggested these trends may be tied to an increasing lack of familiarity with the Services and relationships with current and former Servicemembers, as well as messages from the media and culture that portray military service in a negative or neutral light. The committee member asked for more information on the Future Servicemember Preparatory Course (FSPC). The Assistant Director (AP) said two Services currently offer the course, and the course covers physical fitness and aptitude, includes in-person and self-based learning, and is intended to bolster academic skills needed to prepare individuals for military service. The Assistant Director (AP) said the courses can accept applicants with AFQT scores between 10 and 30. A representative from the U.S. Army Research for the Behavioral and Social Sciences (ARI) said close to 90% of FSPC participants achieve an acceptable AFQT score. The Assistant Director (AP) said the Army's course was established first, and initially the Navy was enlisting applicants with lower scores than the Army. A Navy representative said the Navy's pass rate is around 70%.

The Director (AP) said up to 4% of all active duty accessions can be CAT IV applicants. To go beyond 4% up to 20% Services would require approval from the Secretary of Defense. The Director (AP) said the classification opportunities for CAT IVs are limited. The Assistant Director (AP) added, if an applicant obtains a CAT IIIB or higher score within the same year they enlisted, they are not counted against the 4% limit.

A committee member asked about the status of those who participate in the program, and the Assistant Director (AP) said they are considered Service Members. They are at Basic Combat Training (BCT) sites for at least 3 weeks, and, as clarified by an Army representative, are in a separate unit from other BCT trainees. The majority move directly into BCT after graduation from FSPC. The Army representative said they have drill sergeants, just like basic trainees, and FSPC staff help them with classification based on their job skills. The Assistant Director (AP) said, because they are in a BCT-type environment longer than other recruits, they frequently are assigned leadership positions in their BCT units earlier than others.

A committee member said the briefing was very helpful for understanding the landscape of the recruiting environment, for example, the demographic cliff (age related), over and above the usual testing-oriented topics. The committee member also thanked the Service members who were present at the meeting.

An Air Force representative asked about the through-put of the Navy and Army courses. A Navy representative said the Navy program began in FY24 and has processed 4,200 personnel, reclassifying almost 1,600 of those into more technical occupations, successfully reducing pressure on recruiting to fill those slots. An Army representative said there have been approximately 40,000 individuals who have gone through the Army FSPC over the last 2 years, and that their subsequent performance is comparable to other accessions. A representative of the Human Resources Research Organization (HumRRO) asked if the Army has looked at post-training attrition for the FSPC cohorts. The ARI representative said attrition has been examined in two ways: ARI looked at whether FSPC accessions performed like other Soldiers and found comparative attrition numbers to be lower within the FSPC cohort, though if attrition during the FSPC course is counted, the FSPC attrition rates are the same or a little higher than non-FSPC cohort rates. A Navy representative said the Navy is conducting analyses to compare those who reclassified through the course against direct accessions in selected technical jobs to determine if the intervention (participation in the course) is consistently successful.

The Assistant Director (AP) said Services can brief the committee on these outcomes in future meetings. An Air Force representative asked if there were differences linked to Tailored Adaptive Personality Assessment Scales (TAPAS) scores, and an Army representative said the Army is examining that.

2. <u>R&D Milestones Brief</u> – (Tab F)

The Acting Director, Defense Testing and Assessment Center (DTAC), presented the briefing.

The presenter began the presentation with an overview of the projects to be covered in the briefing, including ASVAB development, ASVAB and Enlistment Testing Program (ETP) revision, Career Exploration Program (CEP), Military Compatibility, and Test Availability.

- ASVAB Development includes item development efforts, new Computerized Adaptive Testing ASVAB (CAT-ASVAB) item pools, new P&P-ASVAB forms, form development methodologies, and implementation of calculators.
- ASVAB and ETP Revision includes Next Generation ASVAB/ASVAB evaluations such as norming investigations and differential prediction analyses; evaluating new cognitive tests/composites for the ASVAB including CR, Computational Thinking (CompT), the Cyber Test (CT), and Mental Counters (MCt); and adding non-cognitive measures for selection and/or classification by creating a TAPAS validity framework and Joint-Service TAPAS.
- Other work is focused on the CEP, a military compatibility assessment (MC), and expanding test availability (e.g., web/cloud delivery of ASVAB and special tests and device expansion).
- Ongoing efforts to develop new items for the ASVAB. This includes developing items and graphics across subtests and converting Assembling Objects (AO) item graphics for alternative device compatibility.
- New Computer Adaptive Testing (CAT-ASVAB Item Pools). The objective of this project is to develop CAT-ASVAB item pools 16 – 20 using new items. The implementation date for these pools has not been determined. Forms 11 – 15 were implemented and documented in a technical bulletin in February and May 2024, respectively.
- Developing new paper-and-pencil (P&P) ASVAB forms 29F/G, 30F/G, 32F/G, and 32F/G from new items. Project completion date is to be decided. The effort now includes updating scanning and scoring systems.
- Evaluation and implementation of calculator use. The objective of this effort is to move forward with incorporating calculator use on the ASVAB. The impact of calculators has been evaluated and a needs assessment is underway, with next steps to be determined subsequently. The project completion date is to be decided.
- Evaluate CAT-ASVAB methodologies and ways to streamline development efforts. A Bayesian item calibration sample size reduction study was completed in June 2024 and an evaluation of Differential Item Functioning (DIF) approaches is underway. Studies may compel adjustments to item seeding, calibration, and analysis practices, as well as form assembly and equating practices. Project completion date is to be decided.
- Evaluate the state of the ASVAB and prepare for the next generation of ASVAB and special purpose tests to be administered on the ASVAB platform in the ETP. These efforts are ongoing and will culminate with development of a roadmap for the next generation ASVAB, a validity argument for the ASVAB, and possible revisions to ASVAB contents.
- Develop and evaluate a non-verbal reasoning assessment (i.e., CR) for possible inclusion in the ASVAB and develop an item generator for the assessment. The 1 October 2024 National Defense Authorization Act (NDAA) CompT requirement to implement the assessment operationally was met and development of new items, CAT pools, and conventional forms, as well as ongoing evaluation, is underway.
- Develop a Computational Thinking (CompT) composite score to meet the NDAA requirement to address computational thinking skills. The CompT composite scores were implemented to meet the 1 October 2024 NDAA CompT requirement. Ongoing evaluations are underway.
- Integrate the use of non-cognitive measures in the military selection and classification process to ensure military compatibility among enlisted and officer populations. A Joint-Service TAPAS (JS TAPAS) and Phase 0 composite have been implemented operationally, with evaluation to follow. These efforts are ongoing.

- Revise and maintain all CEP materials (website and print materials) and conduct program evaluation studies and research studies as needed. These efforts are ongoing.
- Program ASVAB and special tests for delivery on DTAC's web-based/cloud-based platform and introduce enhancements. These efforts, which now include the launch of a new research initiative to investigate potential future program enhancement, are ongoing.
- Expand the Internet version of the CAT-ASVAB (*i*CAT) test delivery application to run on additional operating systems and browsers for desktops/laptops. Expand the Pending Internet Computerized Adaptive Test (P*i*CAT) and AFQT Prediction Test (APT) to run on tablets and smartphones. Completion date is Summer 2025, however, monitoring of operational performance across desktops/laptops, tablets, and smartphones will be ongoing.

There were no questions or comments after the briefing.

3. Update on Committee Recommendations – (Tab G)

The Acting Director, DTAC, presented the briefing.

The DAC-MPT makes recommendations to AP and DTAC following each bi-yearly DAC-MPT meeting, and the recommendations received between December 2022 and June 2024 are documented in this presentation. The presenter began the briefing by addressing DAC-MPT recommendations that it be updated regularly on AP activities, changes to testing programs, and the results of research efforts. AP provides a routine briefing to the DAC-MPT members, updating them on the current challenges. DTAC will work with AP to keep the DAC-MPT apprised of relevant changes and research efforts. The remainder of the briefing focused on recommendations on 28 topics associated with general research, program support, construct coverage, methodological matters, and other critical issues. The Acting Director (DTAC) presented responses from AP and DTAC for each recommendation.

Recommendations regarding major milestones and schedules for ASVAB R&D efforts:

1. In December 2022 the DAC-MPT said it appreciated the scope of research efforts and requested a curated list of technical reports (and access to them as appropriate) and updates regarding progress on this research.
    - DTAC believes the best resources for a "curated list of technical reports" for the DAC-MPT are the ASVAB, AFQT, and TAPAS validity frameworks. DTAC can work with AP (as allowed per FACA guidelines) to provide the most current documentation, and updates to the validity frameworks will be provided as they are completed (anticipated to be on a biennial basis).
2. In August 2023, the DAC-MPT said it wishes to be updated on the results of the research efforts being conducted and the plans for new research. The DAC-MPT also recommended that DTAC monitor developments in Generative Artificial Intelligence (GAI) and innovations in virtual proctoring.
    - DTAC agrees and said it will continue to keep the DAC-MPT apprised of research efforts. DTAC has recently begun a new effort to review AI, generative AI, and technology capabilities for testing and will plan to brief the DAC-MPT on the effort at a future meeting. DTAC continues to monitor trends in virtual proctoring and investigate new virtual proctoring technologies as they arise.
3. In June 2024 the DAC-MPT said it remained impressed by both the number of projects OPA/DTAC manages and the quality of the research produced. The Committee voiced a potential concern about the high workload of this group.
    - DTAC is dedicated to maintaining the highest quality testing program and is continually looking to investigate and refine its products and practices, standardize procedures, and introduce efficiencies, so they can alleviate workloads for their small but mighty team.

Recommendations on ASVAB/AFQT the validity framework:

1. In December 2022 the committee agreed that Theory of Action [TOA] was applied very successfully in the AFQT selection context presented in developing, justifying, and empirically supporting the claims that were tested and recommended continued use of the TOA as an organizing framework for validity.
    - DTAC agreed, saying it has continued to use the Theory of Action as an organizing framework for validity. DTAC is continually updating its AFQT, ASVAB, and TAPAS validity arguments based on their respective TOAs.

Recommendations on Device Expansion Plans:

1. In December 2022 the DAC-MPT asked about research on the interaction of item features and device variability to determine if different performance was observed for different items and tests when delivered on different devices, taking into account interactions among familiarity with the device, the task to be performed, response action, and device. Another question was raised about mode comparability research and the studies that were done or planned to ensure comparability of results across devices, operating systems, and browsers.
    - DTAC agrees and did take into account the interaction of item features and device variability and determined that these were not drivers of performance and response time differences. Familiarity of device was the only significant factor that sometimes (depending on device and subtest) resulted in significant response time and performance differences. Likewise, the past device evaluation efforts did address various device, operating system, and browser conditions. Again, familiarity was the only factor with any significant interactions.
2. Also in December 2022 the DAC-MPT made several recommendations regarding future research into alternate devices and their effects on test scores.
    - DTAC agreed, saying it has developed a device expansion maintenance plan that includes the collection of data from examinees regarding their test-taking experience, including how familiar they are with the device used. Examinees are encouraged to use a device they are familiar with before beginning the APT or PiCAT. DTAC plans to continue to research the impact of device expansion on performance differences, especially for new subtests added to the ASVAB battery.

Recommendations regarding adverse impact:

1. In December 2022 the DAC-MPT said it recommends regular analyses of adverse impact and exploration of potential reasons for differences in test performance to aid in promoting accessions into the Military Services. Future assessments of adverse impact should also consider whether English is the examinee's first language.
    - DTAC agreed and said it is developing a standardized analytic tool to evaluate adverse impact on an annual basis. DTAC does not currently have access to a standardized demographic question on language proficiency or English as the applicant's first language but can explore potential proxy variables.

Recommendations on the AFQT Differential Prediction Study:

1. In December 2022 HumRRO requested input from the DAC-MPT in three areas: the modified Cleary approach to assess differential prediction, other factors that may explain overprediction and underprediction, and approaches for dealing with limited power for analyses involving occupations with small sample sizes. Committee members noted that overprediction was expected and asked questions regarding combinations of outcome measures, the effect of the scores of individuals who did not make it into the study, the use of multilevel modeling for these multi-group analyses, other ways to probe differential prediction, (e.g., using the Johnson-Neyman regions of significance approach; Preacher, Curran & Bauer, 2006), and the use of multilevel modeling to address selection artifacts and comparisons involving technical and non-technical occupations.
    - DTAC said it appreciates the input received from the DAC-MPT.
2. In the same meeting, the DAC-MPT made several suggestions regarding modifications to this research that might be considered: using performance measures that are broader and more direct

than job knowledge tests, clustering related jobs or sorting jobs into technical and non-technical positions, using multilevel modeling as an analytic approach be considered going forward, and evaluating the effect of the test taker's native language. The DAC-MPT is aware that the data needed for these initiatives may not exist at all, may not be reliably collected, or may not be available for a sufficient sample of test takers.

- DTAC agreed, saying the use of broader and more direct job performance measures rather than job knowledge tests is being looked into by the Services, particularly the Army in terms of military fitness and suitability. However, for criterion measures intended to be predicted by outcomes appropriate for ASVAB and other cognitive ability tests, it will require extensive planning and execution that would take a lengthy amount of time to run through the course of development. Clustering related jobs or sorting jobs into technical and non-technical positions is something that could and should be done. We are looking into this as a possible extension on previous studies. Using multilevel modeling as an analytical approach is something that could be explored and utilized in the next study, such as differential prediction. It would be interesting to know which multilevel techniques (e.g., HLM) the DAC-MPT has in mind, and DTAC would appreciate further elaboration. Evaluating the effect of a test-taker's native language would be an interesting application for DLI Foreign Language Center students or English Language Center students. As of yet, this has not gone past the conceptualization stage. Also, it could be a challenge gaining cooperation with DLI as these students are engaged in rigorous courses of study in language acquisition involving full-immersion learning. DTAC appreciates the DAC-MPT's acknowledgment of limitations to their recommendations in that: (a) data may not exist, (b) data may not be reliably collectible, and (c) data may not be available for a sufficient sample of test taker.

Recommendations on the Non-native English Speakers analysis:

1. In August 2023 the DAC-MPT recommended considering how this report informs the development of the NextGen ASVAB. In addition, it may be useful to determine what level of proficiency is needed for Military Service. Appropriate MOS-relevant levels of language proficiency and criteria for measuring those levels should be revisited for the benefit of expanding recruitment and enlistment efforts.
    - Concur, AP explained that military training and operations are conducted in English. DoD supports programs such as Foreign Language Recruiting Initiative (FLRI) for non-native English speakers (NNES) to improve their English skills. To ensure all requirements are considered and to provide for the maximum ability to affiliate with the military, work on NextGen ASVAB will take into account the needs of the NNES within the constraints of the training and operational requirements. Furthermore, when developing classification standards, Military Services take into account training and job requirements to include minimum level of English proficiency required for all servicemembers, to include both NNES and Native English Speakers. Finally, the Department has developed additional non-verbal assessment of cognitive ability, which should aid with identifying individuals who have the potential to benefit from immersive English proficiency training provided by the DoD. DTAC/AP will share this recommendation with the MAPWG Service representatives for consideration by their respective Military Services when designing enlistment programs and developing classification standards.

Recommendations on Complex Reasoning:

1. In December 2022, the DAC-MPT said it valued the development of a complex reasoning measure because such a measure is lacking in the ASVAB, and virtually all jobs in the military require complex reasoning. The DAC-MPT suggested that future research consider including non-English speakers in the pilot study to increase the potential to validate the test for those populations.
    - DTAC said DoD policy currently requires applicants to speak, read, and write English fluently. Non-verbal assessment of cognitive ability should aid with identifying individuals who have the potential to benefit from immersive English proficiency training provided by the DoD. Recruiting non-English speakers for pilot studies poses some

exceptional challenges as general information about the studies and instructions are presented in English. Nevertheless, DTAC has included demographic questions about English proficiency in subsequent pilot studies in an attempt to address this recommendation. Very few (less than 1%) of participants report that they do not speak English well or not at all, which limits analysis. DTAC will continue to work to increase representation of non-English speakers in research and development efforts but must acknowledge logistical obstacles.

2. In August 2023, the committee made three recommendations, one each on measure development, nomological net, and validation. On measure development, the committee asked DTAC to determine why CR scores were "spiked" at a score of 11 across the three forms (this is unlikely to be coincidence). It also suggested continuing to expand the item bank. On the nomological net, the committee suggested correlating CR with ASVAB subtests to understand the nature of CR, where shared and unique sources of variance occur between the measures. Regarding validation, the committee recommended supporting the CR measure further with validity evidence drawn from sources such as past military studies involving similar CR measures, or research literature when the results are generalizable to the military setting, as well as from new studies with the current CR measure.
    - DTAC agreed with all these recommendations. On measure development, histograms presented at the August 2023 DAC-MPT were based on incomplete results. This "spike" at raw score of 11 appears to have smoothed out somewhat in the final sample that is twice as large as what was included in the DAC-MPT presentation. Follow-on work includes additional item development efforts to expand the item bank. On nomological net, DTAC said the analyses would be presented at the January 2025 DAC-MPT meeting. Regarding validation, DTAC has task orders in place for continued development and validation of CR and Computational Thinking composites to include plans for construct validation and criterion-related validation work.
3. In August 2023, the DAC-MPT recommended locating existing military data with CR-related data, in addition to conducting new validation work on the current CR measure (both selection- and classification-oriented validation). To this end, job analyses, O*NET data, and other resources may speak clearly to the need for an agenda for CR research across a wide range of MOS's.
    - DTAC agreed, saying criterion related validity evidence is typically the purview of the Services and that it will provide support with proposed research designs to facilitate cross-Service comparisons.
4. Also in August 2023 the DAC-MPT said future research might consider how CR might work in tandem with a recruit or enlistee's profile of ASVAB scores given potential implications for classification that considers each enlistees' current interests and future goals alongside broader recruiting and labor demands.
    - DTAC said criterion related validity evidence is typically the purview of the Services, but DTAC will provide support with proposed research designs to facilitate cross-Service comparisons.
5. In June 2024 members of the DAC-MPT commented on several aspects of the results of this work and voiced concern about the need for practice items for test takers who are not experienced with this item type. Aware of the time limitations for any individual test, the DAC-MPT recommends careful consideration of the impact of practice on the difficulty of the items.
    - DTAC is evaluating the impact of practice in the context of item presentation order and potential impacts on a Computerized Adaptive Test (CAT) version of CR. CR items are traditionally presented in order of increasing difficulty, which provides additional opportunity for experience and learning with these novel stimuli. This may necessitate a constrained CAT algorithm to accommodate such impacts.

Recommendations on Computational Thinking:

1. In December 2022, the DAC-MPT said it supports the development of the Computational Thinking [CT] measure via a composite and the plans for doing so. The Committee suggested increasing the representation of non-English speakers in the pilot study sample and reviewing the work of Zach Hambrick, who has developed a similar measure.

- DTAC concurred, but noted the same points made in regard to CR (e.g., requirement for applicants to be fluent in English, potential for identifying individuals for English training, small numbers of pilot study participants who are not English speakers). DTAC will continue to work to increase representation of non-English speakers in research and development efforts but must acknowledge logistical obstacles.

2. In August 2023, the DAC-MPT made recommendations on validation, fairness, and the Electronic Data Processing Test (EDTP). Regarding validation, given that a new measure solely designed to assess CT is not being developed, it could be useful in the time allowed to consider approaches that might refine the validation of CT composite further. Regarding fairness, a question important to the Services is, "Will selection/classification outcomes based on CT be fair in terms of minimal adverse impact?" This information was not provided, but given that there are some subgroup mean differences on ASVAB and other cognitive tests examined here, subtest composites can increase these mean differences. Regarding the EDPT, given that EDTP components look like ASVAB + CR subtests, and given the EDPT will not be given to all enlistees, consider removing EDPT from further research.
    - DTAC agreed with all these recommendations. Construct validation analysis results will be presented at the January 2025 DAC-MPT meeting. These will not include MOS-specific results. Nevertheless, DTAC will incorporate similar strategies in research design templates developed to assist the Services in further validation work. Fairness evaluation is part of planned analyses. EDPT is not part of future DTAC research plans.
3. In June 2024, the Committee appreciated the time-urgent need for developing the CT test and recommended that additional work should investigate subgroup differences and other fairness issues and conduct further validation research.
    - DTAC said updates on subgroup differences and construct validation plans will be presented at the January 2025 DAC-MPT meeting.

Recommendations for item analysis in the ASVAB item development process:

1. In August 2023 the DAC-MPT acknowledged the challenge of identifying suitable methods for evaluating dimensionality of ASVAB tryout items under sparse data conditions and proposed the potential use of basic CTT-based statistics, such as item-total correlations, as a viable option. The Committee also noted that planned missingness can be acceptable when researching the overall dimensionality (correlational structure) of measures; however, planned missingness is definitely not recommended when using scores for estimating individual scores in operational settings. Suggested solutions included the potential use of machine learning and inspection of the content of items to identify themes.
    - DTAC said it uses item-total correlations to evaluate item characteristics and quality. Tryout items administered under the planned missingness design do not contribute to operational scores.

Recommendations on CAT-ASVAB pool and P&P ASVAB form development:

1. In December 2022 the DAC-MPT inquired about the transformation steps taken in terms of equating to better understand the processes used and to ensure that variability was not being introduced as a consequence of methodology. Additional information on the efforts to detect and manage multidimensionality in data from CAT-ASVAB forms is also requested. The DAC-MPT also requests more information about the nature of the PC Test stimuli (length, content focus on informational vs. literary reading), given the research to meet operational constraints and ensure comparability between P&P and CAT.
    - DTAC agreed. A comprehensive briefing of CAT-ASVAB equating methodology and rationale was presented to the DAC-MPT on August 16, 2023 (Reeder; 2023a). A briefing specifically targeted toward addressing DAC-MPT concerns over potential of the equating procedure to produce biased or more variable scores at the individual level was presented on June 12, 2024 (Dahlke, 2024). Analysis results presented in both the 2023 and 2024 briefings indicate that the equating process serves its intended purpose without detrimental impacts on examinees' scores. The DAC-MPT was briefed on analytic methods for evaluating and managing multidimensionality in CAT-ASVAB tests on August 16, 2023 (Reeder, 2023b). Further investigation into dimensionality of the

Assembling Objects test will be briefed at a future DAC-MPT. A briefing on the comparability of P&P-ASVAB to CAT-ASVAB is planned for the January 2025 DAC-MPT.

2. In August 2023 the DAC-MPT recognized the importance of using the pool-specific scale transformation, in addition to relying on the IRT measurement invariance property, for the purpose of improving the congruity of composite distributions and qualification rates across different pools at a group level. However, the Committee recommended examining the potential bias that could arise from the pool-specific scale transformation when estimating applicants' abilities at the individual level. The committee suggested that a simulation study relevant to the question be designed to explore this issue. The DAC-MPT also raised a question regarding the consistency of using the same operational IRT scoring method that is used in scaling, equating, and other psychometric analyses. Additional rationale may be necessary if consistency was not maintained. The Committee also highlighted the importance of contemplating the implications of the project's outcomes that align with potential developments of NextGen ASVAB.
    - DTAC agreed. A briefing specifically targeted toward addressing DAC-MPT concerns over potential of the equating procedure to produce biased or more variable scores at the individual level was presented on June 12, 2024 (Dahlke, 2024). Analysis results presented in both the 2023 and 2024 briefings indicate that the equating process serves its intended purpose without detrimental impacts on examinees' scores. *Note: DTAC does not understand the questions regarding consistency of scoring methods and believes those questions to be a misunderstanding of the materials presented. DTAC uses Bayes modal estimation consistently in scoring.*
3. In June 2024, the committee praised the thoroughness of the equating simulation study, viewing it as a valuable confirmation that the two-stage equating process works effectively at both the group and individual levels. The Committee recommended examining whether the results without the second stage produced similar outcomes. If the procedures with and without the second stage yielded comparable results, the possibility of simplifying the entire equating process in the future, if desired, could be contemplated.
    - DTAC has previously presented results indicating that relying solely on the IRT invariance property (i.e., without the second stage) does not produce similar outcomes with respect to the equipercentile objective of qualification rates (Reeder; 2023a). Follow-up analyses will be presented at the January 2025 DAC-MPT to illustrate these impacts within the same simulation framework as the June 2024 presentation.
4. Also in June 2024, the committee made two recommendations on calibration sample size. First, it recognized the importance of examining alternative calibration methods with smaller sample sizes. The differences in calibration results between flexMIRT and BILOG-MG were generally small, suggesting that the calibration program could be suitably replaced. The Committee raised a question about whether these differences could be further minimized by aligning the calibration settings of the two programs as closely as possible. In addition, the Committee recommended that DTAC consider the implications of switching the calibration program, including the need for recalibration of the current pools with the new program. Second, the study on sample sizes showed that the psychometric properties, particularly reliability, did not change substantially across different sample sizes ranging from 700 to 1,200, supporting the use of a smaller sample size in the future. The practical benefit is clear, in the sense that a calibration sample size of about 970 would reduce the current data collection period by 8.3%. However, the Committee believes it is prudent to examine the impact of a smaller sample size on other aspects of the test, such as examinees' scores, DIF analysis, and more.
    - DTAC agreed. Although there is not an immediate need to replace the current operational calibration procedure, DTAC is poised to replace BILOG-MG if and when circumstances dictate it is necessary. DTAC does not believe recalibration of the current pools is necessary given current robust scaling and equating procedures. Additionally, DTAC is currently engaged in research to evaluate impacts of smaller calibration sample sizes for DIF and other item-level analyses that are part of the pool development process.
5. In June 2024, the DAC-MPT responded to the use of machine learning and natural language processing. First, the committee praised the proposed system's use of modern technology and its

potential to streamline ASVAB form development and inquired whether generative artificial intelligence had been considered or used in this process.
    - DTAC is evaluating the security-related implications of incorporating generative models into this process but believes they can add value if content and process security can be assured.

December 2022 recommendations (6) on norming requirements and plans and June 2024 recommendations (2) were made on continuing norming efforts.

1. One Committee member asked for a plot of trend results for AFQT scores.
    - Select AFQT trends were presented during the June 2024 DAC-MPT (McCloy, 2024). DTAC has developed a template analysis to monitor AFQT and other ASVAB score trends over time.
2. The Committee discussed the possible effects of COVID on test scores, noting that some groups were more affected than others. The DAC-MPT recommends that efforts to re-norm should be deferred until the effects of COVID on propensity to serve have abated.
    - DTAC responded that the technical working group (TWG) noted post-pandemic drops in student scores on NAEP, Measures of Academic Progress (MAP), and other standardized tests. They noted the effects of school closures and remote learning could take a decade or more to rectify as most K–12 students were affected.
3. The DAC-MPT recommended that DTAC be sensitive to changes resulting from more vulnerable groups being differentially affected and wait until more time has elapsed before initiating a major re-norming effort. In addition, the methodology used for re-norming the ACT and SAT should be considered as plans to re-norm the ASVAB are developed.
    - DTAC presented a summary of re-norming options and contingencies during the June DAC-MPT (McCloy, 2024) that include considerations for (a) the disruption to schooling that took place during the COVID pandemic, (b) differential impact of disruption to schooling, and (c) multiple methodological approaches to potential re-norming. DTAC agrees that waiting for the full impact of schooling disruptions is understood.
4. The Committee also explored the development of norms based on the applicant pool instead of the customary approach of using the entire population. The DAC-MPT recommends that the DTAC consider the relative advantages and disadvantages of each approach before deciding which approach to use.
    - DTAC agreed. The TWG considered five options for renorming the ASVAB, including applicant-based norms. DTAC will consider the arguments for and against each approach as summarized in the June 2024 DAC-MPT (McCloy, 2024) briefing.
5. The DAC-MPT agrees with the presented results and does not believe that the age of the scale alone is a reason to renorm, noting that there may be public resistance to changing the long-standing interpretations of the scale. The DAC-MPT felt that the TWG had carefully considered a number of different advantages and disadvantages and had no suggestions for further work to inform the decision regarding renorming. The costs and common interpretations of scores further limit interest in renorming.
    - DTAC is aligned with the DAC-MPT and TWG in believing there are few if any substantive reasons to renorm at this time.
6. The DAC-MPT agrees with the TWG that renorming is not needed at this time; however, the Committee recommends continued monitoring of ability and demographic changes in the population.
    - DTAC is working with a data monitoring/visualization tool to assist in evaluating NAEP, SAT, ACT and Census data trends in relation to ASVAB/AFQT scores and demographics.

Recommendations on the use of calculators on the ASVAB:

1. In December 2023 the DAC-MPT recommended continuing the planned research approach, which it said should incorporate: a clear articulation of the problem, planned needs analysis, impact on psychometric properties, a thoroughly designed transition including potential need for training of test administrators and applicants on calculator use and standardized roll out across the Military

Services, continuous program monitoring, and careful definition and collection of appropriate outcome data.
    - DTAC agreed. Substantive updates on the research plan, including empirical impact analyses and needs analysis, will be presented at the January 2025 DAC-MPT meeting.
2. In the June 2024 meeting, committee members and other participants asked a number of questions, including concerns about adverse impact and individual differences when a calculator was used, the responsibility for bringing calculators to the test administration session, the need for training on the use of a calculator, the process of equating all applicable forms, and the potential need to examine calculator use and score differences by MEPS location. Committee members also raised questions about the relationship between the nature of Arithmetic Reasoning (AR) items and the effects of calculator use, the introduction of test anxiety when calculators are allowed, alternative analytic approaches (e.g., correlational studies), and the impact of calculators when the ASVAB is administered on tablets.
    - DTAC shares these concerns and will present further detail at the January 2025 DAC-MPT meeting.
3. Also in June 2024, the DAC-MPT said the research presented was well done and informative. The DAC-MPT looks forward to seeing the full results from Study 2 and Study 3. Given the study results and logistical concerns, the DAC-MPT does not find value in allowing the use of calculators on the ASVAB and does not anticipate that this effort would increase the number of qualified applicants.
    - DTAC agreed. More comprehensive findings from the empirical impact study and needs analysis will be presented at the January 2025 DAC-MPT meeting to address many of the DAC-MPT's concerns, which are shared by DTAC. Given the ambiguity of the problem definition, challenging timeline, administrative barriers, and potential scope of the impact, DTAC's capacity to address emerging issues revealed by these studies may be limited.

December 2022 and June 2024 recommendations on the Next Generation ASVAB/Testing evaluation plan and stakeholder focus group study:

1. The DAC-MPT asked how DTAC defined improvements in selection (e.g., increases in validity or satisfaction). The answer will require another look at the philosophy or purpose of the ASVAB. The DAC-MPT recommends careful consideration of the criteria for "improvement."
    - DTAC agrees that careful consideration of the criteria for improvement in selection is needed. DTAC has been actively considering the criteria and process for making changes to the ASVAB since 2011. A detailed plan for NextGen ASVAB was presented to the DAC-MPT in 2020. Regarding the philosophy of the ASVAB question, DTAC completed a thorough review in 2023 of all the ASVAB philosophy discussions that took place over the past several decades and concluded that the DAC-MPT's 2011 recommendation to articulate the ASVAB philosophy might have unintentionally led to an impasse between DTAC and the Services regarding ASVAB content decisions due to competing philosophies. As such, current thinking is to remove references to a specific philosophy and reframe ASVAB content discussions to focus on guidelines and evaluation processes that have been mapped out. DTAC continues to solicit input from stakeholders to ensure that the different purposes for which they use the ASVAB continue to be met.
2. Committee members recognized the diversity of needs among stakeholders. Although a completely shared vision for the ASVAB is likely impossible, there are no major complaints, and DTAC is hoping to meet most of the stakeholders' goals. The DAC-MPT encourages continued efforts to evaluate stakeholder perceptions and to educate them on the compromises that must be made.
    - DTAC continues to communicate with stakeholders to learn their differing needs and perspectives, build a shared understanding, and help identify a way forward for Next Generation Testing. Most recently, DTAC held a 3-day workshop with a variety of ASVAB stakeholders in November 2024, as well as conducted interviews with additional stakeholders not participating in the workshop.

3. Committee members discussed the issue of the length of the tests and briefly explored alternatives such as changing the CAT stop rules, moving item seeding requirements from proctored testing to VTest administrations, employing psychometric refinements, using a multidimensional approach (e.g., multidimensional IRT), and initiating a taxonomy content review to identify redundancies. Although the tests are already short, the DAC-MPT recommends that DTAC continue to explore various ways to shorten the length of time required for administering the ASVAB and special tests.
    - DTAC agrees and continues to consider avenues to reducing testing time to alleviate the burden on MEPCOM resources. Testing time was a focus of one of the exercises conducted at the November 2024 ASVAB stakeholder workshop.
4. The Committee also discussed applicant perceptions of the ASVAB. The available data were collected from individuals who had taken the ASVAB but had not yet completed training and did not include high school students taking the CEP or applicants who were not accepted. The DAC-MPT encourages efforts to understand a broader range of applicant reactions to the ASVAB.
    - DTAC agrees that it would be useful to get the perspectives of CEP participants and applicants that do not qualify for entry into the military but also notes that these are difficult populations to get access to. In focus groups that were conducted with qualifying applicants, a number discussed taking the ASVAB via the CEP. If there are future focus group efforts, we will make every effort to speak with as broad of a swath of the test-taking population as is practically feasible.
5. Regarding the stakeholder focus group study, the DAC-MPT asked about the representation of the study participants relative to the populations. Demographic information about the participants was limited. Consequently, the sample did not meet strict sampling conditions. The DAC-MPT recommends that future focus groups ensure adequate representation of all critical groups.
    - If there are future focus group efforts, DTAC will make every effort to speak as broad a representation of relevant subgroups as is practically feasible.
6. In June 2024, DAC-MPT supports the systematic approach to considering future changes to the ASVAB and has no substantive comments to make, other than that the focus groups of panelists should consider the needs of the Services in the future, as they make their judgments on which tests should be included.
    - DTAC concurs. DTAC has recently conducted a Next Generation ASVAB workshop with various stakeholder groups, including technical representatives, policy representatives, recruiters, classifiers, and trainers from the Services. DTAC plans to keep the Services involved as Next Generation ASVAB efforts unfold.

August 2023 recommendations on the High School Curriculum Study:

1. The DAC-MPT would like to hear more about this research and understand how the NextGen ASVAB and the Critical Thinking and Complex Reasoning Tests support alignment with common high school curricula. The DAC-MPT also suggested that researchers consider multilevel analyses on variables like school and state to test the hypothesis that schools with more resources provide more courses. Another suggestion was to consider the extent to which such information could be used to assess schools from a workforce development perspective. Another possibility to investigate was whether or not schools offering curricula aligned with ASVAB subtests and better resources offered better recruiting environments and produced more eligible students with a propensity for Military Service.
    - For clarification, DTAC is using the common high school curricula study as a separate source of information to support the NextGen ASVAB work. That is, we are not looking for the high school curricula study to support the inclusion of Complex Reasoning and Computational Thinking. While the data collection plan has already been established and implemented, DTAC will take a multilevel approach, to the extent possible, with the existing data to explore the hypothesis that schools with more resources provide more courses. Likewise, DTAC will consider an extension of the work to assessing schools from a workforce development perspective but would like to hear more from the DAC-MPT on what they envision and how this work could improve the composition of the ASVAB for selection and classification purposes. Another follow-up effort DTAC will consider is collaborating with other DPAC teams to determine whether schools with

better resources that offer curricula aligned with ASVAB subtests offer better recruiting environments and thereby also produce more eligible students with a propensity for military service.

December 2022, August 2023, and June 2024 recommendations on the ASVAB CEP:

1. The DAC-MPT suggested that the "Bring ASVAB CEP to your school" program be looked at closely to determine if the scheduling forum has pushed people away since 2019 and if the demographic questions on the forum should be revised.
    - The form was revised to allow the user to input only critical information to allow for ESS follow-up.
2. The DAC-MPT also suggested that students be assigned an identification code (e.g., pseudo name or number) to reduce the concerns about Military Service.
    - DTAC agrees. Collaboration with USMEPCOM is required to modify the score sheet.
3. Other suggestions included using social media to facilitate a culture of interest in schools, emphasizing the focus on exploring jobs and work as opposed to college and stressing the "whole-person" nature of the assessment.
    - DTAC launched social listening activities and social campaigns tailored to educator sharing and promotion among the educator community.
4. The Committee also felt that strong student testimonies placed on the homepage might engage more users and should be considered.
    - DTAC agrees. This information is gathered when possible (challenge: multiple layers of approval required to contact students but Educational Services Specialists (ESSs) can and do encourage student self-posting).
5. Other efforts to engage more users include working jointly with programs like Upward Bound and offering the program to undeclared freshmen in college and those in the TRIO program.
    - A new business strategy was activated in 2023 to engage underserved populations and broaden efforts in community colleges and other organizations with relevant populations.
6. YouTube videos that are aligned with the topics in the "Student Articles" would also be helpful.
    - Relevant videos have been created and are being developed that align with this suggestion.
7. Understanding other programs in high school that compete with the ASVAB-CEP could help direct marketing efforts, and the use of social media tools such as Kahoot could enable better connections among educators, students, and the military.
    - An in-depth Competitor Analysis is underway. A white paper was provided to Accession Policy that outlined specific comparisons between ASVAB CEP and SchoolLinks.
8. No major concerns were uncovered; however, the DAC-MPT would like to see more information regarding the Army's success in using CEP scores for enlistment. The DAC-MPT also requests that the following questions be addressed in future meetings: Should ASVAB-CEP be mandatory for high school students, and what will be the ramifications for the military services? What methods will best persuade students to take the ASVAB-CEP and take it seriously? How can the military promote, "Do you know people like you who took the ASVAB-CEP"? How should non-cognitive measures be incorporated into the selection and classification programs? How does the content in high school curricula align with the ASVAB, and what are implications for changes to either or both?
    - Where not yet briefed to the DAC-MPT, recommend adding to the agenda for future meetings.
9. The Committee endorses the idea of the Committee members working through the website to better understand the program.
    - A walkthrough was provided at the 08/23 DAC-MPT meetings, and login credentials have been provided to Committee members
10. In August 2023 the DAC-MPT complimented the tool and made a recommendation to identify ways to evaluate user engagement that goes beyond merely counts of accessing the website, such as by measuring frequency of return users. The Committee also endorsed the idea of better explaining the program, so that more participants take advantage of the Post-Test Interpretation service.

- DTAC agrees. “Return User” has been added to the Key Performance Indicators on website analytics. The team is also exploring a pop-up survey to be administered to gather more specific feedback. Better explaining the program correlates closely with ongoing efforts to standardize training and program delivery, disseminate marketing communications, and introducing the Ambassador Program.

11. In June 2024 the DAC-MPT continues to believe that the ASVAB CEP is an important tool for identifying potential recruits for the Services and provides a public service to youth in America. The biggest shortfall in this program appears to be its limited use. Consequently, the DAC-MPT encourages continued marketing efforts to inform the public in general and high school leadership specifically.
    - DTAC agrees and expanded and refined marketing efforts to reach target audiences including school board members, community colleges, superintendents, and state- and district-level decision makers.

August recommendations on the TAPAS validity framework and Joint Enlistment Composite:

1. The DAC-MPT suggested that the feasibility of a synthetic validity approach should be explored as a way to make the most of the available data given their variability and sparseness. A further suggestion was to consider strategies to collect validity data retrospectively (i.e., concurrent validity). The Committee also asked about the use of the TAPAS composite scores and the weights for its multiple components. For the purpose of the Joint Services (JS) Composite, the weights might be common across all Services, but individual Services might build additional composites and each assign unique weighting schemes. The DoD is tasked with producing the weightings. Another suggestion was to include other TAPAS facets for future research.
    - DTAC has a plan in place to explore suitable criteria, which begins with reviewing past work by contractors to establish a common set of criterion measures. The goal is to map any existing measures within the existing framework to military compatibility efforts. Likewise, we are asking the Services to also offer their criterion measures and experiences. Finally, DTAC will propose additional avenues for validating the TAPAS military compatibility composite, including possible synthetic and concurrent validity approaches, as suggested. Included in the validity research will be a review of the facet weighting schemes applied. As TAPAS development evolves, facets will be refined, and new facets will be developed to better support the assessment of military compatibility. Refinement efforts are planned for FY25, and new development will begin in FY27. The Services have the flexibility to introduce new Service-specific facets within the JS TAPAS.

August 2023 recommendations on the TAPAS for Military Compatibility:

1. The members of the DAC-MPT questioned the definition of military core values and the extent to which they are incompatible with counterproductive behaviors, which are also difficult to define and measure. Another member of the Committee suggested that the challenge of measurement might be addressed by identifying a criterion more proximal to the actual counterproductive behaviors (if those were specifically elaborated), which would sacrifice generalizability for fidelity to specific trait identification/prediction. The committee also suggested considering the possibility of deconstructing counterproductive work behaviors into essential components (e.g., making verbal comments as a prelude to physical altercations) as a strategy to address the low base-rate issue. A further question was raised about the relative stability of the characteristics to be assessed and the extent to which pre-accession assessment of these constructs might be useful for the prediction of later behaviors. Multi-level unit of measuring these constructs over time was suggested as a possible alternative.
    - DTAC agrees and has ongoing plans to explore suitable criteria, which begins with reviewing past work by contractors to establish a common set of criterion measures. The goal is to map any existing measures within the existing framework to military compatibility efforts. Likewise, we are asking the Services to also offer their criterion measures and experiences. There are 10 categories of misconduct that the military compatibility composite will address. It is these 10 categories for which we will focus on finding suitable criterion measures. Unfortunately, research shows that the military core

values across Services are not correlated with (or a reverse measure of) the 10 categories of misconduct. DTAC plans to explore multi-level measurement models to address possible issues with stability.

2. The DAC-MPT expressed a great deal of concern about the extent to which TAPAS could defensibly predict CWBs, adherence to military core values, and military compatibility in the general case or at a more specific, granular level targeting more clearly articulated CWBs. The ongoing work to establish a validity argument for TAPAS for varied purposes and uses suggests that the outcomes associated with TAPAS use are variable, and considerable work will need to be done around construct definition (including specificity), the stability of the construct at pre-accession and over time for various examinee groups, the validity argument for the use of this measure for purposes such as disqualifying enlistment candidates or identifying potential issues, and interpretation and use generally. The DAC-MPT recommends that considerable attention be paid to determining what should be measured in a compatibility assessment for articulated specific purposes. In addition, Accession Policy should be open to instruments other than TAPAS that provide targeted information that could predict counterproductive work behaviors in general or specific counterproductive work behaviors, adherence to military core values, and military compatibility.
    - The development of the military compatibility composite based on TAPAS facets is a phased approach. Phase 0 makes it possible to collect data across all Services on the Army Conduct Composite, which is our first military compatibility composite. With these data, we can begin to explore issues related to validity, subgroup differences, and stability. DTAC has developed a targeted definition of military compatibility that focuses on 10 categories of misconduct. The JS TAPAS Military Compatibility Composite is not planned to be the sole source of evidence for disqualifying candidates from the Services. Instead, it will serve as a flagging tool that would invoke further investigation via a clinical psychological interview by a licensed clinician who would provide a Service eligibility recommendation. This two-stage approach is currently being refined and will undergo various levels of validity studies before implemented operationally. Phase 1 JS-TAPAS development will focus on refining the facet pools and the military compatibility composite. Phase 2 will focus on introducing new facets into the JS-TAPAS that support the increased validity for the military compatibility composite. Research in the area of military compatibility assessment is also ongoing for the Officer population where 13 existing assessments are being evaluated for their appropriateness. Findings from this research will inform the Enlistment Testing Program (ETP). Accession Policy and DTAC are open to instruments other than TAPAS. DTAC also plans to develop and pilot its own Situational Judgment Test, intended to address the assessment of military compatibility defined by the 10 categories of misconduct.
3. A final suggestion involved the use of a clinical assessment to follow up on high scores on facets predictive of counterproductive work behaviors. This two-stage process could save money by limiting the clinical evaluation to high scorers only.
    - DTAC agrees. The current plan is to structure the military compatibility assessment into two parts: 1) Use TAPAS Military Compatibility Composite (or equivalent composite for officers) to flag individuals at risk for deviant behaviors; and 2) Use the clinical assessment to obtain professional judgment on those flagged by the test in part 1.

June 2024 recommendations regarding updates on non-cognitive tests and the Best Practices project:

1. Members of the DAC-MPT applauded the careful approach to developing these measures and encouraged future research to pay careful attention to the criteria used for deviant behaviors, particularly those that occur less frequently. The literature on honesty and integrity may be a useful source of information.
    - DTAC has a plan in place to explore suitable criteria that begin with reviewing past work by contractors to establish a common set of criterion measures. The goal is to map any existing measures within the existing framework to military compatibility efforts. Likewise, we are asking the Services to also offer their criterion measures and experiences. Finally, DTAC will propose additional avenues for validating the TAPAS military compatibility composite, including possible synthetic and concurrent validity

approaches, as suggested at the Aug 2023 meeting of the DAC-MPT. Literature on honesty and integrity is a key resource that DTAC has been reviewing within the Best Practices Project, as that team contains an expert researcher in the area.

2. Members of the DAC-MPT voiced similar concerns regarding the weaker prediction of extreme forms of counterproductive work behaviors and advised attention to the criterion used, given the implications of using such an instrument to reject potential officers. A second recommendation is to examine demographic differences in future work, as well as the effects of providing warnings to keep respondents from minimizing or ignoring past misconduct.
    - DTAC has a plan in place to explore suitable criteria, which begins with reviewing past work by contractors to establish a common set of criterion measures. The goal is to map any existing measures within the existing framework to military compatibility efforts. Likewise, we are asking the Services to also offer their criterion measures and experiences. Finally, DTAC will propose additional avenues for validating the military compatibility facets/scales administered for the officer population. Presently, DTAC is exploring various scales within 13 existing assessments for their utility with an officer population. Likewise, DTAC will begin the development of a Situational Judgment Test aimed at addressing the 10 identified focus areas for military compatibility assessment. Validation research will include an exploration of demographic differences.

June 2024 comments on a review of legislation and policy and made a firm recommendation on resources:

1. The DAC-MPT has no comments on the law but noted that the legal requirements for minimum scores emphasize the importance of accurate equating of forms.
    - AP concurs. This is the normal practice of the Department and will continue to be followed.
2. The Committee asked about the maintenance of current funding levels, cautioning that the description of levels as "healthy" may result in future reductions or future "leveraging" of funding for other purposes.
    - DTAC continues to monitor funding levels to ensure that an optimal level of funding is maintained or that steps could be taken to secure additional funding, if needed.

At each DAC-MPT meeting, the committee makes recommendations on topics for future meetings. All appropriate topics recommended by the committee have been briefed or will be covered in future meetings. At the end of the briefing, the presenter asked if the committee had comments or questions.

Regarding the status of the CR test (slides 13-16), the Assistant Director (AP) said the Services are already administering the test for validation and, as more results come in, the Services may be able to incorporate it properly into their composites.

Regarding the form equating methodology, the Assistant Director (AP) asked the committee if DTAC had understood their recommendation and responded appropriately. A DTAC representative said there may have been a misunderstanding, stemming from prior briefings, that two different scoring methods were being used (maximum likelihood and Bayes Modal Estimation [BME]) but that was not the case. A committee member said the slide cleared up any lingering questions.

On slide 41, the Assistant Director (AP) said the TAPAS for Military Compatibility (MC) was still in the experimental and feasibility assessment phases, especially in regard to the integration of clinical assessment for those who scored high on selected facets. The Assistant Director (AP) said in addition to methodological issues there are many logistical factors, to include the timeframe in which medical consults would need to occur, and that processing guidelines and timelines must always be considered.

Regarding the DAC-MPT recommendation on best practices in relation to predicting extreme forms of counter-productive work behaviors (CWBs), the Assistant Director (AP) said DoD has a separate Advisory Committee that deals with the prevention of sexual misconduct, and that AP has briefed that committee on project progress. The Assistant Director clarified that multiple federal advisory committees cannot work on the same topic, so AP is trying to delineate responsibilities. The DAC-MPT will focus more on testing and other committee's will focus more on processes, because they include retired military service members who are more familiar with the military. AP is looking into the legal aspects of conducting joint DAC-MPT meetings.

At the end of the briefing, a committee member asked whether differential prediction was a priority. S/he[2] then proposed that O*NET could be used to facilitate job clustering. The Assistant Director (AP) agreed and explained that the CEP includes job crosswalks between the military and civilian domains, which are also provided to O*NET. A committee member reinforced the committee member's comments, saying that HumRRO's experience with O*NET would be very useful, and that O*NET could be a good resource on the synthetic validity work, specifically to determine how measures align and encouraging broader thinking about pathways across miliary occupational specialties (MOS); the O*NET data would serve as a guide between operations and the ASVAB.

A committee member said s/he appreciated how the DAC-MPT, DTAC, and HumRRO work together, specifically the responsiveness of DTAC and HumRRO's close consideration of and detailed feedback on the DAC-MPT's recommendations. The Assistant Director (AP) expressed appreciation for the comments and partnership, adding that the relationship has helped ensure AP's efforts are efficient and effective, in particular for obtaining the most qualified applicants for military service and maximizing fit with occupational specialties, and improving the CEP and how it is used. The Acting Director (DTAC) emphasized how the work done by the partnership lives in the design of the ASVAB program.

4. <u>Update on P&P Forms</u> – (Tab H)

A HumRRO representative presented the briefing.

> The presenter began by providing some background information on the P&P ASVAB. It is a linear fixed-form version of the ASVAB administered using physical test booklets and answer sheets. It is administered in the ETP and the CEP and produces standard scores on the same dimensions as CAT-ASVAB. It represents a very small share of the testing volume for ETP but a large share of the testing volume for CEP. HumRRO has developed new P&P-ASVAB forms for both ETP and CEP to replace the current sets of forms. Due to P&P-ASVAB being administered in a group setting (as opposed to individually, like CAT-ASVAB), testing time is at a premium. Exceeding the current total testing time is not viable for ETP or CEP.
>
> All items available for P&P-ASVAB forms were developed for and tried out in CAT-ASVAB. Items for some subtests were not directly compatible with the P&P-ASVAB design. For instance, Auto & Shop Information (AS) scores are computed as a composite of Auto Information (AI) and Shop Information (SI) scores for CAT-ASVAB. However, AI and SI must be administered and scored together as a single AS subtest for P&P-ASVAB. Based on dimensionality research that informed the development of CAT-

[2] This document uses the "s/he" convention to prevent the association of comments with specific Defense Advisory Committee members.

ASVAB, AI and SI items are calibrated, scaled, and administered separately for CAT-ASVAB. All CAT-ASVAB AI and SI item parameters are on their respective subtest scales, and the items are field tested with non-overlapping groups of examinees. As another example, in the past, P&P-ASVAB Paragraph Comprehension (PC) sections used a testlet design, with multiple items referencing each passage. In CAT-ASVAB, PC items use a stand-alone passage for each item. The presenter then displayed a chart showing the item count and time limits for the ASVAB subtests when administered in P&P format.

Given the differences between the two formats, adjustments are required for the P&P specifications. The transformations that link the AI and SI Item Response Theory (IRT) scales to a single AS scale must be estimated by defining a target AS scale that closely approximates the scores examinees would earn if it were possible to score AS as a composite of AI and SI scores. The number of items in the PC item sets have to be updated to account for the use of items with stand-alone passages instead of testlets by decreasing the item count to reduce the reading load and limit testing time demands while maintaining an acceptable level of score reliability. In addition, the number of items in Arithmetic Reasoning (AR) item sets need to be updated to mitigate speededness by decreasing the item count to limit testing time demands while maintaining an acceptable level of score reliability. Finally, potential changes to subtest-level time limits need to be identified to accommodate an increased time limit for PC.

The presenter first addressed the IRT rescaling method for AS. All IRT item parameters for available AS items are on separate AI and SI scales to support the CAT-ASVAB, where AI and SI get scored separately and those scores are combined into an AS composite. As mentioned, P&P-ASVAB must administer and score AS as a single subtest, so the AI- and SI-scaled items must be translated to the P&P-ASVAB AS scale before they can be used. AI and SI items are tried out with non-overlapping samples, so the data used to calibrate them cannot support a combined AS-scaled calibration. The initial plan was to collect new data to recalibrate a set of AI and SI items, but the intention now is to use a custom-built rescaling procedure to accomplish this. The first plan would involve administering CAT-scaled AI and SI items and anchor items from past P&P-ASVAB AS item sets to examinees, then calibrating all items together , scaling them on a single dimension. The anchor items' IRT parameters would be used to link the newly estimated parameters to the historical AS scale, and all items would be rescaled to it. The drawbacks are that this process is psychometrically suboptimal, it would be time consuming and expensive, and it is unclear how it would turn out given that it would violate an IRT assumption. The question then became, how can the AI and SI item parameters be shifted onto the AS scale without collecting new data?

The presenter then showed a chart displaying an example of Stocking-Lord (S-L) test characteristic curves (TCCs), and posed the question, if anchor items can provide the scaling information needed to rescale item parameters, can alternative scale-anchoring information be used to achieve the same effect? That is, instead of using anchor items' parameters to define the scale of a test, relevant scale information can be obtained from latent ability distributions of person parameters? AI, SI, and AS have latent means and standard deviations (SDs) from past research on the scaling of P&P-ASVAB and CAT-ASVAB. There is also an estimate of the correlation between latent AI and SI distributions derived from operational CAT-ASVAB data. Using these, a complete variance-covariance matrix relating AI and SI to AS, where AS is a composite of AI and SI, can be constructed. The variance-covariance matrix and means allow the parameters on one scale to be reflected onto another scale while accounting for their shared variance. Instead of anchor items, all that is really needed for rescaling is a relevant target TCC. S-L uses a target TCC based on item parameters that are already on the test's scale. The Modified Stocking-Lord Procedure (MSLP) constructs a target TCC by reflecting AI and SI TCCs onto a composite scale. The presenter then showed a chart illustrating this process. After using a multivariate density distribution to estimate the expected TCC for a subtest, that TCC can be used as a target in a rescaling procedure. From this point onward, the MSLP functions exactly like the traditional Stocking-Lord procedure in how it iteratively estimates coefficients: (1) identify a set of provisional linear rescaling coefficients, (2) use the provisional coefficients to rescale the item parameters, (3) use the provisionally rescaled item parameters to compute a TCC, (4) subtract the provisionally rescaled TCC from the target TCC, (5) compute the density weighted sum of absolute-value differences, and (6) repeat steps 1–5 until Nelder-Mead optimization reaches convergence with a relative tolerance criterion for TCC-matching objective function = 1e-8.

The presenter then discussed a simulation conducted to evaluate the MSLP. The purpose was to benchmark the MSLP's performance against relevant comparators and evaluate the accuracy of expected TCCs against empirical TCCs. TCCs from MSLP were compared to other calibration methods. These included co-calibration of subtest items with BILOG-MG and, after calibration, rescaling the item parameters to match the composite AS scale. Another option was fixed-theta calibration with MULTILOG, which is conceptually the most similar to what the MSLP is meant to accomplish because it allows item parameters to be expressed on the composite AS theta metric without strict dimensionality assumptions. AI- and SI-like item parameters were simulated based on multivariate-normal distributions (*a* and *c* parameters were scaled as logits). AI-like items were designated "Test A" and SI-like items were designated "Test B." In all 200 items per test were used to reflect current item-seeding practices. Person parameters were simulated from bivariate-normal distributions. Ability distributions were based on latent means and SDs for AI and SI estimated from recent operational CAT-ASVAB data. The correlation between Tests A and B were varied from 0.0 to 1.0 in 0.1 increments with16k simulees per correlation condition. Composite ability was an unweighted average of ability on Tests A and B. For each simulee-item combination, the simulee's true theta and the item's true IRT parameters were used to estimate the probability of a correct response. To introduce measurement error, simulee's probabilities of correct responses were compared to randomly generated values from a [0,1] uniform distribution. A simulee got an item correct if their probability of a correct response was greater than or equal to the random value. BILOG-MG parameter estimates were rescaled using latent means and SDs. Results were highly consistent across the 100 replications that were run.

The presenter then showed a series of charts showing the outcomes of one of the simulations. Based on those results the presenter concluded that the MSLP's expected TCCs were closely aligned with the empirical TCCs associated with the composite theta dimension, which supports their use as targets in the rescaling procedure. The MSLP performed well, even when the dimensions contributing to the composite scale were uncorrelated. MSLP-rescaled item parameters produced TCCs that were closely aligned with the expected composite-scaled TCCs. The MSLP solutions were quite similar to the results from fixed-theta calibrations and were better at recovering expected TCCs than were co-calibrations with BILOG-MG (especially when abilities were correlated < .7). Therefore, the MSLP appears well-suited for this use case.

Separate AI and SI items sets have been assembled that will be administered in the AS sections of the new P&P-ASVAB forms. The IRT parameters for the items assigned to the AI and SI solutions require rescaling before they can be combined into usable AS sections. To ensure that item parameters (and resulting theta estimates) are scaled consistently across forms, a single MSLP rescaling was applied to the complete sets of items instead of rescaling each form separately. The rescaled TCCs were plotted against the expected TCCs for these item sets. As a point of comparison, the "provisional" TCCs that ignore the differences in scaling and naïvely presume that the AI, SI, and AS scales are equivalent, were also plotted. The presenter then displayed charts showing the expected, provisional, and MSLP-rescaled TCCs. The presenter concluded this portion of the presentation by indicating that the MSLP is the recommended approach for obtaining AS-scaled item parameters. The MSLP's target scale can be defined as a composite scale. The scores produced using MSLP-rescaled item parameters represent the expected scores examinees would receive if it were feasible to score AI and SI separately and combine them into a composite and will increase the alignment of AS scaling between P&P-ASVAB and CAT-ASVAB. The MSLP is effective at mapping item parameters onto a target IRT scale. It is more accommodating of multidimensionality than co-calibration of items and does not require item-level data as would be the case with fixed-theta calibrations.

The presenter then turned to a discussion of length-reduction analyses for PC. Compared to past testlet-based PC sections, constructing new PC sections from items with stand-alone reading passages requires reducing the number of items administered to control the reading load. When shortening a test, the primary objectives are to maintain an acceptable level of score reliability and adequate coverage of the construct to support score validity. An additional goal was to minimize total word count. To evaluate the effect of form length on reliability, the P&P-ASVAB automated test assembly (ATA) procedure was run using varied PC specifications: form length: 9, 10, 11, 12, 13, 14, and 15 items *fully crossed with* Quadrature-Weighted Average IRT information: 2.0, 2.1, 2.2, 2.3, 2.4, 2.5, and 2.6. Not all combinations of length and information were possible due to the impact of length on test information. Forms with 9 items could not achieve average information greater than 2.3. Forms with 10 items could not achieve average information

greater than 2.5. Simulated test-retest reliability coefficients for PC scores (BME theta estimates) and composite scores that include PC were run using 10k simulees with abilities based on latent means and SDs. The presenter then showed a series of charts displaying the results. They suggested reducing the P&P ASVAB items sets to 10 stand-alone items and targeting the highest average information during form development. PC is already the shortest P&P subtest and administering 10 items allows for coverage of the blueprint categories. The 10-item solution offers competitive levels of reliability compared to other form lengths with a substantially lower reading load. Using forms with the highest average information corresponds closely to maximizing reliability. Reducing the length of PC had a trivial impact on the reliability of the composites that include the subtest. PC is never used in isolation for selection or classification, so the impact on composite reliability is more important than stand-alone reliability.

The presenter next discussed length-reduction analyses for AR. When exploring the impact of the recommended changes to PC on time limits, trends from all P&P-ASVAB subtests were benchmarked to provide context for the PC trends. The results suggested that AR appeared to be much more speeded than the other subtests. Overall, 3.5% of ETP P&P-ASVAB examinees failed to complete the AR section, while only an average of only 1% failed to complete each of the other subtests. This trend generalized to the CEP P&P-ASVAB, but with higher overall non-completion rates (6.5% for AR, 2.75% average non-completion rate for other subtests). This is likely due to a less-motivated examinee population. The CAT-ASVAB time limits are designed to target a 99% completion rate. The presenter then showed a series of charts displaying the results of these analyses.

Before evaluating the impact of reducing the number of items in P&P-ASVAB AR sections, six new 30-item sections were assembled which had gone through all necessary reviews, were free of enemy items, and had passed all other content checks. Rather than start over and repeat the form assembly/review process, these existing sections were used as the basis for reduced-length sections. The impact on simulated score reliability when the least reliable item was iteratively removed from each form was explored, examining solutions with between 5 and 30 items. The shortened item sets were used to simulate test-retest reliability coefficients for AR scores (BME theta estimates) and composite scores that include AR, using 10k simulees with abilities based on latent means and SDs. The presenter then showed charts displaying the test-retest reliability estimates for AR item sets used in various Service composites. Based on these results it appears that 25 items appears to preserve the reliability across all scores evaluated. This length also works well in covering all blueprint categories. As with PC, AR is never used in isolation for selection or classification decisions, so the impact on composite reliability is more important.

The next step was to reexamine the speededness trends for past P&P-ASVAB forms, omitting the last 5 AR items from the analyses, which can give a sense of whether shifting from 30 to 25 items is enough of a reduction to mitigate the speededness observed. This is not a perfect approach given that the last 5 items are also among the most difficult. The results are slightly optimistic, especially for the CEP where examinees have lower motivation. The presenter then showed a series of charts displaying the results. The presenter concluded that the 25-item sections allow scores to retain high levels of reliability, allow for good coverage of all blueprint categories, and seems a sufficient length to mitigate speededness concerns. The recommendation is to use the 30-AR item sets that have already been built and reviewed as the basis for the reduced-length forms, removing 5 items from each. The items should be removed based on their contribution to reliability, with a balance of removals across content areas.

Even after reducing the number of items in the new P&P-ASVAB PC sections, the 10-item sets had higher word counts than past PC sections. The greater reading demands of the new sections requires allocating more time to PC to avoid introducing speededness. The reading demands of the new PC sections and the PC sections from past forms were examined to estimate the necessary time limit adjustment. Consideration was also given to whether any other subtests could be donors of this additional time to avoid increasing the total battery-wide testing time. The PC sections were evaluated on five common reading metrics: (a) word count, (b) Flesh-Kincaid Age, (c) Flesh-Kincaid Grade Level, and (d) Flesh-Kincaid Reading Ease. Because estimates of the Flesch-Kincaid and Flesch metrics can vary across programs, two programs were used to compute them: TreeTagger (a part-of-speech tagger and lemmatization program) and Microsoft Word. The current P&P-ASVAB time limit for PC is 13 minutes. Because P&P-ASVAB is timed for

groups of examinees rather than individuals, the time limit should allow most examinees to finish. However, the time limit should not be set too high or examinees who complete the section more quickly will have to wait longer for others to finish. Both word count and overall reading complexity were considered. Based on word count alone, a time limit of 17 minutes would be appropriate (13 x 1.2794 = 16.6322 minutes). However, the reading complexity metrics suggested a roughly 10% change in the reading ease compared to the past forms. Based on both word count and complexity, 18 minutes should be appropriate (13 x 1.2794 x 1.10 = 18.295 minutes).

A response latency score was computed for each examinee on each subtest based on their responses to try out (i.e., unscored) items. Mean and SD of response latencies for each item were computed. The means and SDs for response latencies were used to convert all examinees' item-level response latencies to Z scores, and each examinee's item-level response latency Z scores across items within each subtest were averaged to get their composite response latency score for that subtest. Examinees' composite response latency estimates were converted to percentiles within each subtest, then organized into twenty equally sized ordinal categories, each of which spanned a range of five percentiles (e.g., the slowest response category included examinees who were at or above the 95th percentile). For each tryout item, the mean amount of time examinees from each response latency percentile category spent answering the item was computed. Some items assigned to the new P&P-ASVAB forms predated the CAT-ASVAB data that were processed in Steps 1 and 2, so linear regression analyses were used to impute missing item-level response latencies for each response latency percentile category. These imputation models based their predictions on items' 3PL IRT item parameters (difficulty, pseudo-guessing, and discrimination) and—for PC only—word counts. The complete database of item-level response latencies was merged with assembled forms' item lists and the sum of item-level latencies for each response latency percentile category within each form was computed. For each subtest, the mean estimated test time across forms for each response latency percentile category was computed to arrive at an overall summary of how much time examinees in each percentile category would require to complete an average form. The presenter then showed a table and graphs summarizing the results.

The presenter continued the presentation by summarizing the recommendations for alterations to P&P ASVAB specifications for new forms. These included using the newly developed MSLP rescaling technique to translate IRT item parameters for AI and SI items onto an AS scale. Also reducing the number of PC items from 15 to 10 to offset the increased text in the passages caused by shifting from testlet design to stand-alone items. Administering 10 items is sufficient to maintain acceptable reliability for composite scores. In addition, reduce the number of AR items from 30 to 25 given that previous P&P forms showed evidence of speededness. As with PC, 25 AR items are sufficient to maintain acceptable reliability for composite scores while mitigating the speededness effects. Finally, adjust the PC time limits from 13 to 18 minutes to account for the increased reading load, which remains greater than past PC sections even after reducing the number of items.

The presenter concluded the presentation by asking for committee feedback on (a) using the modified Stocking-Lord procedure to resolve the AS scaling problem for P&P ASVAB, (b) the recommended lengths of 10 PC and 25 AR items on the P&P ASVAB, and (c) the adjustment of P&P time limits to account for the new PC section's increased time requirements.

When the presenter said PC would include shorter passages and only one item per passage (slide 4), a committee member asked when that decision had been made, noting the stark difference between that approach and the more common design in which multiple questions are used to measure a person's understanding of the depth of each passage. The presenter said the shorter passages are a compromise to attain independence of items in a CAT environment. The issue with longer passages and multiple questions would be finding a way to use multiple questions that may demonstrate co-dependency in CAT. A DTAC representative explained that passages were also shorter to eliminate the need for scrolling, which is not supported in the CAT-ASVAB

interface. The presenter elaborated that scrolling would be more problematic with the expanded use of mobile devices.

On the Modified Stocking-Lord Procedure (MSLP) evaluation simulation (slide 16), a committee member asked why DTAC was looking at BILOG and MULTILOG when newer, more popularly used tools (e.g., flexMIRT) are available. The presenter agreed that MULTILOG is "ancient" but said existing analysis scripts were available, which could be used in fixed-theta calibration analyses. BILOG was considered because it is used in ASVAB work. Research indicates there are minor differences between BILOG and other programs like flexMIRT, but the team agrees those differences are not compelling, and it is agnostic on the matter. When a committee member reiterated the importance of looking ahead, a DTAC representative said, if BILOG were to disappear tomorrow, DTAC would still have a plan. That is, DTAC knows the software is old, but research shows it is not that different from other packages, and they are already paying for BILOG.

A committee member asked about the severity of observed multidimensionality and whether assessments had been conducted with recent data. The presenter said they do not find multidimensionality in recent data because the items are administered to different people; otherwise, it is high (i.e., .75), but not extreme, and they want to determine if it matters. The committee member mentioned that the typical limit is .80 and asked if it is worth using a complex method for anything lower. The presenter said the method they are using does not require a complex study, which would include collecting data just to find out it does not matter.

When the presenter briefed the conclusion of the MSLP investigation (slide 31), a committee member said s/he recognized DTAC is using a straight-forward composite because it will be used operationally; but s/he asked whether multiple weights could be used if that constraint were not in place. The presenter said they want the composite to be equally weighted, so they determined to use that approach to see if it would work. The committee member said this was excellent work.

At the end of the briefing, the presenter posed three questions to the DAC-MPT (slide 64). On Question 1, a committee member suggested that the name of the procedure should no longer be Stocking-Lord because anchor items are not used, though s/he said this was an improvement. A committee member then asked about the use of a single calibration. The presenter replied that, because AS was already treated as two separate subtests, it would be more problematic to recombine them. The presenter explained that the information they have is largely dedicated to the CAT program.

On Question 2, concurrence with recommended lengths of PC and AR, a committee member said the recommended lengths make perfect sense, although shorter tests put more pressure on the items to be particularly informative. S/he said DTAC's item development processes are good, but they must be sustainable for new form development. The presenter was unsure how many forms would be needed in the future, saying they examine how different items in new forms would be balanced with psychometric objectives. The Assistant Director (AP) said that the P&P forms are relied on mainly by the CEP and less for enlistment. The Assistant Director (AP) said the P&P forms need to be available within ETP for continuation of operations if the CAT forms

become unavailable. The Assistant Director also mentioned that the use of the P&P forms in the CEP are experiencing decreased usage as *i*CAT testing has increased to 50% or more of total testing.

A committee member requested clarification on the time pressures experienced on the P&P versus CAT forms. The presenter explained that in P&P testing, people are given a specific amount of time to complete the test and the time limit must but be appropriate for a group setting, such that people have enough time to complete the test without making those who finish quickly sit idly for an excessive amount of time. Examinees must decide how they are going to use that time, such as deciding how much time they can spend per item. In CAT, however, the time limits are more generous due to the individually-paced nature of the test, and examinees can use the time that is needed to answer the questions as long as they do not use more time than is allowed. The Assistant Director (AP) clarified that, in the CAT-ASVAB, they want to designate a time limit that will allow 99% of people to complete the test. The presenter said they do not want to stray too far from that in P&P ASVAB out of fairness to examinees across test environments. The presenter went on to say that they do not know how the current P&P-ASVAB time limits were set, but that they are trying to ensure they are appropriate. The Assistant Director (AP) said high schools are very attuned to time requirements for testing, and that is why it is so important to give this consideration. A committee member commented that the last five items in the P&P forms are the most difficult. The presenter said the P&P items are administered in increasing difficulty; however, though they are progressive, they are not adaptive. In the case of both P&P and CAT versions of the ASVAB, users may run out of *ability* to respond before running out of *time* to respond. People should be able to answer items in their ability range on P&P forms and, from there, the c-parameter governs the day. A committee member noted that the number of AR items decreased from 30 to 25 but the time remained the same. The presenter said they adjusted item count and time to alleviate speededness concerns for PC, but only the item count needed to be adjusted to achieve the same effect for AR.

A committee member asked if the presentation of only one question per passage in the PC test, in addition to eliminating the issue of local dependence, is also a psychometric improvement. S/he also asked if reducing the number of items on the AR test from 30 to 25 would have implications in an environment that allowed calculators. That is, were any of the removed items calculator-relevant? The presenter said that was a big question and explained that all the items were constructed to be answered without a calculator, but they do not know if the removed items were more or less calculator sensitive than the remaining items. A DTAC representative said the matter would be covered soon, and that there is a laundry list of second- and third-order effects of calculator use, including some that have negative impacts. The Assistant Director (AP) said P&P form development is continuing without regard for possible calculator use, but the committee will hear more on DTAC's recommendations on how to proceed in terms of calculators. The presenter said calculator sensitivity is not really an issue here, regardless of the mix of items affected, because it would be a back-end consideration and fix, given it cannot be addressed at the item level.

Regarding Question 3 requesting concurrence with time limit adjustments, a committee member expressed comfort with the solution, saying research supports it. The result is increasing PC by 5 minutes and decreasing AS by 2 minutes and MC by 3 minutes. The committee member then

asked how else DTAC leveraged measurement. A committee member said s/he did not see any alternative. The committee member then asked if 5 minutes really makes a difference to schools. A HumRRO representative said the MEPCOM representatives who work with students at schools reported that the CAT ASVAB and P&P CEP fit into a 90-minute time block and extending past that would push into a third period, which would be an issue for schools.

5. Form Equating Sampling Design – (Tab I)

A HumRRO representative presented the briefing.

The presenter began by providing an overview of current CAT-ASVAB scale maintenance procedures. The consistency of scaling for newly developed CAT-ASVAB forms is maintained via a two-stage process. IRT rescaling maintains the scale for IRT item parameter and person parameter estimates. After new items are calibrated, their IRT parameters are rescaled to match the scaling of parameters for existing operational items. Standard Score Equating maintains the scale of standard scores (the reporting metric for scores) to ensure they are linked to relevant norms (currently, 1997 Profile of American Youth [PAY97] norms). New forms are administered with a reference form in an equating study to derive linear transformation constants (TCs) for converting IRT theta-metric scores to standard scores. Equating ensures the means and SDs of standard scores for the new forms equal those of the reference form.

Linear equating methods are used to derive TCs to transform IRT-based theta scores ($\hat{\theta}$) on new forms to match the scale of the reference form in a phased approach. This is done for each subtest and for the AS and Verbal Expression (VE) composites. A random groups design is employed. Each applicant is assigned to a single form with equal assignment probability. These include the reference form (administered only during equating studies), an operational form (a form from the previous set of CAT-ASVAB forms), and a new form. New forms initially inherit the TCs from the reference form and these are progressively adjusted over three phases as their sample sizes increase. The final sample size goal is 10K per form. TCs for the reference form and operational form do *not* undergo adjustment during this process. The objective is to arrive at a final set of TCs for each new form that will produce standard score distributions with the same mean and SD as the reference form.

A set of pre-established reference form TCs exists for each standard score consisting of intercept and slope coefficients. One slope is used to determine standard scores for individual subtests, and two slopes for composites (AS and VE). These serve as the starting point for establishing the new forms' TCs. When new forms are administered during equating, distributions of theta estimates for the new forms and the reference form are collected. These distributions inform adjustments to the reference form's TCs to fit the new forms. The presenter then presented the formulas used for this process and noted that it is identical to the process one would use to adjust regression coefficients to account for a change to the scaling of predictors/features used in a model. The process for AS and VE is similar but also accounts for contributing subtest scores' covariance. The presenter then presented a table showing the sample size targets per form and data collection phase (total N = 70,000). A second series of graphs showed the unequated and equated qualification rate differences for CAT-ASVAB forms 11-15 compared to the reference form for AFQT and various Service composite scores from the equating study for forms 11-15.

The presenter continued by posing several research questions.

1. Would the use of unequated standard scores from new CAT-ASVAB forms result in biased scores relative to the scores examinees would get if they took the reference form? The equating briefing presented at the June 2024 meeting of the DAC-MPT already showed that equated scores are not biased.
2. Could the sample size for an equating study be reduced from 10k per form to a smaller sample size target while achieving functionally equivalent equating results?

3. Could the current equating design be updated to change the allocation of the sample across phases, the use of pooled vs. form-level equating analyses in early phases, or the use of three phases vs. two phases?

After their June 2024 meeting, the DAC-MPT requested follow-up analyses to address the question of whether the use of unequated standard scores from new CAT-ASVAB forms result in biased scores relative to the scores examinees would get if they took the reference form. A simulation pipeline infrastructure was used, as described in that meeting. Simulations were run for 9 out of 10 CAT-ASVAB subtests, with AO omitted due to ongoing research evaluating the dimensionality of that test. That briefing covered the results of the entire process: (1) construct reference form Y, (2) construct target forms A-E, (3) simulate equating study with forms A-E and Y, (4) simulate evaluation sample, (5) compute standard scores for evaluation sample, and (6) compute composite scores for evaluation sample. This presentation focuses on the results of a reduced process where Step 3 is omitted and the reference form's TCs are used to compute all standard scores.

Conditional bias analyses were performed in two ways. First using true-score z scores rounded to 1 decimal place. The results are detailed but estimates at the tails of the ability distribution are impacted by large amount of sampling error. The second method employes true-score deciles. This is less detailed but allows for much more stable estimates of average bias across segments of the ability continuum due to equalized sample sizes across deciles. Each combination of composite by form by replication by true score was evaluated. The scores evaluated in the bias analyses were centered and scaled using the mean and SD of the true scores (generating thetas converted to composite scores using generated TCs). The presenter then showed a series of plots depicting the mean bias effects across forms and replications. They suggested that bypassing equating and computing standard scores using the reference form's TCs introduces bias into the composite scores. In the simulation, lower scores tended to be overestimated and higher scores tended to be underestimated. This bias results in qualification rate differences. Performing equating nullifies the biases observed in unequated scores. Equated scores are not biased at any point along the ability continuum. Equated scores produce qualification rates that are aligned with the reference form's qualification rate. The key conclusions are that equating serves its intended purpose without biasing scores and is a remedy for biases that could occur in unequated score distributions.

Next, The presenter discussed the evaluation of the sample sizes needed per form. Data from the equating study for CAT-ASVAB Forms 11–15 were reanalyzed using different specifications: (a) form-level sample sizes varied from 500 to 10k in increments of 500. Samples were formed by selecting the first *N* records for each form in the order they were collected. In a corresponding set of 100 bootstrapped analyses per sample size, equating analyses were based on the first *N* records for each form in the order they appeared in each bootstrapped sample. For each equating analysis, TCs were estimated based on form-specific equating solutions and pooled equating solutions with all five forms equated together. Form-specific equating solutions are the focus of the sample size evaluations. Pooled equating solutions were developed to support evaluations involving the number and allocation of equating phases. The presenter then showed a series of graphs depicting the convergence of the TCs and the qualification rate differences within form convergence. A holdout sample was prepared containing 10K records per form for each of the four new forms that have been operationally administered since being equated. A series of graphs displayed the qualification rate differences relative to the condition with 10 K per form equating across all composites and forms using 5, 6, 7, 8, and 9K cases. The results suggest that a target sample size of 6K examinees per form appears to achieve functional convergence with analyses based on 10K examinees per form. Solutions based on as few as 5K examinees per form were quite stable, but using 6K per form allowed the solutions to stabilize even more.

Having identified a recommended form-level target sample size for forms' final equating analyses, the presenter next discussed evaluations of how other aspects of the equating study design might be altered to (a) streamline the administration of the study, and (b) reduce differences between scores recorded for examinees who test during an equating study and the scores they would have received if the final equated TCs could be used to recompute their standard scores. The design factors considered in these evaluations have no additional impact on the final TCs estimated for each form beyond the reduction of the total form-level sample size. For the evaluations, each sample was constructed by selecting examinees from the

equating data set from CAT-ASVAB Forms 11–15 in the order their results were recorded. A series of four sequential evaluations were carried out to identify a recommended configuration for future equating studies: (a) using a final form-level sample size of 10k vs. 6k (rehash of sample size evaluation), (b) using pooled equating vs. form-specific equating in early phases, (c) using existing early-phase sample sizes vs. increasing them, and (d) using a three-phase design vs. a two-phase design. The recommended design feature from each evaluation was carried forward in subsequent evaluations. The primary basis for making these evaluations is their impact on the qualification rate differences (and the SDs of differences across forms) between (a) the equated scores examinees would have earned if the final TCs could be applied retroactively and (b) the operational scores examinees would have earned at the time they tested, as determined using the TCs specified by the design features in the evaluation. To enhance the realism of these evaluations, a form-level sample size lag of 500 examinees was included between equating phases. This accounts for the additional testing that occurs while temporary equating solutions are being computed, replicated, implemented, and released. For instance, although the current Phase 1 *N* is 500 per form, the processing lag in the analyses means 500 additional people take each form before the provisional TCs can be replaced with temporary, equated TCs. The additional testing volume that accumulates while the TCs are being updated represents an additional group of people who are not benefiting from the gradual updates made to the TCs during the study period.

The presenter continued by showing a series of graphs displaying qualification rate differences for reported scores compared to scores based on final equating constants (a) under the current design, and (b) using a final form-level N of 10K vs 6K. The presenter indicated that using 6K examinees per form allows for a substantial reduction in the duration of the equating study while having a minimal impact on the overall quality of examinees' scores.

The second evaluation examined the use of pooled versus separate equating in early phases of the study in terms of the qualification rate differences across forms. These were also displayed in a series of charts. The conclusion was that using form-specific equating analyses in Phase 2 improves the overall quality of reported scores by reducing the variability in quality across forms during Phase 3. The third evaluation examined the effect on overall qualification rate differences using existing early-phase *N*s versus increased *N*s. The results suggested that current procedures are optimal, with the sample size targets effective at mitigating the impact of provisional TCs on the quality of reported scores. The fourth evaluation examined the effect on overall qualification rate differences across forms using a three-phase versus a two-phase design. The results suggested a three-phase design is superior because it allows an additional opportunity to refine the temporary TCs, which improves the quality of the reported scores.

The presenter concluded by summarizing the recommended alterations to the CAT-ASVAB equating design. Future CAT-ASVAB equating studies should continue using a three-phase design with a target of 500 examinees per form in Phase 1 and estimating temporary TCs using a pooled equating analysis across forms. Phase 2 should continue to include 1,500 examinees per form, with temporary TCs estimated using a separate equating analysis per form. Phase 3 should have a target of 6,000 examinees per form, with final TCs estimated using separate equating analysis per form. This design will reduce the duration and number of examinees involved in equating studies, while converging well with the results of a 10K per-form equating solution and improving the quality of scores reported during Phase 3. The presenter asked for committee input on the change to 6,000 examinees per form in the phase 3 data collection.

At the end of the briefing, a committee member asked if the scores of those who took the new forms are used for the record (i.e., operational). The presenter said they are and that was why Phase 1 had such a small sample size – to minimize impact. A committee member then asked if, after the transformation is applied, a person has the same likelihood of qualifying as they would have had if they had taken the referent test. Referring to slide 9, the presenter said there were differences. If a person takes a new form in an early phase, it does impact that person's likelihood of qualifying; this result is the reason for conducting equating studies. A committee member said the differences appear to be very small, but there may be somewhere along the scale where a small difference matters, that is, at the cut points. S/he asked if the Services have

an obligation to explain to applicants how their scores may vary from those of other applicants. The presenter said two factors weigh against explaining the situation: (1) applicants do not know they are taking a new test and (2) DTAC does not know how their scores may have been different if they had taken an existing form. It is just a function of how the testing system operates.

A committee member asked if degraded performance for pooled equating is a function of the non-invariance issue, saying one could equate forms that are completely unrelated. S/he then asked why pooled equating is less effective. The presenter said it is the failure of the IRT invariance assumption; although one can build forms to be parallel, they are still slightly more or less informative, at a minimum having different SDs, and that will flow down through the equating process. The presenter noted that differences between the pooled and form-specific equating solutions could reveal a lack of invariance, and that those differences are why they equate the forms separately at the end of the study. A committee member commented that the rate of assignment is important and then said the committee appreciated the work.

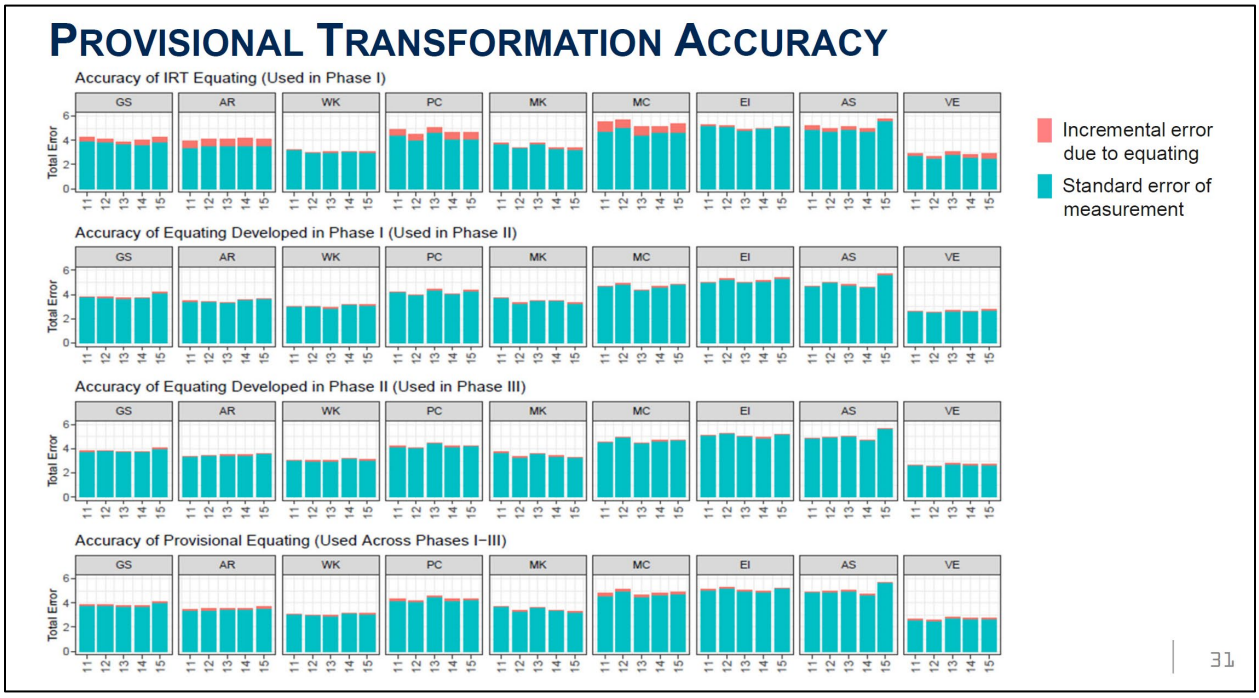
Continuing the discussion, the Assistant Director (AP) said applicants have the opportunity to retest. A committee member asked if there are ethical implications, given that there are no alternatives. Is there an obligation to explain after the fact? The presenter said the testing program is unusual in that it includes this process at all; other CAT programs do not. The presenter speculated that not including the process may be an even greater ethical problem than including it.

A DTAC representative made a similar point. First, this process has always been characterized as an insurance policy. Most of the work is done in the IRT transformation step. Equating error is evaluated in terms of the difference between the provisional score and what it would have been. The amount of that equating error, compared to measurement error, is very small, especially as seen when looking at the axes on plots.

A committee member thanked DTAC for revisiting the topic and said the results appeared to be firm and informative. S/he said the process appears to meet the goal of producing higher levels of standardized testing: higher comparability among forms and to a reference form. This is more than just adjusting test difficulty. The presenter said, specific to the mission at hand, all forms must be as consistent as possible to maintain qualification rates and eliminate variability in the Services' decision-making processes. A committee member said the differences between forms appear to be effectively removed by this process; therefore, any issues with measurement invariance are not a problem. The committee member then said the criterion in the simulation study may be whether each examinee recovered their true performance; however, comparability is the focus.

Regarding the committee's concern about the impact of using provisional scores for enlistment, The presenter showed slide 16, Evaluation of Composite Score Bias by True-Score Decile, and commented on the relatively small degree of difference between the unequated and equated forms and the true scores. A DTAC representative then referred to the CAT-ASVAB Pools 11-15 Equating briefing, which was presented at the DAC-MPT August 2023 meeting. These slides answered the question, related to difference in scores based on provisional constants from what

they would have been if based on final constants? The investigation (a) rescored all applicants who took Forms 11-15 using final TCs, (b) compared rescored values to those used operationally based on provisional constants, (c) calculated total errors as the sum of equating errors and measurement errors, and (d) compared total error with standard errors of measurement. The chart on slide 31 from the 2023 briefing (shown below) illustrates the relatively small incremental error due to equating compared to the standard error of measurement.



6. <u>Complex Reasoning Update</u> – (Tab J)

A HumRRO representative presented the briefing.

The presenter began by defining CR as non-verbal reasoning characterized by the ability to analyze visual information and solve problems using visual reasoning. Complex (non-verbal) reasoning is one element of fluid intelligence, which has been found to be a strong predictor of training and job success. The 2006 ASVAB Review Panel suggested that DoD consider adding a test of fluid intelligence to better balance the ASVAB's composition (between fluid and crystalized intelligence). The potential benefits include better prediction of training and job success, lower susceptibility to compromise, and increased qualification rates. The presenter then showed an example of a transformation item, which included various item features (i.e., types/orientation/size of shapes, number of shapes, line weighting of shapes) and directions of transformations (i.e., vertical, horizontal, diagonal).

The presenter continued by indicating that CR was launched on the ASVAB platform on August 13, 2024. There are four static forms and the 24 items constituting each form are administered in a specified presentation order. It became available to applicants on September 16, 2024, and a total of 9,837 took the assessment between September 24 and November 4 2024. A chart showed descriptive statistics based on those data, as well as correlations with the AFQT and other ASVAB subtests. These ranged between .26 with AS and .56 with AFQT and AO. The presenter then summarized the lines of effort required to complete the work, which include (a) designing CR items and piloting procedures, (b) piloting new items and assembling CAT pools, (c) recommending refinement procedures, (d) evaluating CR and CompT scores, and (e) documenting CR and CompT.

The presenter continued by discussing the third CR pilot test, which is taking place in four waves. The objective of the first wave is to determine whether non-progressive item order impacts item functioning and test performance. These findings will influence the feasibility of a CAT CR. The design calls for 5 static forms of 24 CR items to be administered along with a pre- and post-test questionnaire. It includes two CR attention-check items to determine level of effort. The sample includes non-military participants representative of military applicants (e.g., ages 18-35, U.S. citizens, high school diploma/General Educational Diploma (GED)/< 1 year of college). The target is 5,250 participants, or around 1,050 per form. CR is being administered on the Qualtrics platform, with participants randomly assigned to one form. There is a 35-minute time limit, time to complete is recorded, and a desktop or laptop must be used. The presenter then displayed a chart showing the data collection figures to date. A total of 502 individuals have participated.

The objective of waves 2-4 is to pilot test 288 CR items for potential inclusion on the ASVAB platform and to evaluate, calibrate, and link new CR items to the new base IRT scale (estimated with operational CR data). Each examinee will receive 24 CR items, with multiple static forms with overalpping items. There will also be a pre- and post-test questionnaire and two CR attention-check items. The sampling frame will be the same as the first wave, with a target of 5,250, or around 525 participants per form and 1,050 responses per item. The method will mirror that used in the first wave (e.g., administered on Qualtrics platform).

The challenge will be to determine how to calibrate and link new CR items to the base scale from operational data on applicants. A simulation study was conducted (100 replications) to evaluate the three data collection designs and the four calibration designs to determine which resulted in the best psychometric solution. The data collection options included:

- Operational and randomly selected new seed items, which represents the gold standard. This would be a comparison group only, and the option is not being considered.
- The fully-crossed option would include every combination of evens and odds of new item sets with operational data.
- The daisy chain option would include chained combinations of even and odd new item sets with operational data.
- The random groups option would randomly assign one of five intact item sets (operational or one of four new item sets).

Scaling options include BILOG scales programs, True-Scaled Params, Fixed OP Params, Fixed OP Params (rescaled), Latent Mu-Sigma scaled, and Stocking-Lord equated. Based on the simulations, the Daisy-Chain design, with 10 combinationis of even-odd item sets across the operational and four experimental item sets is the recommended approach. All designs performed very similarly on psychometric metrics. This option allows for common items and guards against deviations from randomly equivalent groups and is less intensive compared to the fully crossed design.

The next steps include collecting sufficient data at military enlistment processing stations (MEPS) from military applicants on operational CR forms (4 versions, same 24 items). MEPS military applicant sample and CR form are used to establish the new IRT base scale. This has been completed. Next, calibrate the operational CR form and derive a new base scale using operational data on the MEPS military applicant sample. This has also been completed. A total of 288 new CR items (96 items per wave) will be piloted using the daisy-chain design with a non-military sample. Finally, calibrate the 288 new items using the data collected in step 3 and link it to the base scale developed in step 2, with the scaling approach to be decided (e.g., fixing parameters to the operational MEPS sample, scaling to latent mu-sigma of operational MEPS sample, Stocking-Lord equating.

The presenter concluded by asking the DAC-MPT if they have any feedback on the daisy-chain design and plan for scaling and linking new CR itesm to the new base scale in waves 2-4. The presenter also sought input on other analyses that should be considered for evaluating the feasibility of an adaptive version of CR from the wave 1 data. Finally, the presenter asked for any other thoughts concerning creating an adaptive version of CR.

As the presenter provided an update of the development effort (slide 5), a committee member asked if any of the descriptives or correlations with ASVAB and special tests were surprising. The presenter said the results were fairly straight forward, noting that AO and the AFQT have the highest correlations with CR.

To close the briefing, the presenter presented questions for the DAC-MPT. The first question asked for feedback on the Daisy-Chain design and plan for scaling and linking new items to the new base scale in Waves 2-4. A committee member asked if the design accounts for item difficulty and provides information on the advantages of moving from easy to more difficult items. The presenter explained that this is why they used the even-odd approach, to allow for balancing out difficulty effects by fully interweaving items.

In response to a question on creating an adaptive version of CR, a committee member suggested it would allow for experimentation on the possibility of creating a shorter, more reliable test.

Regarding the pairing of even and odd items, a committee member suggested that there might be some accumulation of errors based on the distribution. A HumRRO representative noted how the distribution "loops around" in a circular design with the last form having half items from the Operational Form and odd items from Form D. The presenter explained that this was not evident in the table.

7. Computational Thinking Update – (Tab K)

A HumRRO representative presented the briefing.

The presenter began the briefing by displaying FY 2021 William M. Thornberry FY21 NDAA requirements for the special purpose computational thinking (CompT) test to be developed as an adjunct to the ASVAB. The test must assess six domains, including problem decomposition, abstraction, pattern recognition, analytic ability, identifying variables involved in data representation, and creating algorithms and solution expressions. The test was required to be available for operational use by October 1, 2024 (as amended by NDAA for FY22). The presenter then displayed a table listing each construct domain and its definition.

Existing measures of computational thinking assess some of the six domains but are typically used within the K-12 classroom environment. Some measures have been developed for job selection; however, they require specific programming language skills. The timeline specified in the NDAA did not support creating a new, valid measure of computational thinking, but the existing ASVAB/special tests and the new CR test were likely assessing all or some of the 6 computational thinking domains.

The presenter continued by providing an overview of the project. The goal of phase 1 was to define a computational thinking score equation by (a) gathering empirical and subject matter expert (SME)-estimated correlations; (b) specifying and analyzing prediction models; (c) generating, evaluating, and finalizing synthetic CompT score equations; and (d) submitting software requirements and specifications. The goal of phase 2 is to verify the validity of the computational thinking scores by: (a) selecting a marker test, (b) developing and implementing a data collection plan at the MEPS, (c) matching shippers' ASVAB and Cyber Test (CT) scores to study data, and (d) conducting analyses and summarizing the results.

The presenter then showed a slide displaying three computational thinking score equations:

1. CompT_AR—2CR + AR
2. CompT_CT—2CR + CT
3. CompT_All—2CR + AR + CT

The scores are a weighted sum of CR, AR, and CT standard scors with X = 50, SD = 10. The AR, CR, and CT standard (T) scores are normed to the PAY97 sample.

The MEPS administered the Qualtrics data collection tool between April 15 and May 20. It included CR, Computational Thinking Assessment for Middle Schoolers (CTA-M), and a background questionnaire. A total of 1,044 shippers completed the instruments. HumRRO delivered the participant IDs from Qualtrics to DTAC on a weekly basis. DTAC used this information to pull ASVAB and CT scores into a de-identified dataset. HumRRO appended this with repsonses on the instruments administered at the MEPS. A total of 922 shippers were matched. Any data that showed lack of motivation were removed (e.g., responses to two CR attention-check items, time spent, careless response patterns). This resulted in 722 cases. The presenter then displayed a chart showing demographic information for the sample and noted that participation was limited to shippers with a pre-enlistement CT score. The Services have different policies regarding administering CT, therefore the distribution across Services is not equal.

The presenter then reiterated the components of the equation-based CompT scores (i.e., AR, CT, and CR) and the formulas displayed earlier. The CTA-M is designed for classroom use with middle school students. It consists of 23 items administered with a 45-minute time limit, including 15 Computational Thinking Test (CTt) items and 8 Bebras items. The items map to two or three of the six construct domains based on consensus judgements by HumRRO team members (i.e., problem decomposition, solving for algorithms, analytic ability). A score was calculated for each shipper on CTA-M (the criterion) and CR (the predictor). Three CompT scores were calculated using the operational equations from phase 1. Additional computations included (a) predictor and criterion descriptive statistics, (b) predictor and criterion reliability estimates, and (c) predictor and criterion subgroup differences. The presenter noted that for AR and CT, the existing estimates of reliability and current estimates of subgroup differences were used. Charts showed (a) the predictor and criterion descriptives, (b) the reliabilities, and (c) the predictor and criterion subgroup differences.

The data analysis involved calculating the zero-order correlations between CTA-M and the three components of the computational thinking score equation (i.e., AR, CT, CR). The results were corrected for range restriction and the results disattenuated for criterion unreliability. Zero-order correlations between CTA-M and the three operational equation-based computational thinking scores developed in phase 1 were also calculated, and the same corrections applied. The emprical validity of non-negative least squares (NNLS) regression equations were estmated using data from the phase 2 validation study. These were also corrected and adjusted for shrinkage. Post-hoc analyses were done to recompute estimates using all nine ASVAB subtests, CT, and CR. The presenter then showed charts displaying the results. Results indicate that all three equation-based scores (CompT_AR, CompT_CT, and CompT-All) are strong predictors of the computational thinking construct, at least as it was operationalized in the phases 2 validity study (i.e., CTA-M). Emprical weights for the score components (AR, CT, CR) derived from the phase 2 validity study did not outperform the operational weights derived from the phase 1 synthetic validity study. Empirical validity estimates using all ASVAB subtests, CT, and CR resulted in relatively small increases in predicting CTA-M scores (delta R = 0.04).

The presenter concluded by noting that CR is available for administration on the CAT platform. When an applicant completes CR, the calculation of CompT scores is triggered. It requires an AR and/or CT score from the last two years, and uses the most recent score when there are multiples. A blank score is submitted if an eligible AR and/or CT score is not found. The CompT score is saved to the applicants CR record. MEPCOM receives all 4 scores: CR as well as 3 CompT scores.

HumRRO is in the process of preparing research designs for CR and CompT that DTAC may consider. Applicant data containing one to three CompT scores is slowly accumulating, which will support additional analyses. This will include demographic information for future subgroup differences research, and occupationial training criteria data. The ASVAB Training Relevance Survey results may be used to identify military occupations with high compuational thinkinig relevance for further research. The presenter asked for input from the DAC-MPT on additional research on fairness and/or validity.

Toward the end of the briefing, a committee member asked if the presenter was comfortable with the CTA-M as the criterion variable. The presenter said the mean and max scores were not particularly high, but that the test has "pac-man-like" problems, some more complex than others. The eight Bebras items are considered more complex, which increased the difficulty of the test without adding another dimension.

At the end of the briefing, the DAC-MPT was asked for suggestions for conducting additional research on fairness issues and/or validity. A committee member said it would be valuable to continue gathering data as the test goes operational; DTAC will want to understand better the time element and how the test performs when implemented. S/he said DTAC should continue to use data to guide decision making even after the test is implemented. The presenter replied that the Services would want evidence of how the test would impact qualification rates, and an occupational composite evaluation tool prototype that HumRRO recently developed and soon will be available to the Services. It has been fed with simulated data and can serve as a sandbox for the Services to explore impact of changes to their composites on validity. The presenter emphasized the importance of having the right data feeding into the tool in order for it to achieve full functionality. DTAC is planning to have a data submission portal for the Services to use, but this is in the early stages of development. A committee member asked if all this was on track with the NDAA. The presenter said the NDAA only required a means to assess the computational thinking construct, and that requirement had been met in mid-September 2024 by offering the CR test and the three corresponding CompT composite scores. The presenter said some Services are considering administering CR to applicants and that they may be accumulating data to investigate outcomes associated with the CR test and/or the CompT composite scores (e.g., impact on qualification rates). The presenter said, based on recent stakeholder workshop results, CR was considered an important test for a future test battery and may, at some point, be added to the AFQT or incorporated into the ASVAB in some other fashion.

The Assistant Director (AP) reiterated that DTAC has met the NDAA requirement by producing a composite that evaluates the six construct dimensions of computational thinking and is now working with the Services to refine their classification composites, how CR will be used for classification, and how it can be used with CompT. The timeline for these activities is more flexible than the initial requirement, which drove the deadline for CR test development, but because they want to provide the Services with a tool for making decisions, they are still pushing hard to continue progress.

8. <u>Calculator Analyses Efforts – Calculator Impact Study</u> – (Tab L)

A HumRRO representative presented the briefing.

> The presenter began by noting that current policy does not allow calculators to be used when taking the ASVAB. Previous research surveyed SMEs across Services about whether Servicemembers are required to apply mathematics knowledge and arithmetic reasoning without having access to a calculator or other tool. Overall, 68% of SMEs indicated that some form of math without a calculator is required in training, and 56% indicated this was true on the job. The concerns expressed about this policy include the fact that other national testing programs (e.g., Scholastic Aptitude Test [SAT], American College Testing Test [ACT], GED) allow calculators on quantitative tests and high schools often allow calculators during instruction and on exams. In addition, exclusion of calculators may result in the perception that the ASVAB testing

program is not keeping up with trends in assessment. Finally, test items requiring manual calculations may result in increased test anxiety as students are not accustomed to performing such calculations without a calculator.

The purpose of the research presented here was to empirically evaluate the impact on examinee test performance and the psychometric properties of the Arithmetic Reasoning (AR) and Math Knowledge (MK) subtests when calculators are allowed. Among the study design considerations were to (a) maximize generalizability to ASVAB applicant population, (b) minimize security risks to existing ASVAB item pools, (c) minimize disruptions to operational testing of applicants, and (d) minimize strain or burden on study participants. Participants were similar to those who take the ASVAB under operational testing conditions, with (relatively) recent operational ASVAB scores. The procedure was designed to be as similar as possible to ASVAB operational testing. The tests were administered in MEPS by Test Administrators/Test Control Officers along with a post-test survey to obtain contextual information about participants, their motivation, and their calculator usage. Shippers completed the study during a waiting period on their ship day. In all, 3,042 participants met all screening criteria (sufficient effort and motivation) and 2,870 participants met all screening criteria and were unequivocally matched to their official ASVAB administration. All participants completed the same 30-item AR form and 25-item MK form. There were two conditions: calculator provided/calculator not provided. To avoid intermingling or "cross-condition" exposure, all participants on a given day were assigned to the same condition: Odd days (11^{th}, 19^{th}, 25^{th} of month) = calculator not provided; Even days (12^{th}, 20^{th}, 30^{th} of month) = calculator provided.

The presenter then turned to a discussion of the first research question which focused on whether calculator availability has a meaningful impact on the dimensionality of the AR and MK subtests. This was addressed by doing parallel analysis, examining bifactor models, conducting multiple groups confirmatory factor analysis (CFA), assessing differential functioning of items and tests, and computing correlations with other subtest scores. The parallel analysis results indicate similar AR and MK dimensionality in the No Calculator and Calculator conditions. The Bifactor model analysis results also support this finding. The configural factorial invariance resulting from the CFA indicated that all items loaded on a single dimension across groups. Metric invariance was fully supported for MK and partially supported for AR (after removing the equivalence constraints for a subset of items that also demonstrated non-compensatory differential function values). Scalar invariance was partially supported for AR and MK. Differential functioning of items and tests (DFIT) for invariance of item parameters across conditions. Overall, participants in the calculator condition were more likely to answer 13 AR items and 2 MK items correctly. Differential test functioning (DTF) was significant for AR but not MK. The pattern and magnitude of AR and MK correlations with other subtest scores were similar for official ASVAB scores for both groups.

The second research question focused on whether psychometric properties differ based on calculator availability. Test-level analyses included mean score and reliability comparisons and examining DTF between conditions. At the item level, DIF and differences in item statistics across conditions were evaluated. The presenter then showed a series of tables and graphs summarizing the results. Calculator availability resulted in modest increases in average AR scores but had little effect on MK scores. Allowing calculators had no notable impact on subtest reliability. Comparisons of study and official ASVAB scores indicated that AR scores were higher for study participants who used a calculator, but there was no significant difference in MK scores. Overall, the results suggest that calculators make some AR items easier but have little impact on the difficulty of MK items. The effects of calculators on scores and item difficulty parameters are primarily linear, and the conditions could be linked through linear rescaling procedures applied to either scores or item parameters to maintain the interpretability of standard and composite scores. This finding is likely limited to the individually equated, fixed linear forms used in this study (i.e., it is not likely to generalize to all P&P-ASVAB forms, nor to CAT-ASVAB forms). Even though the mean effects of calculators on item parameters were nullified via IRT equating, there was considerable variance in the differences in AR items' equated *b* parameters between conditions, and a few items had outlier *a* parameters in the Calculator condition. There was less variance in MK items' parameter differences between conditions, but DIF analyses showed that a small proportion of MK items are likely to be calculator sensitive. This variance in equated item parameters means that a CAT assessment based on equated parameters might encounter inefficiencies due to items' actual parameters differing from the

equated parameter estimates. Equating would be an essential component of introducing calculators to operational ASVAB testing (to maintain continuity of scores), resulting in no systematic advantage gained by examinees from using calculators.

Research question 3 focused on whether calculator availability had an impact on subgroup performance differences. This involved examining mean score differences across subgroups, assessing adverse impact by condition, and conducting within condition DIF analyses. The presenter then showed tables and graphs illustrating the results. They indicated that the magnitude of effect sizes between conditions was consistent across subgroups and allowing calculators does not appear to alter the potential for adverse impact. Significant differences in DIF between conditions were uncommon across subgroup contrasts.

A fourth research question was whether calculator availability had an impact on the amount of time needed to complete each math subtest. The results indicated that calculators do not appear to differentially impact time spent by demographic subgroups. All subgroups completed AR more quickly when a calculator was available, and the magnitude of the time spent difference was similar across subgroups. The impact of calculator availability on MK time spent was trivial-to-small for all subgroups.

Finally, there were trivial to small differences between the conditions on some of the post-test questions. Participants in the calculator condition reported feeling slightly more motivated and slightly less anxious than those in the no calculator condition.

The presenter summarized the findings. There is no discernible impact of allowing calculators on the factor structure or dimensionality of AR and MK. Parallel analysis, bifactor CFA analysis, and correlation analysis indicate no meaningful dimensionality differences between conditions for AR and MK. DTF results indicate some AR items are easier in the Calculator condition. Allowing calculators had no notable impact on item discrimination and subtest reliability. Some AR items were easier for participants in the Calculator condition than in the No Calculator condition; overall, AR scores were higher in the Calculator condition. Differences in AR TCCs between conditions were minimal after using linear rescaling to account for the impact of calculators (overall impact of calculators on IRT parameters is primarily linear). Scores of examinees who test with a calculator can be linked to the score scale of examinees who test without a calculator with a high degree of accuracy using linear transformations. However, this finding is likely limited to the specific, fixed linear forms used in this study, and may not generalize to all P&P-ASVAB and CAT-ASVAB forms.

MK items tended not to be impacted by allowing calculators; overall, MK scores were not significantly different between conditions. The impact of allowing calculators was similar across demographic subgroups. Mean differences between the No Calculator and Calculator conditions were comparable across subgroups for both subtests. Where there were apparent differences across subgroups in potential performance gains in the Calculator condition, the subgroup sample sizes were small (meaning that sampling error cannot be ruled out as an explanation for the pattern of results observed). All subgroups completed AR more quickly when a calculator was available; this difference was statistically significant for all subgroups except non-English proficient participants. The numbers of non-English proficient participants were small for both the No Calculator and Calculator conditions, so this finding should be interpreted with caution. There were no significant mean differences in testing times between conditions for MK.

The presenter continued by pointing out some of the limitations of the study. It included only 30 AR and 25 MK items, which is a very small subset of the total inventory of AR and MK items (approximately 10,000). It is possible the impact of calculators on other fixed-length, linear forms composed of different subsets of AR and MK items could be stronger or weaker than the current results. In addition, other subtests that could be affected by calculator use, such as Mechanical Comprehension (MC) and Electronics Information (EI), were not included. Use of a fixed-length, linear form limits the ability to infer impact in CAT-ASVAB administrations, or even on other fixed-length, linear forms that may include a different mix of calculator-sensitive items. It seems reasonable to assume there will be a range across examinees in the number of calculator-sensitive items administered (i.e., some examinees might see significantly more calculator-sensitive items than other examinees). If calculators are permitted on the ASVAB, it will be important to

account for the variability in calculator sensitivity across items to minimize the possibility that any given applicant could be advantaged or disadvantaged based on the number of calculator-sensitive items received. It would be inappropriate to apply a single scaling constant to all applicants provided with a calculator if some applicants receive fewer calculator-sensitive items than others. All AR and MK item parameters, regardless of P&P or CAT format, would need to be rescaled based on a linkage of parameter estimates derived from larger samples of both examinees and items. This rescaling would involve a universal scale transformation for item parameters on all forms, such that all item parameters for a given subtest would be adjusted via the same linear transformation, not form-specific transformations. The P&P-ASVAB and CAT-ASVAB could be impacted by this universal rescaling in different ways. P&P-ASVAB forms, although psychometrically parallel at the time of their design, may contain different numbers of calculator (in)sensitive items. Variation in form-level calculator sensitivity could result in forms producing scores impacted by systematic biases, even after the average effect of calculators is taken into account. Forms with more calculator-sensitive items would produce overestimated scores, while forms with fewer calculator-sensitive items would produce underestimated scores. CAT-ASVAB forms could also be impacted by residual errors in parameter estimates after item parameters are rescaled, as those errors would impact the efficiency with which the CAT algorithm selects items.

An equating study will be necessary to maintain statutorily required AFQT qualification rates. USC, Title 10, Sec 520, mandates a limitation on enlistment of applicants with an AFQT score between 10 and 30. This implies an ability to accurately estimate aptitude—allowing use of calculators on the ASVAB could result in changing the definition of the AFQT scores. Calculator use would affect both the CAT and P&P formats and multiple administration purposes (e.g., Armed Forces Classification Test [AFCT], P*i*CAT, Verification Test [VTest], ETP). It will have implications for the score scale as forms are recycled for different purposes. Between AR and MK, approximately 10,000 items have been developed, calibrated, and scaled under non-calculator conditions. All item parameters will need to be rescaled. A complementary study suggests relying on SME judgments of impact would be insufficient. The linear TCs used to convert theta estimates to standard scores are based on linking form-specific score distributions to the PAY97 norms under no-calculator conditions. These constants will need to be adjusted to account for calculator effects on score distributions. New specifications for item development would be needed to guide item writing for use on future ASVAB administrations if calculators are allowed. A new testing time would need to be determined to account for possible changes in the amount of time needed to complete AR, MK, and the remainder of the ASVAB. Even if equated, many uncertainties persist. Decades of validity evidence is based on ASVAB administered without the use of calculators. There is also a potential concern of accurately assessing the ability of examinees at the high-end of AR achievement. Calculators could create a ceiling effect on AR for higher ability applicants such that the AR subtest may no longer be able to accurately measure/assess the ability of examinees at the high end of the ability distribution. We have or will have only some knowledge (a snapshot based on 30 AR & 25 MK items) of psychometric impacts on *difficulty*, *dimensionality*, *response time*, *fairness*, *norms*, and *composite cut scores.*

Logistical concerns include determining when and how to distribute and collect calculators during ASVAB administrations as well as distributing and maintaining calculators (including for overseas testing). There is also an issue with determining who will provide and maintain calculators for each Service for AFCT administrations. In addition, there are test security concerns associated with monitoring the use of the approved device (including the possibility that individuals might attempt to alter their calculator to use as a recording device). Training and guidance will have to be developed for test administrators, including guidance on enforcement of approved calculators and determining how/if to prevent use of calculators on non-math tests (e.g., MC, EI). Given the parallelism between conditions' equated TCCs, allowing calculators could put some examinees at a disadvantage if they choose not to make full use of the calculators. Choosing not to (consistently) use a calculator could reduce examinees' expected rates of correct responses (but they would be evaluated relative to calculator users). Examinees who prefer not to use a calculator would effectively test under no-calculator conditions but be scored according to calculator-based standards. Scores would be a function of both math ability and individual differences in calculator use.

When the presenter briefed the conclusions for research question 2 (slides 25 and 26), a committee member commented on the difference in the effect of calculators between MK and AR. A HumRRO representative said AR effectively requires more work; it requires a person to set up the situation and extract information, rather than just punching information into a calculator. On the other hand, MK is more straightforward, in that items require knowledge about how to solve problems, in lieu of actually solving the problems.

Regarding the summary of findings presented on slide 39, a committee member proposed that there may be some cases in which linear transformation does not hold; it may be important to look at the results of linear and nonlinear approaches. A HumRRO representative said their main concern about this equating is that the findings may not generalize to other forms. If one equating solution is rolled out, and one universal set of linking constants is set, but it is applied to two forms, there may be issues due to forms having different numbers of calculator-sensitive items. If forms have different numbers (or severities) of calculator-sensitive items, there is not a single set of linking constants that would generalize to all forms. The approach works with a single fixed form, but it may not work for other fixed forms unless they go through form-specific linkage analyses. A committee member asked what makes an item more calculator-sensitive, because adjustments should not be added across the board. A HumRRO representative said they do not yet have a good sense of that, empirically, because there were only 55 items. There are many blueprint categories, and each category could be impacted differently. An example of a vulnerable item is one with a stem that requires computing a square root. The test-taker sees the square root symbol and then push the button that looks like the symbol. This measures only their ability to push a button that looks like a symbol. A HumRRO representative reiterated that some items require you to know what you are doing and others do not. The bottom line is that we do not yet know the features of the items that make them more sensitive. However, we have pause about rolling out universal rescaling because it impacts items differently.

A committee member asked, if the forms are not equivalent, what methods would you use to adjust? The Acting Director (DTAC) said they tried to predict which items would be sensitive using SMEs, but their predictions were not correct. A HumRRO representative said there was a method available, which was not the best idea but perhaps the best option. This method would begin with triaging the provisional scores obtained immediately following the introduction of calculators and by linking those scores to the scores observed in the period immediately prior to the introduction of calculators. This assumes the random equivalence of the examinees who tested during the pre- and post-calculator periods and would rely on analyses similar to those used for CAT-ASVAB equating. The P&P EPT forms have low volumes, so that makes it difficult to apply this linking strategy to those forms. The concern is trying out the thousands of items that have been calibrated and potentially not being able to use any of them if calculators are allowed. The HumRRO representative said addressing the overall situation was not a simple matter; there are many moving parts and no guarantee that we would ever get back to where we are now with the existing tests. The Assistant Director (AP) said the recommendation would be that DTAC cannot equate out the effect of the use of calculators for existing items and that it does not make sense to allow the use of calculators on the current items. The presenter further said rescaling would be necessary to maintain the qualification rates as statutorily required.

9. Calculator Analyses Efforts – CAT Simulation – (Tab M)

A HumRRO representative presented the briefing.

The presenter began by referencing the previous presentation, indicating that it demonstrates what might happen with fixed-length, linear forms if calculators are allowed when taking the ASVAB. What happens with CAT-ASVAB, however, remains an open question. The work in this presentation aimed to evaluate what might happen to CAT-ASVAB composite score distributions after AR and MK item parameters are rescaled to account for the impact of calculators on latent ability distributions. It is based on the assumption that the results from the previous study generalize to CAT-ASVAB. The work employed the simulation pipeline infrastructure described in the June 2024 meeting of the DAC-MPT to evaluate consistency between a reference (i.e., unmodified) condition and different experimental conditions.

The data available are from the impact study with a small sample size of items (i.e., 30 AR and 25 MK) which underrepresent the universe of items. In addition, not all items are expected to have equal calculator sensitivity. MK alone has 40-plus taxonomies and 200-plus identified enemy groups. The impact study evaluated fixed-length, linear forms, which are constructed differently than CAT forms. CAT, by definition, adaptively selects items from the form and has explicit content balancing for only two subtests (AO, GS). Due to the "greedy" selection algorithm, discrimination plays a larger role than content area in item selection. This study evaluates what *might* happen after a formal linking study is completed to rescale existing CAT-ASVAB AR and MK item parameters onto a metric that is compatible with calculators *if that study's findings converge with the Impact Study*

Because of the characteristics of the Impact Study data, instead of focusing on a single condition, a range of counterfactuals was evaluated, each of which answers what can be expected would happen if different types of error were introduced. To generalize from the available data, a 3D Gaussian copula was fit to the Impact Study's item parameter data and sampled values from the copula. *a* and *c* parameters were converted to the normal metric for 1) the Impact Study data and 2) the generating parameters used in the simulation pipeline. We fit the copula to residuals between without-calculator parameters and equated with-calculator parameters from the Impact Study, added these residuals to the transformed generating parameters, transformed the altered *a* and *c* parameters back to their natural metrics, and estimated new composites for the holdout sample from the simulation pipeline infrastructure. Several conditions modify the *b* parameters deflections to address plausible scenarios for how the universe of items may differ from the sample in terms of calculator sensitivity. The research questions addressed were (a) how do empirically informed, copula-based deflections to item parameter estimates affect composite score distributions for CAT ASVAB? (b) how do biased difficulty parameter deflections affect composite score distributions from CAT-ASVAB? and (c) if effects are present, which composites and which ranges of those score distributions are most affected?

The presenter then discussed the various conditions.

- The test condition (Condition 0) consisted of running the final stage of the simulation pipeline from Heinrich-Wallace (2024) to compute composite scores for the holdout sample; 10 replications (700,000 cases per composite per condition) were evaluated. All other conditions are evaluated relative to the test condition. This is conceptually similar to decision consistency (comparing two estimated scores). In this case, decision consistency is preferable to decision accuracy (comparing an estimated and a generating score) because all composites are based on Bayesian modal estimate theta-hats, which are subject to shrinkage.
- Retest (Condition 1) is the same as the Test condition, but with a different random seed.
- Random Error (Condition 2) *a*, *b*, and *c* parameters have copula-based deflections based on the Impact Study data.
- Alternating Tail-Sampled Error (Conditions 3–7) a and c parameter deflections are the same as Condition 2, but the *b* parameter deflections are sampled from the top and bottom 5% of copula-based deflections. Different proportions of items (3/15, 6/15, 9/15, 12/15, and 15/15) have the manipulation while the remaining items have no manipulation. These conditions evaluate

counterfactuals where different proportions of items have higher or lower sensitivity to calculators than the average items included in the Impact Study.

- Alternating Tail-Sampled Error, Moderate (Condition 8) is the same as Condition 7 (15/15) but all *b* parameter deflections are halved. This assesses the same counterfactual as Condition 7 (15/15 items are manipulated), but items varied less in their calculator sensitivity.
- Systematic Error in *b* Parameters (Condition 9) shows the effect of systematic error on composite scores. The largest simulated deflection for *b* parameters is added (which was negative) to the difficulty of each item, indicative of an item that is more calculator sensitive than the average error from the Impact Study sample of items. This emphasizes the importance of equating (which removes systematic error) and represents a proof of concept that the pipeline is working properly. It allows for simulating extreme results.

The presenter then displayed a series of charts presenting the results for AFQT and various Service composites. The presenter concluded that across bias, RMSE, reliability, mean score conditional bias, and qualification rate differences, in all conditions, calculator error introduces the same pattern of effects while the degree of these effects depends on the condition. Low-ability simulees have inflated scores while moderate-to-high ability simulees have deflated scores, with a larger effect for high-ability simulees. For AFQT, there is very little conditional bias at the IIIB cut score (31 on the percentile AFQT scale) across conditions. The degree is linear on proportion of items with manipulation, Random Error most like Alternating Tail Error (6/15). The effect varies across composites and is predicted by the proportion of the composite that is contributed by AR and MK (see slide 16). The most affected composite is Navy: Basic Electricity and Electronics.

There were no questions or comments at the end of the briefing.

10. <u>Calculator Analyses Efforts – Calculator Needs Assessment</u> – (Tab N)

A HumRRO representative presented the briefing.

The presenter began by stating that the purpose of the work being described is to conduct a needs assessment to determine whether a test assessing math content with a calculator is warranted and, if so, use the findings to inform them of what the taxonomy/blueprint would be. A needs assessment survey was administered on the HumRRO platform from June through October 2024. It requested input on the types of math needed in training and on-the-job, and the role of calculators in performing that math. A meeting was held with Manpower Accession Policy Working Group (MAPWG) technical and policy representatives to identify training staff and occupational managers across Services who could respond to the survey. The needs assessment sample was based on the 2022 Training Relevance Survey sample and included training courses and occupations covering a variety of content, including some with intensive math requirements (e.g., Air Force Precision Measurement Equipment Laboratory). In September, additional training courses and occupations were added for greater representation in some job clusters. The responses from each training course or occupation were averaged to weight each equally. The data were clustered into eight areas: Electrical, Infantry/Combat, Information Technology, Intelligence, Logistics and Administration, Mechanical, Medical, and Science/Engineering.

The presenter then displayed charts showing the number of training responses and on-the-job responses by occupational area. A three-point rating scale was employed: (a) 0 a given type of math is not required or is only done with a calculator; (b) 1 the type of math is required with a calculator, but those who enter training or their first job knowing how to do this math do not perform better than those who do not, and; (c) 2 the math is required with a calculator and those who enter training or their first job knowing how to do this math do perform better than those who do not. Additionally, an average rating of less than 1 indicates performing the type math is not need in training/on the job, while greater than 1 but less than 1.5 suggests being able to perform the type of math with a calculator is not a prerequisite for successful performance, and an average rating of greater than 1.5 indicates being able to perform the type of math with a calculator is a prerequisite for success performance in training or on the job.

The presenter continued by displaying tables summarizing the results by AR and MK content areas and occupational clusters. Additional charts included math types that are not included in the AR or MK blueprints. The overall results indicate that there are relatively few types of math where calculator use is a prerequisite for successful performance in training or on the job, and these are generally limited to three occupational clusters: Logistics/Administration, Science/Engineering, and Medical. There are other types of math where calculators are used in training and on the job, but calculator use is generally not a prerequisite for success. The presenter concluded by noting that the sample was purposefully selected to include a range of occupations and math requirements. However, due to limited participation in specific occupational areas and in some Services, the sample is not as robust as planned. Work continues to augment the sample to the degree possible.

The presenter concluded the presentation by asking if the DAC-MPT had recommendations to address any of the implications identified in the calculator impact study. The presenter also asked if there are other complications resulting from calculator error that could affect CAT tests, specifically, that were not addressed in the simulation study? Finally, the presenter asked if, based on the results of the needs assessment, whether the DAC-MPT believes the results support the need for a special purpose test that assesses math ability with a calculator for use in classification.

Regarding interpretation of the needs assessment results (slide 9), a committee member asked how the team handled training and jobs that require heterogeneous types of math; that is, did it specify formulating responses based on the highest form of math? The presenter said they did not provide instructions to that effect, so it was left to interpretation. They knew variance existed in the types of math required, if for no other reason than different types of math are included in MK and AR, which reflect the types of math required by jobs.

The Assistant Director (AP) noted that a rating of 2, as shown on slide 8, indicates that calculators are required and those who use them perform better on the job, however, on slide 9, a rating of greater than 1.5 is sufficient to indicate a calculator is a prerequisite. The presenter said they selected 1.5 as the cutoff because there were no average scores of 2. The Director (AP) asked how the team made the leap to "successful performance" based on the questions that were asked. A HumRRO representative responded that the scaling technique was drawn from prior research and can be described as a branching process. That is, they first answer whether the math is performed with a calculator and, if so, if people already know how to do it, and finally do those who use a calculator perform better than those who do not. If the average response of multiple respondents was 1 or less, that was an indication that there is no advantage to using a calculator. The HumRRO representative said calling the use of calculators a "prerequisite" on slide 9 is using too strong language; the term "beneficial" would have been better. The Director (AP) sought confirmation that a rating of 1.5 or greater meant that those who knew how to do the math with a calculator got more out of the training, and a HumRRO representative said yes. A committee member asked if the team had defined "successful performance." The HumRRO representative said performance is viewed as relative, performing better or worse than others. A MEPCOM representative said success is "GO/NO GO;" a person makes the standard or does not. Another HumRRO representative commented that performance measurement in the Army is oriented effectively on performing job tasks to standard (i.e., do they meet the standard).

When the presenter said the sample was not as robust as had been planned (slide 16), the Assistant Director (AP) said DTAC had made several attempts, through the Army G1, to get more Army SMEs, but the effort has stalled. An Army representative said it is a matter of reaching the people who are interested and the Army will continue to reach out to obtain the requested support.

At the end of the briefing, the presenter asked three questions of the committee. In response to the third question (on the need for a special purpose test that assesses math with a calculator for use in classification), a committee member asked if creating such a test would be useful to applicants who want to enter a specific MOS and if it would help AP accomplish its objectives. The Assistant Director (AP) replied that AP is heading toward recommending the special test as a course of action. If calculators are useful for some jobs, then it would be helpful to be able to identify people who would be a good fit for those jobs. Instead of testing with calculators for all jobs, the testing would be more targeted. The Assistant Director said AP is not yet ready to finalize the course of action, but if there is value in calculator use, the special test is probably the path forward.

The Acting Director (DTAC) said the SME study, which was part of the impact study, suggested that SMEs cannot help much here. There is differential impact across jobs. Additionally, the CAT environment makes introducing calculators even more complicated. Regardless of what DTAC does, there will still be degradation in the precision of scores. The Acting Director (DTAC) said that at this time the needs analysis did not show a lot of need for calculators. In addition, it was difficult to obtain input. The Acting Director then mentioned the calculator impact briefing and the potential impact on maintenance of the item bank, as well as logistical considerations such as managing the devices. In reference to costs and benefits, the Acting Director (DTAC) mentioned benefits such as perception of the test, staying current, and not creating anxiety for test-takers. It is important not to discount those, but the costs are heavy, and if they can handle the optics sufficiently with the special test option, that is the preference. The Acting Director (DTAC) clarified this was a summary of DTAC's perspective but there are other stakeholders with different perspectives.

The Assistant Director (AP) summarized AP's perspective on the possibility of allowing calculators and using equating procedures for current items to retain measurement precision as unlikely, at least at this time. However, it may make sense to develop new items designed to be answered with a calculator for MOS where calculators are more useful. The result would be a special test that could be used for classification into certain MOS. Importantly, this would protect the current testing program and allow it to stay on par with other programs while adding value for relevant MOS.

A committee member asked for clarification on the additional types of math referenced on slide 15. The presenter said these types of math are at a higher level than what is covered in AR and MK and are from a taxonomy generated by Waugh et al. (2015), which includes statistics, for example. The responses from SMEs suggested that these types of math were not needed for entry-level training and occupations.

The presenter mentioned one caveat on the SMEs (training developers, trainers, and career managers): They were asked to respond about entry-level courses and jobs in the needs assessment tasking to the Services. However, the needs assessment directions were not specific about the level of training or job and some SMEs were affiliated with higher-level training or jobs. We tried to weed out input regarding courses and jobs that were not entry level based on the background information provided.

The Director (AP) said they can share this information from the calculator study briefings with leadership as a rationale for not moving forward too quickly. The question will not go away, however, so DTAC should think about what it would do if given the luxury of time.

A committee member suggested continuing to examine the consistency between (a) what is taught in school and how it is assessed and (b) what is required on the job. That would include focusing on the courses taken by students who are more inclined to join the military. A HumRRO representative commented that the high school curriculum study would address this subject to some degree.

The Director (AP) proposed that if they pursued the special test option, the test might experience limited use at first; however, if a strong argument for its utility emerged, it could make its way into the AFQT over time.

The Acting Director (DTAC) commented on the concern about mitigating applicant anxiety over testing without calculators, saying HumRRO found that anxiety with shippers did not appear to be an issue, so perhaps it is not a viable concern with applicants either. The Director (AP) said leadership is concerned about losing applicants, and the shippers are past the point of testing. The Acting Director (DTAC) conceded that point, however, expressed a larger concern. That is, if a person is not able to sit comfortably for an exam because they are anxious for the lack of a calculator, their personality might not be a good match for military service, where they will have many stressful experiences. The Assistant Director (AP) said the point had been made before, and it continues to be a rich topic for discussion, perhaps one the committee could address. That is, this is not the end of the conversation, and AP will continue working through issues and asking the committee's advice.

A committee member said the comment about fit with military service resonated with him. S/he added that the discussion so far had presented all the various angles – the costs and benefits – nicely. S/he agreed that the perceived needs, which may or may not be empirically based, are appreciated, but the DAC-MPT is also interested in the impact on test scores and the likely requirement for equating. The committee member then mentioned drawbacks, such as the need to have calculators that are not dirty and that work – more practical aspects of the requirement. The committee member also commented on the similar pattern of results from the CAT simulations and the results on composite score bias, asking if there might be a connection between the two.

11. Public Comments

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

12. Refinement of the Joint Service TAPAS Instrument – (Tab O)

A HumRRO representative presented the briefing.

The presenter began by explaining that the goal of this work is to develop a TAPAS composite for military compatibility designed to predict alignment with military core values and ability to predict various forms of misconduct. It arises from a DoD directive that applies to enlisted personnel. A second goal is to develop a composite for enlisted selection designed to predict first-term enlisted performance which will potentially expand the qualified applicant pool without compromising valued outcomes. In the end, this will be a Joint-Service (JS) TAPAS instrument. It will be modular and will include a common core of facets that support assessing military compatibility (MC) and enlisted (ENL) composites. The instrument will also allow Service-Specific (SS) facets to support Service-specific use cases. The presenter then displayed graphics illustrating possible configurations of the instrument.

The work is being carried out in a phased approach. Phase 0 involved work designed to address the immediate OSD tasking. It resulted in interim MC and ENL composites. Facets were added to the Air Force and Marine Corps TAPAS needed for scoring of the interim MC composite. This was implemented in September 2024, although the composites were not used for operational decision making. Phase 1 refines the recommendations from the earlier research. Content development and psychometric work is occurring in FY 2025 and will (a) refine the composition and facet weighting of the Phase 1 MC and ENL composites, (b) update the TAPAS statement pools, (c) calibrate TAPAS statement pools with a joint-Service sample, and (d) develop provisional joint-Service norms for the JS and SS facets. The programming required to enable implementation at the MEPS will take place sometime in FY 2027.

Phase 2 will involve the evaluation and refinement of the Phase 1 JS composites for operational decision making. The norms for the JS and SS facets will be updated based on FY 2027 applicant data and subsequent evaluation work. The composition and weights for each Phase 1 composite will be revisited and adjusted as needed. This will establish an evidentiary base for the use of the final Phase 2 composites for enlisted and military compatibility-related screening decisions (e.g., criterion-related validity study for the enlistment composite). The presenter then displayed a chart showing the various phases and timelines.

The presenter then discussed the focal criterion for the MC composite, which include 10 categories of misconduct (e.g., violent behavior, sexual assault, unethical behavior). These were informed by a literature review and expert review. SMEs evaluated the conceptual and empirical evidence of alignment between TAPAS facets and the 10 categories of misconduct and rated the alignment as strong, moderate, or weak. They then reached a consensus on the facet composition weighting for the preliminary Phase 1 MC composite.

The ENL composite was based on performance dimensions from a taxonomy developed previously and captured “overall performance” from Service stakeholders. There were 10 dimensions (e.g., task performance, organizational support, support for peers, physical performance, and safety/security consciousness). Archival and SME data were gathered to support development and validation of the composite. A subset of facets for predicting first-term enlisted performance was identified based on regression models.

The FY 2024 research focused on refining the preliminary Phase 1 JS TAPAS recommendations and identifying needs for FY 2025 development work. Multiple research efforts were conducted to evaluate the TAPAS facets and their statement pools. In addition, there were multiple rounds of discussion with OSD and the Services to arrive at an agreed-upon set of JS facets and a JS instrument design. Finally, plans for recalibration of TAPAS statements with a joint-Service sample were established. Efforts included:

- Retranslation of facet statements
- Bias and sensitivity review of facet statements
- Evaluating susceptibility of facet statements to transient error
- Revisiting the marginal IRT reliability of facet scores
- Evaluating the equivalence of facet scores across TAPAS versions
- Conducting composite shortening analyses

This work provided additional perspectives on the functioning of TAPAS facets beyond what was known when the Phase 1 composites recommendations were made in FY 2023.

The presenter then discussed the retranslation of the facet statements, the goal of which was to evaluate whether they are clear indicators of the intended facets. Natural Language Processing (NLP) methods were used to identify the items most in need of review by SMEs (482 of the 1,200+ statements in the DoD TAPAS statement pool). The focus was on statements that were more semantically similar to those of another facet than the intended facet. These were independently rated by 8 SMEs who indicated which facet each measured. At least 6 of 8 SMEs had to agree that a statement aligned with its intended facet for it to be considered "translated" to that facet. The presenter then displayed a table showing the target facets and the percentage of statements related to each that were assigned to the facet, assigned to a non-target facet, or where the results were equivocal. The presenter noted that the facets varied in the percentage of statements translated by the SMEs to the target facet, with some (e.g., physical conditioning) exhibiting perfect retranslation, while others (e.g., attention to detail) exhibit relatively poor retranslation. The recommendation for FY 2025 is to have humans (a) retranslate the remainder of statements in the pool, (b) move statements to the proper facet as needed and recalibrate, and (c) revise statements so they have a clear translation and recalibrate.

Another activity involved reviewing the TAPAS statements to identify any that may be problematic from a bias or sensitivity perspective. Each statement was reviewed by two external SMEs with expertise in this area, with four SMEs participating. Statements flagged by at least one external SME underwent a second round of review by three internal experts, who indicated whether the statements should be revised or dropped and the reason for doing so. The presenter then showed a slide outlining the bias and sensitivity categories (i.e., unfamiliar term, colloquial, unfamiliar situation, controversial language, discrimination), and their definitions. A table showed the percentage of statements for each facet that were judged fair or identified for revision or elimination. Almost all facets had statements that were flagged for one or more reasons. Most flags were related to the use of unfamiliar/colloquial terms rather than use of controversial or discriminatory language. The recommendation for FY 2025 is to have internal experts (a) review the remainder of the pool, (b) write new statements to replace those that are dropped and recalibrate, and (c) revise statements flagged for revision and recalibrate.

Susceptibility of the facet statements to transient error was evaluated by having eight SMEs independently rate each statement on the following scale:

> *Please rate how much you think applicants' responses to the following statements would be influenced by their psychological/physical state at the time of testing (e.g., based on their mood, how they feel physically, etc.), using a scale of 1 (not at all influenced), 2 (slightly influenced), 3 (moderately influenced), and 4 (very influenced.*

The presenter then showed a table listing the various facets with the percentage of statements for each falling into various mean rating categories (e.g., 1 to 1.5, 1.6 to 2.0). Overall, SMEs viewed responses to TAPAS statements as not very susceptible to transient error as evidenced by low mean ratings. Statements rated as slightly more susceptible were consistent with expectations, given affective elements associated with those facets (e.g., optimism, adjustment, even tempered). The recommendation for FY 2025 is to revisit/revise statements with ratings of 2.0 or higher, if deemed warranted, and recalibrate.

The next task was to provide updated estimates of the marginal IRT reliability of facet scores based on large, current sets of applicant data (or published data when applicant data is not available). This was accomplished using Army, Air Force, and Marine Corps versions in use from 2021-2023 that were current as of February 2024. The analyses are based on applicant records where no more than one TAPAS response check item was incorrect. The slide provided the number of cases for each Service (Army 212,726; Air Force 108,063; Marine Corps 82,794). The presenter then showed a table with the results. The facets exhibited relatively low to middling reliability compared to suggested reliability for high-stakes testing (average estimates = .40 to .76). This suggests not using individual facet scores for decision making, given that composites would be more defensible. The recommendation for FY 2025 is to carefully examine statement pools for low reliability facets during FY 2025 content development (e.g., evidence of

heterogeneity, multiple clear dimensions within a facet) and aim to bolster/refine the statement pool for those facets.

The next step was to begin to evaluate the comparability of facet scores from TAPAS versions that have different facet compositions. This was done by examining the comparability of TAPAS facet intercorrelations (e.g., comparing the same facet A-B correlation across versions). The comparability of TAPAS facet correlations with other composites (e.g., AFQT) was also examined. In all, seven different versions of Army TAPAS used at the MEPS over time that partially overlapped in their facet composition were examined. The presenter continued by outlining several approaches that could be taken to accomplish this work and noting the limitations of each. In the end, given time limitations, a simpler but more limited approach was taken that focused only on the similarity of TAPAS facet intercorrelations and TAPAS facet-other correlations (e.g., AFQT, 6- and 24-month attrition) across versions that differed in their facet composition. TAPAS fact intercorrelations and TAPAS facet-other correlations were generally quite similar across versions, indicating that the facet mix may not have notable impact on a target facet's measurement. When differences were found, they tended to be for TAPAS facet intercorrelations between TAPAS versions from different Army TAPAS development stages (e.g., Stage 2 with least use of cleaning/quality flags and Stage 4 with most use of cleaning/quality flags). The average absolute differences between same-facet correlations across versions was .014 within stages and .054 across stages. Between-stage differences in facet intercorrelations did not translate into differences in TAPAS facet-AFQT or TAPAS facet-attrition correlations.

The last task was to evaluate the possibility of shortening the preliminary Phase 1 MC and ENL composites. This was accomplished by performing best subsets regression using Phase 1 MC and ENL composites as criteria (separate models for each criterion) and the facets that contribute to those composites as initial predictors. Regressions were based on the facet intercorrelation matrices developed during the FY 2023 research. This allowed for the identification of facets that were consistently retained in models as the number of features in the predictor subset was reduced and the Multiple *R* achieved by those reduced models. The results indicate that there appears to be room to shorten the Phase 2 MC and ENL composites and still achieve a very high correlation with the full versions of those composites.

The presenter continued by noting that only a limited number of facets can be administered as part of the JS TAPAS due to testing time constraints at the MEPS and the cognitive load associated with assessing more of them. There is a tradeoff between the number of facets and the number of statements per facet. More facets means more flexibility to cover JS MC and ENL composites and Service-specific uses. However, more facets also means fewer statements per facet given constraints on testing time. This may result in a less reliable instrument. The greater the number of statements per facet, the higher marginal IRT reliability of the facets. The literature suggests 20 statements per. The target is no more than 17 facets for the JS TAPAS instrument, so a key decision point is how many to reserve for the Joint-Service and Service-specific facets.

Key considerations in identifying JS facts included (a) use in and importance to Phase 1 composites, (b) use in and importance to Service-specific composites, (c) performance of the FY 2024 research metrics, and (d) relevance to outcomes considered of broad interest to the Services (e.g., attrition, leadership potential). On the set level, considerations include (a) balance in terms of the personality construct mix, (b) more JS facets or more Service-specific slots, and (c) more facets overall or fewer facets and more statements per. SMEs from HumRRO, Drasgow Consulting Group, and DTAC reviewed the information for each facet in light of these considerations and developed recommendations for potential sets of facets to include in the JS instrument. The goal was to identify a single set of facets that could be used to support scoring of the refined Phase 1 JS MC and ENL composites. It must be decided in upcoming work whether different facets from the set would be used to score each composite or all used for each composite but weighted differently. After reviewing the considerations, research findings, and recommendations with Service representatives, a consensus was reached to include 12 JS facets and reserve 5 slots for Service-specific facets.

The presenter then discussed the next steps for the JS TAPAS. Preparations are underway for the implementation of the Phase 1 composites. The statement pool has been developed and existing statement recalibration and new statement calibration has been performed using a joint-Service sample. The

composition and weighting of the Phase 1 composites are being finalized and the development of provisional joint-Service norms for facets is underway. Programming work to allow for instrument delivery at the MEPS is scheduled for FY 2027. Ongoing research and development (R&D) work includes examining the effects of practice and coaching on TAPAS, reviewing the potential role of artificial intelligence (AI) in bringing efficiencies to non-cognitive assessments (e.g., statement development), and examining the potential for TAPAS and supervised machine learning for predicting attrition.

The presenter concluded by asking for DAC-MPT members' opinions about the acceptable minimum level of reliability for defending use of TAPAS composite scores for making high-stakes selection decisions. The presenter also sought input on the tradeoff between narrowing the construct covered by TAPAS, which should improve reliability, and the difficulty in developing a statement pool of sufficient size. Finally, the presenter asked for suggestions for mitigating coaching effects if it is found that TAPAS is susceptible to coaching.

At the end of the briefing, a committee member thanked the researchers for a very clear presentation and responded to Question 2, on how to deal with the tradeoff between narrowing the constructs measured and producing a sufficient number of items to measure the narrowed constructs. The committee member then recommended focusing on the purpose of the testing, which is selection, and prioritizing predictive validity over reliability. Reliability statistics might not be as useful a tool as standard error of the mean (SEM) and the range of the scores. A committee member then asked about the optimal value in relation to SEM, in adaptive testing, and commented on the stopping rule. Committee members asked, for clarification, how many statements were associated with reliability estimates and how does that compare to the 17 statements that are proposed for the operational test. Additionally, if composite measures are used, will the operational setting constrain you in relation to what type of composites will reach your goals?

A committee member said if it were possible to identify people who had been coached, what could be done about it? The Acting Director (DTAC) said an information variable could be provided, which would tag the score as suspect; the Services would be responsible for deciding whether to use the score or to require a retest. This is one reason questions about test-retest reliability are important. A committee member said, while not precisely the same thing as coaching, a colleague has conducted research on the effects of applicant faking. The article is titled "Effects of applicant faking on forced-choice and Likert scores" (Pavlov, Maydeu-Olivares, & Fairchild, 2019) and can be found here. The presenter referenced the large literature on faking detection, saying research approaches it from the IRT and Classical Test Theory (CTT) perspectives. The presenter was unsure how much work had been done on explicit coaching as opposed to simple misrepresentation.

13. <u>Adverse Impact</u> – (Tab P)

A HumRRO representative presented the briefing.

The presenter began by defining adverse impact (AI) as the unintended discrimination of a protected class that is the result of a selection procedure. It is not a property of a test, however AI may occur when a test's scores are used as the basis for selection. A selection test may potentially demonstrate AI when it shows sizeable mean test score differences between a majority group and a protected class (minority). Effect sizes of the standard mean difference give us an index to examine a test's potential AI. Adverse impact does not mean a test is biased. The presenter continued by citing several sources supporting the validity and fairness of the ASVAB. The presenter then addressed how adverse impact is assessed. The four-fifths rule is

frequently used. It states that "a selection rate for any race, sex, or ethnic group, which is less than four-fifths (80%) of the rate for the group with the highest rate, will generally be regarded by the Federal enforcement agencies as evidence of adverse impact." The presenter continued by showing the formula for the impact ratio (IR) used to compare selection rates. Additional formulas were presented showing how statistical significance of IR and its confidence intervals are computed.

The four-fifths rule and accompanying statistics are applied to the AFQT by comparing qualification rates across the focal and reference groups, including qualification for entry into the military (i.e., those scoring in AFQT Category IIB or higher), and qualifying for enlistment incentives (i.e., those scoring in AFQT Category IIA or higher). AI is assessed using initial test scores only. The presenter noted that significance testing is not necessarily useful in analyses with very large numbers of applicants (i.e., > 2,000). The presenter continued by explaining that effect sizes (ES), or standard mean differences (commonly Cohen's *d*) can be plotted and classified with respect to Cohen's standards of evaluation (i.e., small $\leq$ 0.20, moderate $\geq$ 0.50, large $\geq$ 0.80). The presenter then showed the formula for computing 95% confidence intervals around the effect sizes. These provide a boundary around an ES point estimate, with small boundaries indicating a more precise estimate.

The presenter next showed a chart showing the ASVAB and special tests under consideration. The latter include the CT and Coding Speed (CS). CS is only used by the Navy. The analysis sample included FY 2023 applicants. Comparison groups were (a) males/females, (b) non-Hispanic Whites/Hispanic Whites, (c) Non-Hispanic Whites/Non-Hispanic Blacks, (d) Non-Hispanic Whites/Non-Hispanic Asians, and (e) Non-Hispanic Whites/Non-White Hispanics. In all cases the first group cited constitutes the reference group and the second the focal group. All groups represent more than 2% of the applicant population. Data were cleaned to include only initial test records with a valid score, name, and social security number. Duplicates were removed for ASVAB only. Records with response times greater than 2.5 SDs below the mean were removed for ASVAB and CT, as were those with less than 2 minutes to complete CS. Finally, records were removed if data were missing on all demographic variables (i.e., sex, race, and ethnicity). The final counts were 241,412 for ASVAB, 49,681 for CT, and 39,213 for CS. The presenter continued by showing a chart of ASVAB, CT and CS applicant numbers by each of the reference and focal groups.

The presenter then presented charts showing:

- Impact ratios for AFQT cut scores FY 2023 IIIB+ and IIIA+
- Comparison of impact ratios for odd-number years FY 09-23
- Comparison of effect sizes for odd-numbered years FY 09-23, AFQT scores
- Comparison of effect sizes for odd-numbered years FY 09-23, non-AFQT tests

He concluded that the magnitude of impact on the ASVAB has remained fairly consistent across fiscal years but still varies in size from negligible to large across tests and groups. A comparison of impact across testing programs gives some indication of whether the observed FY 2023 magnitudes are reasonable. Sufficient information for estimating effect sizes is available online for two other large-scale testing programs, the SAT math and reading tests, and the National Assessment of Educational Progress (NAEP) grade 12 reading, math, and science tests. A series of charts presented these results. The presenter concluded that for the AFQT tests and GS, the direction and magnitude of overall impact is generally consistent with comparable SAT and NAEP tests, which suggests that impact on ASVAB tests is reflective of differences in job or training performance. Comparisons across programs may be somewhat restricted due to differences in such factors as group definitions, testing populations, and test content. NAEP is effectively an unrestricted sample and those selecting into the Armed Services likely differ from SAT test-takers in terms of personality, motivation, and other characteristics. Adverse impact does not reflect test 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 test 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. CT and CS generally exhibited small to moderate effects and were usually as low or lower than most ASVAB tests. Effects for CT and CS were also generally consistent with those found in FY 2021, with the exception that the CS non-Hispanic White-non-Hispanic Black effect size was near 0 in FY 2021 and near .30 in FY 2023.

The presenter concluded by asking if the members of the DAC-MPT had any general feedback or recommendations based on the results and if there are other results that they would be interested in seeing.

At the end of the briefing, a committee member complimented the charts as being easy to understand. Noting the reference group is always the group with the highest selection rate, s/he asked if there are any selection composites for which the reference group is smaller than the comparison group and would there be any effect. The presenter said it would be the non-Hispanic Asian – non-Hispanic White comparison, if it exists at all. The presenter said there are not instances where the AI ratio would be higher for the focal group. A committee member clarified, saying, though it may not apply to the Services, another way of looking at the impact of composites is to do so by looking at relative sizes of groups in the general population. The Assistant Director (AP) said DTAC would take that into consideration.

A committee member commented that the Air Force has jobs that require a wide array of intellects and asked if DTAC has data that shows which military branch the highest scorers enter. The presenter said the information is available. The Assistant Director (AP) said, anecdotally, she sees higher AFQT scores in the Navy and Air Force, though the FSPC has changed that to some degree in the recent years. However, most FSPC participants enter with one score, but their official score, which is taken at the end of the course, is higher than their original score.

The Acting Director (DTAC) confirmed awareness that the referent group is the group with the highest selection rate but asked if the counts were done within a year. That is, could the referent group change year-to-year? The committee member said yes, in the private sector, one would never look at every single person who took the test, but it would be analyzed within a specified timeframe. Additionally, one would look at relevant groups as well as by geographical area. S/he said much of that does not apply to the military's processes. The Assistant Director (AP) thanked the presenter for a great presentation and for providing information to consider going forward.

14. <u>Curriculum Alignment Study</u> – (Tab Q)

A HumRRO representative presented the briefing.

The presenter began by presenting the goals of the high school curriculum study which are to determine how ASVAB subtests align with content taught in high schools, explore how ASVAB content is taught, and map ASVAB content to other relevant sources. The study design should include (a) a review of previous high school curriculum and high school assessment alignment studies with ASVAB content, (b) a review of previous mappings between ASVAB and other tests, (c) a review of any available NAEP transcript studies, and (d) a method for assessing if there are differences in course-taking behavior patterns between military applicants and the general high school population.

The presenter continued by providing an overview of current trends in teaching practices. The development that had the most significant potential impact on educational approaches in the past 20 years was the introduction of the Common Core State Standards (CCSS) for English Language Arts and Mathematics in 2009 and the Next Generation Science Standards (NGSS) in 2011. The common core recommended an emphasis on complex texts and writing assignments that called for the use of evidence to support arguments. In regard to math, the goal was to encourage teaching practices that support gaining a conceptual understanding of underlying principles. The NGSS place an emphasis on developing an understanding of core underlying principles, using that information to generate and apply models to explain various phenomenon, and treating science as a progression that builds throughout a student's time in school. In both cases, research has produced mixed results regarding impact.

The presenter next addressed other trends in teaching practices including integrated instruction, where content is blended within and across disciplines. Research has shown mixed results, with more positive results at lower grades. Other trends in teaching practices include:

- Identifying and applying learning progressions, which starts by specifying the ultimate learning objective and moves backward to identify all the prerequisites.
- Microlearning involves breaking instructional material into small chunks and incorporating assessments throughout to ensure that students understand fundamental content before moving to more complex content.
- Flipped instruction moves the introduction of content outside the classroom so that class time can be spent discussing and developing an understanding of it.
- Project-based instruction involves having students, individually or in groups, apply what is learned in the classroom and what they discover through their own research to develop solutions to real-world problems.
- In a National Center for Education Statistics (NCES)-funded study of the use of technology in the classroom, 84% of schools indicated that technology was being used for activities normally done in the classroom, with 54% suggesting that the activities would not be possible without employing technology.

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

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

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

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

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

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

The presenter then discussed studies examining high school course taking behavior. These largely fell into one of four broad categories: (a) course-taking behavior and changes in course taking over time, (b) the impact of course taking on future outcomes, (c) changes in and the impact of Career and Technical Education (CTE) course taking, and (d) methodological studies. Much of the research is based on NCES-funded studies, including the High School Longitudinal Studies (HSLS) and the High School Transcript Studies (HSTS).

Overall, the results suggest that, over time, students were earning more credits and pursuing more challenging curricula. However, there is evidence that course titles may not accurately reflect content. Data from 2019 suggest only 12% of students followed a rigorous curricula and 23% were below standard.

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

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

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

Several approaches are being taken to achieve the goals of the current research. The alignment work done in 2015 is being reexamined to see if there have been shifts in the sources used that indicate a greater or lesser alignment with the ASVAB. Another type of alignment study is being conducted in which course catalogs from a sample of high schools across the country will be collected, and SMEs will be asked to review ASVAB test blueprints along with relevant high school courses and make judgments regarding the degree to which the ASVAB content is covered. Questions were included in the Futures Survey conducted by the JAMRS branch, asking respondents to indicate courses taken. Analyses were run to compare results for respondents who indicate a propensity for enlisting to those who are not propensed to see if there are differences in course taking. Another question focused on extracurricular activities that may be relevant, such as participation in clubs or special interest groups. Finally, data from the 2019 High School Transcript Study will be explored to see if there are relevant results that have not been reported in the literature. This work is in progress.

In regard to the review of comparable taxonomies, the ACT Curriculum Study included a survey of high school English Language Arts teachers asking them to identify the topic areas most frequently taught. The highest rated were composing skills and strategies, vocabulary comprehension strategies, analysis and evaluation of texts, and inferential comprehension of texts. HumRRO PC item editors reviewed the findings and agreed that vocabulary is covered (by Work Knowledge [WK]), inferential comprehension is addressed, and analysis and evaluation of texts is partially covered (no evaluation). Composing skills and strategies are not addressed. ACT includes standards for various ranges of scores for their reading test. The comparisons with ASVAB are not clear-cut due to the inclusion in the standards of “somewhat challenging” and “challenging” passages. ASVAB PC passages are limited to 100-180 words to eliminate scrolling, while ACT passages average at around 800 words. Nonetheless, the PC editors agreed that most standards are addressed, with the exceptions including determining cause-effect relationships and making comparisons between passages.

Comparisons with the NAEP reading assessment and achievement level definitions are also not straightforward. NAEP includes items that require comparisons between two or more texts, and passage length can range from 1,000 to 1,500 words. Seven item types are included, only one of which is used in PC (i.e., single selection multiple choice). The PC editors agreed that the Basic Achievement Level Standards are addressed in ASVAB. Those at higher levels (i.e., proficient, advanced) are only partially covered or not covered. Common characteristics of standards not covered include presenting diagrams and charts, comparisons between texts, and items requiring analysis, evaluation, synthesis, and critique of texts.

Turning to MK/AR, the presenter reported that the ACT Curriculum Study also asked math teachers to rate the most important skills to be developed. The four skills identified by HumRRO math editors as not covered by ASVAB were higher level (e.g., Math 3, Algebra 2). ACT also sets standards for various skill ranges on their math test. Editors indicated that 12 skills at the lowest level covered (13-15) are addressed and the remainder could be assuming they could be assessed through multiple-choice questions (e.g., locate positive rational numbers on a number line). All skills at the 16-19 level are or could be addressed except for one involving probability which is not in the existing blueprint. Several skills at the 20-23 level were judged to be outside the AR/MK blueprint, and others were judged to be included or candidates for inclusion in AR/MK.

The 2022 and 2024 NAEP Mathematics Assessment Framework includes objectives deemed appropriate for assessment by subtopic and grade. Math editors agreed that all objectives in Numbers, Properties, and Operations are covered, partially covered, or could be covered, except for measurement in triangles. Most objectives in Geometry, Algebra, and Data Analysis/Statistics/Probability were judged outside of the AR/MK blueprint and would require more extensive item types (e.g., describe, analyze, explain).

The Next Generation Science Standards cover three broad areas: Physical Sciences, Life Sciences, and Earth/Space Sciences. These are also addressed in the ASVAB. Subareas within each define skills high school students should be able to demonstrate, with an emphasis on application of knowledge rather than retention. As a result, most would require alternate means of assessment (e.g., conduct a project, write a paper) or more expansive item types (e.g., develop a model).

The ACT Science Test also covers three broad areas: Life Science/Biology, Physical Science/Chemistry/ Physics), and Earth/Space Science. HumRRO's GS editor judged all to be addressed in the ASVAB. The ACT Science and Readiness Standards describe what students at various score levels should be able to do. There are three broad areas: Interpretation of Data, Scientific Investigation, and Evaluation of Models/Inferences/Experimental Results. HumRRO's GS editor indicated that the descriptors do not represent the way in which content is covered by ASVAB (e.g., compare, determine), although certain topic areas are addressed.

The National Academy of Sciences, National Research Council for K-12 Science Education covers four broad areas: Physical Sciences, Life Sciences, Earth/Space Science, and Engineering/Technology/ Application of Science. HumRRO's GS editors judged nearly all to be covered in GS, except the latter. The 2028 NAEP Science Framework addresses the first three listed above, and the GS editor identified all topic areas as addressed by ASVAB except Evidence of Common Ancestry and Diversity.

The presenter summarized by indicating that ASVAB addresses the preponderance of content covered in the sources reviewed. Possible additions to the blueprints were identified, although many skills not addressed by ASVAB would be difficult to assess through a test or would require more complex item types. The differences in the underlying purpose of the ASVAB (selection/classification) and other tests (diagnostic/developmental) may obviate the need to assess knowledge and skills in similar ways.

The sampling plan for the course catalog portion of the work involved (a) randomly selecting one state from each of the nine Census regions; (b) creating an extract of data from the Common Core of Data for each state that lists all schools in each state; (c) sorting the schools by level and eliminating Pre-K, elementary, and middle schools; (d) sorting schools by type and eliminating special education, unknown, and alternative schools; and (e) generating random numbers to select five schools from each state. The results of this process led to an underrepresentation of City/Large schools given that three of the selected

states had no City/Large schools. As a result, one City/Large school was randomly chosen from two of the remaining four. Texas and Florida were added to represent high recruitment states. The websites for the selected schools were reviewed for course catalogs, which were found in 30 of 49 cases. The schools that did not supply catalogs typically were quite small. Additional samples within the state/size jurisdiction groups were drawn until course catalogs were located. This could mean that smaller schools will be underrepresented in the sample. ASVAB item writers/editors were identified to serve as SMEs. A ratings spreadsheet was created and virtual meetings were held to discuss the purpose of the task, explain how the schools were selected, and provide guidance on using the spreadsheet. The task is ongoing; the results presented here reflect findings to date.

SMEs indicated that all ASVAB AR/MK content was covered, either in prerequisite courses (to those in the catalogs) or by basic courses in the catalogs. One possible exception was Time/Temperature, with SMEs identifying few explicit mentions. For GS, all topics were covered in a mixture of basic and advanced courses, with the exception of Botany, which was not addressed in approximately 60% of the catalogs. Of the 56 catalogs reviewed thus far, 34 were identified as having no automotive technology/repair classes. Shop Information content was available in about two-thirds of the catalogs reviewed thus far. All six blueprint elements of the MC test were covered in the catalogs reviewed. Regarding the CT, 10 schools offered no related classes, and 8 only provided courses in the use of information technology and software. All test components were covered in 14 schools. The topics most likely to be omitted were Network Configuration, Offensive Methods, and PC Configuration and Maintenance.

The presenter then turned to the results from the items inserted in the JAMRS Advertising Tracking Survey. Overall, there were a small number of propensed respondents (89 of 880). Significantly higher proportions of respondents *not* considering military service reported taking biology, chemistry, calculus, and statistics/probability. A significantly higher proportion of those in the "definitely not enlist" category compared to those in the "probably not enlist" category took chemistry and statistics/probability. Finally, a significantly higher proportion of those in the "definitely not enlist" category took business/marketing compared to those in the propensed group. The presenter then showed a table summarizing the results.

In regard to extracurricular activities, participation was generally low; below 10% in most cases. The highest participation levels were in social service/volunteer efforts, sports, cheerleading, and computer-related pursuits. There were few significant differences, with the most notable being higher percentages of those in the medium- and high-propensity groups taking part in automobile and construction activities, both of which are relevant to the ASVAB.

The presenter concluded by stating that ASVAB content is largely addressed in the relevant frameworks reviewed. Some suggestions arose for additions to the blueprints, although assessing some content would require an expansion of item types. ASVAB academic content (e.g., GS, AR/MK) is typically covered in high school courses, but technical content coverage is spottier. There is some indication of course-taking differences between propensed and non-propensed youth, with the latter taking higher level courses. In addition, there is some indication that propensed youth are more likely to take part in extracurricular activities relevant to the ASVAB (e.g., automotive, construction).

The presenter concluded by asking if the DAC-MPT had recommendations for how this work can improve the composition of the ASVAB for selection and classification purposes. The presenter also noted that the ASVAB currently assesses both knowledge learned in school and knowledge and skills needed in the military that may not be addressed in formal education. The DAC-MPT was asked for their thoughts on how the next generation ASVAB can continue to bridge that gap.

At the end of the briefing, a committee member provided a list of questions, which included: What changes to the ASVAB might make it more effective at assessing knowledge gained through modern educational trends (e.g., integrated or flipped instruction)? How do Reserve Officers' Training Corps (ROTC) students differ from high school students in their course patterns? Should the ASVAB include other more open-ended or project-based question formats,

despite the challenges in scoring and implementation? Would you consider involving educators, recruiters, and SMEs in workshops to refine crosswalks and ensure content reflects both school and military needs? Would you consider conducting follow-up studies on test-takers to evaluate how changes influence career trajectories and military readiness?

The presenter said the last item was a very good idea, and then welcomed the committee member's thoughts on some initiatives they are launching under the CEP. A committee member said s/he appreciated all the work it took to do the comparisons, which showed great alignment with content. The committee member reported seeing this with a lot of the standards in science and engineering, as well as some in reading and math. S/he asked if the current assessments require a depth of knowledge that can reflect the different kinds of cognitive skills covered by the ASVAB? S/he noted that there have been changes in the way certain knowledge and skills are taught. The Next Generation Science Standards (NGSS) have been out for 13 years; is DTAC finding sufficient measurement of the higher order thinking skills? The presenter attempted to answer saying General Science (GS) is a declarative knowledge type test and asked if they needed to assess the higher cognitive skills as well. That is, would that provide value for selection and/or classification? The presenter asked if anyone who has been working with item writing could comment. A DTAC representative said item writing is his division in DTAC, and his thinking aligns with the presenter's: predictive validity is the objective. They would be interested in alternative item types measuring more in-depth thinking if they could provide additional validity. Much effort is poured into developing new item types, but without a lot of payoff in predictive validity, but that does not mean they should stop seeking improvement. A DTAC representative said, as a parent, when working through math problems with kids, it is possible to reach the correct answer but use the wrong technique, at least according to how it is taught in school.

A DTAC representative said the Assistant Director (AP) had mentioned the ASVAB modernization effort. The DTAC representative said it would not be out of balance for them to adjust at the margins of the blueprints – such as they have done with the CT – but these types of changes risk changing the nature of the test. A minor shift or new content area that does not seem too risky would be reasonable. The Assistant Director (AP) said DTAC would share more about ASVAB modernization in the future, and that it could include actions such as moving CR from the special test domain into the AFQT. The overall effort will include discussions about all the tests, what is to be done with them, and the cost/benefit of changes. The Assistant Director (AP) said it is a long-term effort but they can start thinking on a smaller scale sooner than later for such things as assessing in-depth thinking and verbal skills. A committee member followed up by saying s/he was not completely sold on the new formats; in reading, for example, there are no charts or graphs, but including these could change the nature of what is assessed, even if the items remained multiple-choice. S/he said the same could be done for reading and evaluating a claim.

A HumRRO representative said a study in 1997 investigated whether different types of items and content would have an impact on validity. Some content was matched to high school curriculum and some to jobs. They developed a skills taxonomy to accompany the content taxonomy. There were many application-type items. Bottom line, the efforts had zero impact on group differences or validity. To the DTAC representative's, it was a lot of effort without any bang; but to the

committee member's point, there seems to be room to develop multiple-choice tests that do more than measure the regurgitation of knowledge, and that should be explored. A committee member clarified that s/he does not believe the current tests require simple recall, but s/he does think there is an opportunity to tap other skills. A DTAC representative said the conclusion seems to be that they need to be writing better multiple-choice items, or at least reoriented multiple-choice items.

15. ASVAB CEP – General, Find Your Interests, and Work Values – (Tab R)

This briefing included three parts: (1) a general CEP presentation by representatives of DTAC and MEPCOM, (2) a Find Your Interests (FYI) Inventory presentation by a HumRRO representative, and (3) a Work Values Situational Judgement Assessment (WV SJA) presentation by a HumRRO representative.

The first part of the briefing focused on the CEP, and began with a review of the mission and vision for that program. The CEP's mission is to provide a career exploration service to American youth and qualified leads to military recruiters. The CEP assesses academic ability and vocational interests, which together help inform career decisions. Personalized career exploration, awareness of career field entry requirements, and future-oriented planning tools help students work with parents and educators to develop postsecondary plans. Eligible participants use their scores to explore enlistment and have no obligation to join the military.

The presenter displayed figures depicting program participation. The number of participating students rose to 619,926 across 13,105 schools. This led to over 339,463 leads. P&P testing declined from 91% of all testtakers in 2018-2019 to 54%, while *i*CAT testing rose from 9% to 46%.

The presenter described the 2024 ASVAB CEP Jamboree, which was a three-day strategic planning session with stakeholders. The session focused on reviewing the past year's performance and achievements and brainstorming for the future. The presenter then displayed a chart summarizing the various components of the CEP ecosystem of integrated business strategies. These included:

- Technology. Objectives are to optimize user experience by enhancing features and addressing bugs; migrate CEP websites into Defense Personnel Assessment Center System (DPACS) boundary to enhance security; consolidate backend systems for operational efficiency; and expand data analytics to inform decision-making. School year (SY)24/25 goals are to migrate ASVAB Program and Careers in the Military (CITM) websites into the DPACS boundary NLT August 2025.
- New Research and Innovation. This includes studies to evaluate and improve CEP measures/processes: (a) students' readiness to benefit from CEP, (b) use of AI to improve occupational crosswalks, (c) evaluation of non-cognitive measures, (d) expansion of post-test interpretation (PTI) delivery, and (e) use of external data to inform program impact. SY24/25 goals are to leverage research and innovation to enhance the ASVAB CEP program, improve occupational crosswalks, and address stakeholder needs and concerns.
- Occupational Website Data and Content. One of the primary benefits of the ASVAB CEP to users is the data contained on the program's websites. This initiative focuses on the activities undertaken to collect, analyze, store, and share occupational data. The SY24/25 goals are to define an occupational crosswalk process and explore utilization of AI to further enhance collection and analysis.
- Promotion and Engagement. Advertising, social media, content marketing, national events, and stakeholder engagement provide opportunities for knowledge sharing and interaction with various customer segments of ASVAB CEP's target audiences. SY24/25 goals are to execute the SY24/25 Social Media Strategic Plan, increase program awareness, and grow social media presence.
- Workforce Multiplier. The personnel responsible for delivering the ASVAB CEP require awareness and training. This initiative seeks to expand the numbers and the knowledge of those who can speak to the benefits of the program. SY24/25 goals are to expand the PTI training

program and work strategic partnerships with U.S. Army Recruiting and Retention College leaders, Junior ROTC (JROTC), and MEPS battalion commanders.

- Legislative Activities. This includes monitoring ASVAB CEP-related legislative activities, systematizing Department of Education connections, and following up on and maintaining connections made at conferences. The SY24/25 goals are to continue tracking state and federal legislation and develop interactive mapping and visualization tools.
- Underserved Populations. The ASVAB CEP benefits young adults. This initiative seeks to expand access to the ASVAB CEP among eligible students in post-secondary institutions, homeschooling, and schools that don't offer ASVAB CEP. SY24/25 goal is to create a pilot program with the goal to increase private and homeschool testing as well as post-secondary institution participation.

The FYI Inventory presentation began with an overview of the assessment. The original form was developed in 2005. It was a 90-item RIASEC measure, with dislike/indifferent/like response options. FYI Inventory scores are reported using total-group and sex-specific norms. In 2017, DTAC convened an ASVAB CEP Expert Panel, which reviewed the components of the revamped CEP, giving particular emphasis to the FYI Inventory. The panel suggested updating the inventory to ensure currency/relevance of items and construct coverage per basic interests (Su et al., 2019). The presenter showed a chart of the 41 RIASEC basic interests.

The presenter then described the FYI form development and analysis, which began in 2019. HumRRO drafted 450 new FYI items for field testing. This effort was driven by expert panel guidance to focus on content validity and construct coverage, identify contemporary content related to emerging economic changes, build on existing items with an enhanced item pool rated by a panel of experts, and identify Basic Interest Indicators, using Su et al. (2019) and the Strong Interests Inventory as frameworks for potential detailed basic interest markers. Enemy and clone items were identified by employing NLP procedures. 230 of the 450 items were field tested, and DTAC used those data to develop a new FYI form. For each RIASEC scale, DTAC retained 7–10 items from the current form, adding 5–8 new items.

Following construction of the form, HumRRO reviewed the form for content, emphasizing the basic interests taxonomy per guidance from the ASVAB CEP Expert Panel. The review revealed only partial coverage (61%) of the 41 basic interests in Su et al.'s (2019) taxonomy: 3 of 10 for Realistic, 3 of 4 for Investigative, 6 of 7 for Artistic, 3 of 8 for Social, 7 of 8 for Enterprising, and 3 of 4 for Conventional. Because the CEP links to O*NET occupational information, HumRRO proposed two other options for the new form that would increase coverage of the basic interests. In the end, three forms were evaluated:

- Form Version 1 was assembled with focus on item statistics and IRT parameters. It retains a majority of original FYI Items.
- Form Version 2 places more focus on basic interests, but retains a mix of original and field test FYI items
- Form Version 3 focuses primarily on basic interests. Most items (79%) were new and selected to ensure coverage of all basic interests. There was no requirement to retain any previous items, though 19 were retained.

The presenter showed a table of the basic interest coverage provided by the original (current) form, followed by a table showing coverage provided by the three proposed forms. The presenter then provided internal consistency estimates at the occupational theme (e.g., realistic, investigative) level for each version of the form. The new forms had lower reliability estimates than the current form, but this was reported to be acceptable and even desirable given the heterogeneity of each RIASEC dimension. The presenter then turned to sex differences across forms. Form Version 3 was found to have the smallest subgroup differences, but higher conventional *d* values due to inclusion of Information Technology. Multidimensional scaling results were shown for each proposed form version, with the presenter discussing the shapes of the points on the graphs and recommending Form Version 3. This version provides strong psychometric characteristics. That is, it has more reasonable (i.e., not too high) reliability estimates, the smallest subgroup differences despite not purposefully selecting items with this criterion in mind, and complete coverage of the basic interests. Next steps are to finalize dimensionality analyses (item-level exploratory factor analysis; CFA models [standard, circumplex]), field test and analyze the new form, and establish norms for new form.

The presenter concluded by asking the committee about their reactions to the new form, including their concerns. The committee was also asked if it wanted to see additional analysis/information before field testing the proposed new form and if it had suggestions for designing the field test or recommendations for establishing norms.

The third part of the briefing, covering the WV SJA, began with the goal of the effort, which is to explore the possibility of creating a work values assessment to add to the ASVAB CEP. Though work values tend to have greater meaning and utility for experienced workers, the original idea was to introduce CEP participants to the concept of work values, for example, to facilitate discussions between students and counselors or teachers. Development began with a systematic review of pinnacle research publications. Based on these reviews, various inventory formats were proposed, including (a) ipsative, IRT-based scoring model pairing work values statements against one another; (b) situational policy-capture approach to measuring work values using regression-based methods for scoring; and (c) multiple-choice items with basic mathematics for scoring. Due to administration time constraints and accessibility with P&P administration, the third option was chosen. The presenter then identified the products of the research, including the WV SJA as well as versions of other proposed activities, including a Realistic Job Preview, Personal Values and Work Values, the Intersection of Work Values and Work Interests, How Has the Pandemic Made You Think About What You Value, and a Structured Interview.

The WV SJA is a situational judgement test (SJT) that assesses the six work values from the Theory of Work Adjustment (Dawis et al., 1964, 1968; Dawis & Lofquist, 1976, 1978). It introduces students to work values. It is linked to occupations (as are the ASVAB and FYI) to permit career exploration in terms of work values. The presenter then showed a list of work values (i.e., achievement, independence, recognition, relationships, support, and working conditions) and their definitions.

The introductory screen for the assessment provides a short set of instructions and informs the user of the six work values. The presenter showed two example SJT items, one that used a school context and another that used a work context. A results page provides a list of the work values selected by the tool based on user responses. It also provides a brief description of workers who score high on the various work values. An option is available to allow the user to explore further or retake the assessment. The assessment results also show the user where they scored similarly (or tied) across multiple work values and asks the user to identify which sounds most like them.

The presenter then showed the preliminary results of an analysis of WV SJA response data. There are currently fewer than 42,000 responses in the uncleaned data. Initial results reveal modal response profiles and differences by sex and context (school and work). After providing demographics of the sample, the presenter showed the top work values profiles in rank order from 1 to 10, including the number of students having each profile. The top ranked profile (i.e., relationships-support-achievement) was assigned to 948 students. Next, a graph showed the number of times each work value occurred in the top 10 profiles. Achievement occurred in all 10, while the next two most frequently occurring work values were support and relationships at 6 occurrences each. Working conditions was the least frequently occurring value. Another graph showed the top work values by sex. Achievement, independence, and recognitions were higher for males and relationships, support, and working conditions were higher for females. The presenter then showed the top work values profile by sex and series of bar charts showing item endorsement in the school context versus the work context. Regarding the average endorsement between contexts, there were significant differences for all values except recognition.

At the end of the briefing, the presenter asked the DAC-MPT if, given the respondent population, the WV SJA should focus on a single context (i.e., school or work). The committee was also asked if it had concerns with using the assessment to identify occupational matches.

As the CEP briefing concluded, a committee member congratulated the CEP National Director on the recognition her team and program are receiving. The Assistant Director (AP) spoke briefly about the goals for testing. The number of students tested prior to COVID-19 was

800,000. After dipping severely during the COVID-19 timeframe, the program is back up to 600,000. AP is hoping to get back to 800,000 soon. The Assistant Director also stressed the importance of making better use of PTIs so that the program is more than just another standardized test.

After the FYI Inventory briefing, a committee member said s/he appreciated the reduction in group differences. A committee member then asked how the team arrived at the third version of the form. The presenter explained that Form Version 1 retains a majority of the original items. Form Version 2 includes a mix of original and new items. Form 3 comprises almost 80% new items. A committee member said, looking at the scaling results (slides 25-27), Version 3 is the least round and seems to do the worst job of approximating the RIASEC model. The presenter clarified that only a few people responded to both items in some of the item pairs, which prevents them from making any strong claims now. The team will have a better standing to consider making adjustments before it goes final.

The Assistant Director (AP) commented that there are no cut scores on the FYI, but that it is strictly for exploration, which provides the flexibility for the test to have lower reliabilities than would otherwise be required. Beyond that, the test seems to effectively provide the intended type of information, especially how they compare to other people like them. The Assistant Director (AP) asked the committee for feedback on the need for sex-based norms, given that the inventory will not be used for selection or classification purposes; that is, it will not be used to compare individuals against others, but only to provide an opportunity for self-exploration. A committee member said a full norm-based tool that allows users to use their subgroup might be more fitting than providing un-normed results. S/he agreed with the importance of helping people learn how much their profiles resembled others like them.

A committee member said s/he appreciated the move to basic interests and noted that it merges with other efforts that align with O*NET. S/he said this gives up a bit of the RIASEC dimensionality, but it allows better career exploration. S/he said the more information you give them, the more differentiated their interests can become. S/he said it should be a good sell to States who are interested in their workforces, and the test should align more toward that than toward the RIASEC. The presenter said, because O*NET is headed in that direction, they do not want the FYI to be left behind. In closing, the Assistant Director (AP) asked if the committee had any reactions to the new FYI form or hesitation about Version 3. There was no concern or hesitation among the committee members.

At the end of the WV SJT briefing, a committee member thanked the presenter and said the work was fascinating. The committee member suggested introducing more variability in context, such as volunteer opportunities. S/he asked if there were any post-survey questions on whether individuals had a difficult time answering the questions if they had not had a job. The presenter said they had discussed this topic to gain insight into internal processes, and that the processes used to judge values in work versus school contexts likely differ. The committee member thanked the presenter for her explanation.

A committee member commented on the WV SJA potentially focusing on a single context, saying that focusing only on the school context reduces the variability in exposure; if both

contexts are considered, the results can recognize more variability as well as shared values across contexts. Another committee member agreed. S/he also commented about the item related to internships on slide 39, asking how many test-takers will have had internships. S/he noted the requirement to examine item content very carefully to determine if the experience is sufficiently common to reference. The presenter said that is a great point and explained that the team had discussed taking a deep dive into the content; the presenter agreed that the internship item was not a great example of the representativeness of items associated with that particular context.

Regarding concerns with using the WV SJA to identify occupational matches, a committee member suggested it was important to consider levels of subjectivity. S/he explained how varying income levels affect students' work experiences; that is, kids in lower income families may be more likely to have to work in addition to going to school than kids of families that are more affluent.

16. <u>Future Topics</u> – (Tab S)

The Acting Director, DTAC, presented the briefing.

The Acting Director presented a list of potential topics for future DAC-MPT meetings:

- ASVAB evaluations
- CAT-ASVAB/Form development methodology
- Unproctored testing
- Super-scoring
- Adding new non-cognitive measures
- Calculator effort
- Validity
- Explore AI/GAI/technology advancements
- Next generation testing
- Adding new cognitive tests/composites

The Acting Director (DTAC) began the discussion of future topics by asking what the committee could support best. The Acting Director said the list of topics to discuss must be narrowed to an attainable number, and a priority listing would be best. The Acting Director mentioned two other topics that might be priorities: development of an SJT for military compatibility and FSPC research by the Services.

A committee member said s/he would like to see more on data collected on the TAPAS, as well as the high school curriculum work and ASVAB CEP. The committee member said the discussion of those could be combined toward the goal of discussing the incorporation of new learning pedagogies. S/he also recommended including a presentation on the FSPC. A committee member requested an update on the FYI and the normative information that is collected, as well as progress on the SJA if there is enough to report at that time. A committee member suggested a

presentation or discussion on how to identify people for the Preparation Courses and who will benefit if there are data relevant to those topics.

A committee member asked if there was a theme or story related to the calculator studies that has unfolded across meetings that would be useful to discuss. S/he mentioned the limited number of items on which data had been collected and asked if they wanted the DAC-MPT to recommend a conclusion at this time. The Assistant Director (AP) said they are still trying to understand better whether a special test will provide value. AP is interested in DAC-MPT recommendation in this realm. Do you see value? Have we exhausted the domain of potential analyses? A DTAC representative agreed with the Assistant Director (AP): This is a difficult study to conduct without intruding upon operations. If the committee has recommendations, let DTAC know, but it seems like we know all that we can know now. A DTAC representative said feedback on the needs assessment could be useful, but otherwise, DTAC has what it needs. The Assistant Director (AP) agreed that the focus should be on the needs assessment.

A committee member expressed interest in hearing more about AI and asked if DTAC is planning to develop AI-powered personalized learning tools, such as the tutoring sites for the ASVAB. S/he asked who runs those sites and the Assistant Director (AP) responded that in most cases it is not the DoD, and that AP cannot comment on the non-DoD sites. The Assistant Director clarified that there is a March2Success program that was developed by the Army and is available to the public. The Acting Director (DTAC) followed up by stating that DTAC does not have personalized learning tools that are AI. The Assistant Director (AP) clarified that DTAC is not responsible for developing educational materials, though the Services can, with approval from the Under Secretary of Defense for Personnel and Readiness.

A committee member said many topics were of interest, but prioritization is the issue. S/he said the committee wants updates on the difficult work DTAC is performing and on their priorities. The committee member mentioned four cross-cutting themes that could guide future presentations. First, ensure all stakeholders are able to see what they need to see. Second, orient on consequential decision making and ask where consequential decisions are being made and how do data inform those decisions. Third, focus on the integration of measures and tools, conceptually and operationally. There has been a great range of presentations, but more synthesis in regard to how measures connect conceptually and operationally would be helpful. Fourth, any developments in AI that impact, particularly, the ASVAB and TAPAS.

The Assistant Director (AP) said they would rethink the organization of the presentations in light of these suggestions. A committee member conceded it would be a balancing act. The Assistant Director (AP) concluded by saying the topics are based on DoD priorities but AP wants to take into account the interests of the committee as well. AP is keenly interested in how to integrate CR into the suite of assessments, TAPAS integration, and redefining recruit quality.

17. Public Comments

At the end of the second day, the Assistant Director (AP) opened the floor to public comments and asked participants to limit their comments to no more than 5 minutes per person. There were no comments.

18. Closing Comments

The Committee Chair said s/he appreciated everyone's commitment to the cause of making the big picture happen. This is a high-quality assessment that supports high-stakes decision making. The quality is evident on many levels, and the DAC-MPT appreciates the attention to detail and desire to do things the right way, especially within the continually shifting landscape. The research and management are much appreciated. The Chair said the meeting covered a wide range of topics and issues to ensure the ASVAB moves deep and broad in enhanced fashion, to include the examination of CompT and tools like FYI that will benefit society in general. Moving forward, the committee will summarize their input on this remarkable work and hope that it is considered carefully and used appropriately as you move forward as a team. The expertise you have in-house (DTAC and HumRRO) is formidable. The Assistant Director (AP) said thank you to all the participants.

Tab A

LIST OF ATTENDEES

Defense Advisory Committee on Military Personnel Testing (DAC-MPT)
January 22-23, 2025

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

Dr. Jennifer Tucker	Assessment Branch Chief	US Space Force
SGM Alan Myers	Senior Retention & Accessions Policy Manager	US Army HQDA, G1
Dr. Tonia Heffner	Chief, Selection and Assignment Research Unit	US Army Research Institute (ARI)
Dr. Cristina Kirkendall	Research Psychologist	ARI
Dr. Kirby Hockensmith	Research Psychologist	ARI
Dr. Sophie Romay	Senior Personnel Research Psychologist	US Air Force Personnel Center
Dr. Andrew Deregla	Research Intern	US Air Force Personnel Center
Dr. Benjamin Gilbert	Personnel Psychologist	US Air Force Personnel Center
Mr. Ken Schwartz	Chief, Testing and Survey Policy	US Air Force Personnel Policy
Mr. David Jackson		US Air Force
Mr. James Johnson	Director, Selection and Classification	US Navy, OPNAV N132
Mr. Robert Tiegs	Testing Director	US Military Entrance Processing Command (USMEPCOM)
Mr. David Davis	Chief, Testing Division	USMEPCOM
Ms. Jaime Clayton	Enlistment Testing Program Program	USMEPCOM
Dr. Claire Vincent	Program Manager	HumRRO
Dr. Kimberly Adams	Program Manager	HumRRO
Dr. Scott Oppler	Chief Scientist	HumRRO
Dr. Rod McCloy	Chief Scientist	HumRRO
Dr. Dan Putka	Chief Scientist	HumRRO
Dr. Monica Gribben	Principal Scientist	HumRRO
Dr. Kevin Bradley	Principal Scientist	HumRRO
Dr. Jeffrey Dahlke	Principal Scientist	HumRRO

Dr. Glen Heinrich-Wallace	Senior Scientist	HumRRO
Dr. Nick Howald	Senior Scientist	HumRRO
Dr. Katherine Klein	Senior Scientist	HumRRO
Dr. Maura Burke	Senior Scientist	HumRRO

Tab B

DEFENSE ADVISORY COMMITTEE ON MILITARY PERSONNEL TESTING
TENTATIVE AGENDA
January 22–23, 2025

January 22, 2025 (Mountain Time)

8:30 a.m. – 8:45 a.m.	Welcome and Opening Remarks	Dr. Sofiya Velgach, OASD(M&RA)/AP
8:45 a.m. – 9:15 a.m.	Accession Policy Brief	Dr. Katherine Helland OASD(M&RA)/AP
9:15 a.m. – 10:00 a.m.	R&D Milestones Brief	Dr. Tia Fechter on behalf of Mary Pommerich (OPA/DTAC)
10:00 a.m. – 10:15 a.m.	*Break*	
10:15 a.m. – 11:15 a.m.	Update on Committee Recommendations	Dr. Tia Fechter on behalf of Mary Pommerich (OPA/DTAC)
11:15 a.m. – 12:15 p.m.	Update on P&P forms	Dr. Jeff Dahlke (HumRRO)
12:15 p.m. – 1:45 p.m.	Lunch	
1:45 p.m. – 2:45 p.m.	Form Equating Sampling Design	Dr. Jeff Dahlke (HumRRO)
2:45 p.m. – 3:15 p.m.	Complex Reasoning Update	Dr. Kate Klein (HumRRO)
3:15 p.m. – 3:45 p.m.	Computational Thinking Update	Dr. Kimberly Adams (HumRRO)
3:45 p.m. – 4:00 p.m.	*Break*	
4:00 p.m. – 5:30 p.m.	Calculator Analyses Efforts	
	a. Calculator Impact Study	Dr. Kevin Bradley (HumRRO)
	b. CAT Simulation	Dr. Glen Heinrich-Wallace (HumRRO)
	c. Calculator Needs Assessment	Dr. Monica Gribben (HumRRO)
5:30 p.m. – 5:45 p.m.	*Public Comments*	

January 23, 2025 (Mountain Time)

8:30 a.m. – 9:30 a.m.	Refinement of the Joint Service-TAPAS Instrument	Dr. Dan Putka (HumRRO)
9:30 a.m. – 10:30 a.m.	Adverse Impact	Dr. Nick Howald (HumRRO)
10:30 a.m. – 10:45 a.m.	*Break*	
10:45 a.m. – 11:30 a.m.	Curriculum Alignment Study	Dr. Rod McCloy (HumRRO)
11:30 a.m. - 12:30 p.m.	ASVAB CEP	
	a. General	Dr. Irina Rader (OPA/DTAC)
	b. Find Your Interests	Dr. Rod McCloy (HumRRO)
	c. Work Values	Dr. Maura Burke (HumRRO)
12:30 p.m. – 12:45 p.m.	Future Topics	Dr. Tia Fechter (OPA/DTAC) on behalf of Mary Pommerich
12:45 p.m. – 1:00 p.m.	*Public Comments*	
1:00 p.m. – 1:15 p.m.	Closing Comments	Dr. Fred Oswald Chair
1:15 p.m. – 3:00 p.m.	*Working Lunch (Administrative Items)*	

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

Tab C

LIST OF ACRONYMS

ACT	American College Testing Test
ADD	Attention-Deficit Disorder
ADHD	Attention-Deficit/Hyperactive Disorder
AFCT	Armed Forces Classification Test
AFQT	Armed Forces Qualification Test
AI	Artificial Intelligence or Auto Information (based on context)
AO	Assembling Objects
AP	Accession Policy
APT	AFQT Prediction Test
AR	Arithmetic Reasoning
ARI	U.S. Army Research Institute for the Behavioral and Social Sciences
AS	Auto & Shop Information
ASVAB	Armed Services Vocational Aptitude Battery
ATA	Automated Test Assembly
BCT	Basic Combat Training
BME	Bayes Modal Estimation
CAT-ASVAB	Computerized Adaptive Testing ASVAB
CEP	Career Exploration Program
CFA	Confirmatory Factor Analysis
CITM	Careers in the Military
CompT	Computational Thinking
COVID-19	Coronavirus Disease 2019
CR	Complex Reasoning
CS	Coding Speed
CT	Cyber Test
CTA-M	Computational Thinking Assessment for Middle Grades
CTE	Career and Technical Education
CWB	Counter-productive Work Behavior
DAC-MPT	Defense Advisory Committee on Military Personnel Testing
DEP	Delayed Entry Program
DFIT	Differential Functioning of Items and Tests
DIF	Differential Item Functioning
DPACS	Defense Personnel Assessment Center System
DTAC	Defense Testing and Assessment Center
EDTP	Electronic Data Processing Test
ENL	Enlisted Selection Composite
ES	Effect Size
ESS	Educational Services Specialist
ETP	Enlistment Testing Program
FACA	Federal Advisory Committee Act

FSPC	Future Servicemember Preparatory Course
FY	Fiscal Year
FYI	Find Your Interests
GAI	Generative Artificial Intelligence
GED	General Educational Diploma
GS	General Science
HumRRO	Human Resources Research Organization
HSLS	High School Longitudinal Studies
HSTS	High School Transcript Studies
*i*CAT	Internet version of the CAT-ASVAB
IRT	Item Response Theory
JAMRS	Joint Advertising, Market Research, and Studies
JROTC	Junior Reserve Officers' Training Corps
JS	Joint-Service
MAP	Measures of Academic Progress
MAPWG	Manpower Accession Policy Working Group
MARP	Medical Accessions Records Pilot
MC	Military Compatibility Composite
MCRG	Military Compatibility Research Group
MCt	Mental Counters test
MEPS	Military Entrance Processing Station
MK	Mathematics Knowledge
MOS	Military Occupational Specialty
MSLP	Modified Stocking-Lord Procedure
M&RA	Manpower & Reserve Affairs
NAEP	National Assessment of Educational Progress
NCES	National Center for Education Statistics
NDAA	National Defense Authorization Act
NGSS	Next Generation Science Standards
NLP	Natural Language Processing
OASD	Office of the Assistant Secretary of Defense
OPA	Office of People Analytics
OSD	Office of the Secretary of Defense
O*NET	Occupational Information Network
P&P	Paper-and-Pencil
PAY97	1997 Profile of American Youth
PC	Paragraph Comprehension
P*i*CAT	Pending Internet Computerized Adaptive Test
PTI	Post-Test Interpretation
RAISEC	Realistic, Investigative, Artistic, Social, Enterprising, and Conventional
ROTC	Reserve Officers' Training Corps
R&D	Research and Development
SAT	Scholastic Aptitude Test

SD	Standard Deviation
SEM	Standard Error of the Mean
SI	Shop Information
SJT	Situation Judgement Test
SME	Subject Matter Expert
SS	Service-Specific
SY	School Year
S-L	Stocking-Lord
TAPAS	Tailored Adaptive Personality Assessment System
TC	Transformation Constant
TCCs	Test Characteristic Curve
TOA	Theory of Action
TWG	Technical Working Group
USC	U. S. Code
USMEPCOM	U.S. Military Entrance Processing Command
VE	Verbal Expression
VTest	Verification Test
WK	Word Knowledge
WV SJA	Work Values Situation Judgement Assessment

Tab D

RICE

Department of Psychological Sciences
Fred Oswald
Professor and Herbert S. Autrey Chair in Social Science

March 2, 2025

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

Dear Dr. Helland,

The Defense Advisory Committee on Personnel Testing (DAC-MPT) is pleased to provide this report on our meeting of January 22-23, 2025, in El Paso, Texas. In addition to myself, the DAC-MPT Committee members are Drs. Sonia Esquivel, Won-Chan Lee, Osvaldo Morera, Nancy Tippins, and April Zenisky. All members attended in person except myself; because airports in Houston were closed due to a rare snowstorm, I attended virtually.

Overall, members of the DAC-MPT found this meeting to be highly productive and well-organized, like previous meetings, and reflective of significant progress made on multiple fronts since our previous meeting. All presentations were informative, as were our interactions during the day between DAC-MPT members and session participants. All stakeholders are highly collaborative and jointly committed to high-quality military personnel testing and the US military workforce.

The DAC-MPT report and recommendations follow in order of the meeting agenda.

Accession Policy Brief – Director, Accession Policy (AP), Dr. Katherine Helland (OASD(M&RA)/AP)

Dr. Helland began her presentation by comparing the recruitment goals for 2023 and 2024. In FY 2024, all military branches met their 2024 recruiting accession missions, except for the Navy Active Duty, Army Reserve, and Navy Reserve. Navy met their contractual goal but fell short of shipping the target recruits to basic training due to capacity limitations.

The general fragility of the recruiting market continues, indicating that all recruiting challenges have not yet been resolved. Historically low numbers of youth are currently propensed for military service due to numerous factors at play, including the lack of familiarity with military service, which is related to the decline in the numbers of veterans, a general decline in trust in the military and other government institutions, the demographic decline of 18-year olds, aggressive recruitment practices on the part of colleges and universities, fear of emotional or physical injury, and reluctance to leave family and friends. Eligibility also remains a barrier. Only

Email foswald@rice.edu · Office 713-412-2147 · Rice University Department of Psychological Sciences-MS 25
6100 Main St. · Houston, TX 77005 · workforce.rice.edu

23% of youth are eligible without some form of waiver. There is an uptick in the number of waivers given, some of which are due to electronic health records that provide more visibility into health problems and require a waiver.

Many initiatives are underway to increase propensity for service. For example, AP is focused on making sure that processing is not a barrier to enlistment. Joint Advertising, Market Research, and Studies (JAMRS) has just launched a new youth-oriented media campaign to encourage youth to consider military service. Adult influencers who view our ads are more likely to recommend military service and discuss the value of service with youth. AP is also working closely with the Department of Education and state education agencies to find ways for high schools to receive credit for graduates who join the military, on par with credit received for attending college. If high schools establish accountability metrics that include military service, they may be more willing to allow recruiters into the high school. Legislation is also in progress to allow recruiters to have higher quality access to high schools. AP is working with public service agencies, such as the Peace Corps and AmeriCorps, to identify effective ways to message national service.

AP is also addressing barriers to eligibility by conducting a pilot study that examines medical conditions and the window of time that the Services can look back at medical records. For example, in the pilot, the time frame for ADHD was shortened to one year with no medication. Because anyone entering the Services must be deployable world-wide, AP must consider the potential long-term impact of allowing people with given medical conditions to enlist.

Dr. Sofiya Velgach, Assistant Director for AP, continued the discussion on AP testing initiatives by defining what is meant by the "quality" of a recruit. In terms of the enlisted recruits, quality refers to the individual's Armed Forces Qualification Test (AFQT) score and education level (e.g., high school graduate, GED). However, whole-person assessment can expand the notion of quality and increase talent management in the Services further. In this vein, the Department is considering the introduction of measures of personality and fluid intelligence (e.g., complex reasoning), identifying and developing the appropriate measurement of such constructs, and ensuring they are sufficiently reliable and valid. A challenge is determining how best to integrate these measures into existing standards and processes (e.g., integrating personality measures with the AFQT). Moving forward, AP will need to work with Congress to ensure members understand future plans and to receive their guidance. Although laws will likely not need to be changed, conversations with key members of the House and Senate will be needed to facilitate understanding and buy-in.

Returning to the discussion of recruiting challenges, the DAC-MPT was updated on the efforts to innovate and modernize the recruitment process without reducing standards. Both the Army and the Navy have deployed the Future Servicemember Preparatory Course (FSPC) to improve recruits' physical or academic readiness, to overcome barriers to accession, and set up

participants for success in service. Both Services report success in these programs. The Army Future Soldier Program has resulted in ~90% of participants increasing their AFQT score. The newer Navy Future Sailor Program has a success rate of approximately ~70%, although the Navy used a lower minimum score for entry, compared to Army. Very low-aptitude individuals have difficulty raising their scores significantly. An important criterion for the success of these programs will be the recruits' performance in basic and technical training. It has been noted that in basic training, those finishing the program tend to be placed in leadership positions, and their attrition numbers seem to be lower.

Comments/Recommendations
The DAC-MPT thanks Dr. Helland for providing a helpful understanding of the many influential factors in the recruiting environment noted above.

The DAC-MPT recommends a more detailed briefing on the FSPC at a future meeting.

R&D Milestones Brief – Dr. Tia Fechter, Acting Director, Defense Testing and Assessment Center (DTAC)

Dr. Fechter presented a birds-eye view of the projects, accomplishments, and timelines for future and ongoing R&D work by the DTAC. Further Armed Services Vocational Aptitude Battery (ASVAB) development includes new item development, scoring, and equating within computer-adapted and paper-and pencil forms; the implications of implementing calculators; evaluation of the ASVAB (e.g., alignment of ASVAB with training and high-school curriculum; and continued monitoring of need for norming efforts). Continued research on the ASVAB includes exploring sophisticated methods for adaptive testing (CAT-ASVAB), such as using Bayesian analyses for item calibration and machine learning for form assembly. ASVAB research is also dedicated to new tests and composites of tests that measure Complex Reasoning (with a tool for generating non-proprietary items), Computational Thinking (to meet the National Defense Authorization Act (NDAA) requirement), Cyber Test, and Mental Counters. Ongoing research also investigates the Joint Service TAPAS personality measure (to combine with service-specific TAPAS versions).

R&D also captures further advances in the ASVAB Career Exploration Program (CEP) (e.g., monitoring usage in schools, refreshing the Find Your Interests inventory, expanding CEP to the Pacific) and Military Compatibility Assessment (for both enlisted and officers), as well as expanding test availability on the cloud and across a wider range of devices. Many aspects of this R&D landscape were covered in the presentations that followed, which themselves are summarized in this document.

Comments/Recommendations
There are no questions, comments, or recommendations to this briefing.

Update on Committee Recommendations – Dr. Tia Fechter, Acting Director, DTAC

Dr. Fechter presented a set of presentation slides that systematically addressed prior DAC-MPT recommendations of the committee and implementation status. Specifics can be found in those slides.

Comments/Recommendations
The committee appreciates how all stakeholders DAC-MPT, AP, DTAC, and HumRRO work together; specifically the responsiveness of DTAC's close consideration of and detailed feedback on the DAC-MPT's recommendations.

There are no specific recommendations to this briefing.

Update on P&P forms – Dr. Jeff Dahlke (HumRRO)

Dr. Dahlke first set the stage by providing context for four research projects carried out in support of the P&P form development. It should be noted that the P&P mode is used far more in ASVAB Career Exploration Program (CEP) than the ETP and is administered in a group setting. The overarching goal of this research is to ensure that the P&P administrations are producing standard scores on the same dimensions as CAT-ASVAB, further supporting the validity argument for the new P&P forms. The four studies are (1) Item Response Theory (IRT) rescaling for Auto and Shop Information, (2) Paragraph Comprehension (PC) test length reduction, (3) Arithmetic Reasoning (AR) test length reduction, and (4) time limit adjustments. With these studies, note that certain design decisions made for CAT-ASVAB do not carry through for P&P ASVAB. For example, Auto Information (AI) and Shop Information (SI) are different across the two modes (in CAT-ASVAB AI and SI are scaled/calibrated/administered separately, while in P&P-ASVAB, these are administered and scored together). The presentation reviewed the necessary adjustments to P&P-ASVAB specifications, where AI and SI is converted to the reporting scale, and time limits on AR need to be reduced to allow for extra time on PC.

The proposed modification is to construct the target Test Information Functions (TIF) using modified Stocking Lord transformation procedure (m)SLP by reflecting AI and SI Test Characteristics Curves (TCCs) onto a composite scale. The yielded results showed that the expected TCCs are closely aligned with the empirical TCCs for the composite theta. The rescaled parameters were then applied to the new SI and AI forms, and the rescaled TCCs were plotted against the expected TCCs. Given the results, this approach was recommended to obtain AS-scaled P&P-ASVAB item parameters, thus increasing the alignment between results across administration modes.

The next study of interest was length-reduction analyses for PC, because the use of single item-per-passage PC items is problematic from a testing-time perspective. Single items per passage

minimize dependence within a set of items, and in fact, item independence after controlling for the construct is a key assumption of IRT. Overall, the goals of the project are to maintain acceptable reliability and ensure adequate content coverage of PC as a construct. A related goal identified was to minimize the total word count across stimulus, stem, and options. The study design fully crossed form length with weighted average IRT information (though some combinations were not possible due to depressed/insufficient test information). The outcome simulated test-retest reliability for PC and composite scores that included PC. Results suggest that 10 items would provide sufficient reliability on relevant outcome scores while covering the construct blueprint.

The third study looked at reducing the length of AR, where any time reduction could be given to PC. Current speededness results show that AR at its present length is speeded in both the ETP and CEP test settings. At this point, six new P&P 30-item AR sections had already been assembled, so rather than start over, simulations of reliability were conducted that varied the numbers of items removed. Analyses showed that reducing AR by 5 items preserved reliability and ensured content coverage. A small caveat here is that on the P&P forms, the last few items are among most difficult, so that may impact simulation results. The recommendation is to remove 5 items for newly developed AR forms, given the presented simulation results.

The fourth and final study was to identify further ways to allocate more time to the P&P ASVAB PC sections. It is very difficult to add more time in the ETP and CEP test administration settings. The main research question was whether other subtests through reduction could donate more time to PC. The analyses carried out focused on reading load, using several common reading metrics.

It was noted that the previous generations of P&P ASVAB forms generally required significantly less time as compared to the new generation (27% shorter). The current time limit for PC is 13 minutes, but in the group setting using paper and pencil, there is also an issue of balancing time, because the time limit cannot be too short (where examinees are not finishing) nor too long (where faster examinees wait impatiently for examinees who use more time to finish). Analyses suggested that a time limit of 18 minutes should be appropriate for the new PC P&P forms. Estimation of response latency was carried out; two options were provided for time limit adjustments in P&P, by either not altering other subtests, or taking 2 minutes from AS and 3 from Mechanical Comprehension (MC) in order to reallocate 5 minutes to PC.

Comments/Recommendations

The DAC-MPT was concerned about whether the choice to administer 10 items in PC will be supported by current and future item development. Fewer items increase the ‘pressure’ on existing items to have good discrimination values across the proficiency distribution to support the requisite inferences. The DAC-MPT was assured that the item development process can support the requirements.

In terms of the final recommendations relating to the P&P ASVAB, the DAC-MPT concurs with using the (m)SLP procedure for equating. Turning to the question of the new 10-item length of PC and 25-item length for AR, the DAC-MPT also concurs with the recommended lengths for these tests, given the research presented.

The AR item reduction and form revisions discussed are operating under the assumption that the ASVAB program is progressing without a calculator. The DAC-MPT recommends this point be held for future consideration, depending on the final decisions regarding the use of calculators.

Finally, the committee concurs with the proposed P&P ASVAB time limit adjustments (i.e., option B: offset the increased PC time limit by reducing time limits for AS and MC).

Form Equating Sampling Design – Dr. Jeff Dahlke (HumRRO)

Dr. Jeff Dahlke of HumRRO summarized the results of the follow-up analyses requested by the DAC-MPT in June 2024.

The simulation study examined three key questions: (a) whether using unequated standard scores from new CAT-ASVAB forms would introduce bias, relative to scores on the reference form, (b) whether the equating sample size could be reduced from 10,000 per form while maintaining results that are equally informative, and (c) whether adjustments to sample allocation across phases could improve the equating design. Simulations conducted for nine CAT-ASVAB subtests revealed that bypassing equating and using the transformation constants (TCs) from the reference form introduced bias. Namely, lower scores tended to be overestimated and higher scores underestimated, leading to qualification rate differences. Equating effectively eliminated these biases. These findings reaffirmed the conclusions from the June 2024 briefing, demonstrating that equating prevents biases inherent in unequated score distributions.

To assess the feasibility of reducing sample sizes for CAT-ASVAB equating studies, data from Forms 11–15 were reanalyzed. Equating analyses were conducted, with form-level sample sizes ranging from 500 to 10,000 in increments of 500. TCs were estimated using both form-specific and pooled equating solutions, with form-specific solutions serving as the focus for sample size evaluations, whereas pooled solutions informed phase allocation considerations. Results supported a target sample size of 6,000 examinees per form for future equating studies.

With this form-level sample size recommendation in place, modifications to the equating study design were explored to streamline administration without impacting final TCs beyond the reduced sample size. The recommended approach maintained a three-phase equating design with revised sample size targets: 500 per form in Phase 1 using pooled equating, 1,500 per form in Phase 2 using form-specific equating, and 6,000 per form in Phase 3 using form-specific

Email foswald@rice.edu · Office 713-412-2147 · Rice University Department of Psychological Sciences-MS 25
6100 Main St. · Houston, TX 77005 · workforce.rice.edu

equating for final TCs. This refined design will reduce the study duration and the number of examinees required, while ensuring strong convergence with the 10,000-per-form solution and improving score quality in Phase 3.

Comments/Recommendations
The DAC-MPT acknowledged the scientific rigor and practical implications from this work. The DAC-MPT endorses the recommended design changes, specifically the use of 500 per form in Phase 1 (pooled equating), 1,500 per form in Phase 2 (separate equating for each form), and 6,000 per form in Phase 3 (separate equating for each form).

Complex Reasoning Update – Dr. Kate Klein (HumRRO)

Dr. Klein presented an update on the development of the Complex Reasoning (CR) measure. CR is similar to other constructs (e.g., fluid intelligence) and, like those constructs, is predictive of both training and job success. However, CR is a component that was lacking from the current accession testing program. In September 2024, a CR assessment, using four static forms, has been launched on the ASVAB platform. It is available at the MEPS to enlisted applicants. As of Nov 4, 2024, ~10K applicants have taken the assessment.

In continuing development work, five new 24-item forms were created from a test blueprint. The goal is to assess feasibility and inform CAT design. Wave 1 targeted 5,250 participants (about 1,050 per form) who were similar to the enlisted applicant population (e.g., prior non-military, ages 18-35, US citizens, and HS degree/GED/< 1 year of college). Waves 2-4 will pilot test and calibrate 288 new CR items for potential inclusion on the ASVAB platform. Test administration incorporates attention-check items and response latencies to identify those participants who might be showing insufficient effort in their responses. Multiple approaches to combining item sets are also being explored, as are multiple algorithms for scaling the test.

Comments/Recommendations
DAC-MPT recommends additional consideration of both benefits and challenges for developing an adaptive CR test. A CAT CR would allow for the possibility of creating a shorter, more reliable test, if desired. Additionally, the committee recommends focusing on equivalent distributional characteristics for items and item composites across different forms.

Computational Thinking Update – Dr. Kimberly Adams (HumRRO)

Dr. Adams began the presentation by reminding the DAC-MPT of the NDAA Congressional directive to implement an assessment measuring six areas of computational thinking. The deadline for this effort was October 1, 2024. Most of the off-the-shelf computational thinking assessments are unsuitable for military use (they are tied to specific programming languages, or they are targeted to K-12 classroom environments and freely available on the Internet).

When the NDAA legislative requirement was evaluated, the decision was made to develop a new composite using a combination of existing tools and new Computational Reasoning assessment described previously. To do so, the first task is to define the computational thinking score equation, and then the second task is to verify the validity of resulting scores. In Phase 2 work, 'shippers' were used (individuals who are recruited but not yet at Basic Training, who took the Cyber Test in order to be in the sample). Three possible computational thinking equations were developed, where all of them apply double-weight to CR. The fact not all applicants take the Cyber Test was a natural barrier to full representation of all Service branches in the sample for this particular study.

A criterion measure of computational thinking consisted of 15 computational thinking test items and 8 Bebras items. All three equations proposed were strong predictors of the computational thinking construct. Software updates to implement the composites have been completed. Thus, the legislative goal of implementing a computational thinking composite has been met.

In terms of future research, to further understand the predictive validity of computational thinking scores, the training relevance survey might be used to identify those occupational specialties with high computational thinking applicability.

Comments/Recommendations
DAC-MPT recommends continuing to evaluate test validity as the composite goes operational, as well as the time requirements for taking the test. This and other data should continue to guide decision-making even after the test is implemented.

Overall, the DAC-MPT acknowledges their confidence in this work, especially under such a tight timeline. Future work in this area is encouraged to help ensure that the computational thinking scores are meeting their intended purposes.

Calculator Analyses Efforts

a. Calculator Impact Study – Dr. Kevin Bradley (HumRRO)

Dr. Kevin Bradley (HumRRO) presented an update on studies involving the impact of the use of a calculator on two of the ASVAB subtests: Arithmetic Reasoning (AR) (k = 30 items) and Math Knowledge (MK) (k = 25 items). The four research questions that were addressed have been presented previously:

- Research Question 1: Does calculator availability meaningfully impact the dimensionality of AR and MK subtests?
- Research Question 2: Do psychometric properties differ based on calculator availability?
- Research Question 3: Does calculator availability impact subgroup performance differences?

- Research Question 4: Does calculator availability impact the amount of time needed to complete each math subtest?

Results from Research Question 1 indicated that allowing the use of calculators did not meaningfully impact the underlying dimensionality of the subset items selected from the AR and MK subtests. The presented results indicated that the use of calculators made AR items easier but had little impact on MK scores. These results, specifically AR, imply that if calculators are to be allowed on the ASVAB going forward, test equating would be necessary to maintain the required interpretability of AFQT scores. Consequently, and critically, potential mean differences involving calculator use would disappear.

In terms of Research Question 2, the reliability of the examination was not impacted, where reliability was indexed both by coefficient alpha from classical test theory and by using a marginal item response theory index. The effects of calculators on scores and item difficulty parameters were primarily linear, and the conditions could be linked through linear rescaling procedures applied to either scores or item parameters to maintain the interpretability of standard and composite scores. This finding is likely limited to the fixed linear forms used in this study and would not be generalizable to CAT-ASVAB forms. There was also no notable impact on item discrimination. Few items did item discrimination parameters in the Calculator condition that were outliers. As previously identified, equating would be an essential component of introducing calculators to operational ASVAB testing, resulting in no systematic advantage gained by examinees from using calculators.

In terms of Research Question 3, all groups benefitted similarly from the use of a calculator. In terms of Research Question 4, the use of a calculator shortened the test length time for the 30 selected AR items (by a few minutes on average) but did not meaningfully shorten the test length for the 25 selected MK items.

The main study limitation was the use of only 55 items from approximately 10,000 available MK and AR items, thus raising the need for testing a larger set of items to extend the generalizability of the findings. The study was also not able to generalize to other testing formats, such as the CAT-ASVAB administration or other fixed-length linear forms. Rescaling would also need to be performed on the basis of larger samples of both examinees and items, resulting in a universal scale transformation for item parameters across all testing forms. If calculators were used, an equating study would need to be performed for both computer adaptive and paper and pencil formats. Finally, given that around 10,000 AR and MK items have been developed, scaled and calibrated under a no-calculator condition, they would then need to be rescaled under the calculator condition.

(continued)

Email foswald@rice.edu · Office 713-412-2147 · Rice University Department of Psychological Sciences–MS 25
6100 Main St. · Houston, TX 77005 · workforce.rice.edu

Comments/Recommendations
The DAC-MPT finds the presentation very informative. Given that research results showed no scoring advantages for the use of calculators, coupled with operational downsides, the DAC-MPT recommends against using calculators on the current versions of AR and MK.

If this effort is pursued further, DAC-MPT advises that there may be some cases in which a linear transformation does not hold, and therefore both linear and nonlinear approaches should be examined.

Calculator Analyses Efforts

b. CAT Simulation – Dr. Glen Heinrich-Wallace (HumRRO)

Dr. Glen Heinrich-Wallce of HumRRO summarized a simulation study that evaluated what might happen to CAT-ASVAB composite score distributions after rescaling AR and MK item parameters to account for the impact of calculators on latent ability distributions. Specifically, the study addressed the following research questions: (a) How do empirically informed, copula-based deflections to item parameter estimates affect composite score distributions for CAT-ASVAB? (b) How do biased difficulty parameter deflections affect composite score distributions for CAT-ASVAB? (c) If the previous effects are present, which composites and which ranges of those score distributions are most affected?

The study concluded that measurement precision decreased across all conditions due to added error in parameter estimates. Higher-ability simulees generally showed greater measurement error and were more likely to be under-classified. The impact of calculator use on composite precision varied by Service, depending on the weighting of AR and MK in their classification composites.

Comments/Recommendations
There were no comments or recommendations.

Calculator Analyses Efforts

c. Calculator Needs Assessment – Dr. Monica Gribben (HumRRO)

Dr. Monica Gribben (HumRRO) presented an update on the math/calculator needs and requirements assessment. The analysis was designed to determine whether a test(s) assessing math content using a calculator is needed across jobs or for certain jobs. If such test(s) are needed, the impact to the blueprint of the examination(s) needs to be considered.

Needs assessment was conducted across eight areas from both training and on-the-job perspectives: (a) electrical, (b) infantry and combat, (c) information technology, (d) intelligence, (e) logistics and administration, (f) mechanical, (g) medical and (h) science and technology.

Based on the sample of training courses and occupations in the needs assessment, it was determined that there were no types of math where calculator use is a prerequisite for successful job performance across all occupations. However, the use of a calculator may benefit few specific occupational areas: (a) logistics and administration, (b) science and engineering, and (c) medical.

The needs assessment included a target sample from a range of occupations, but it was noted that limited participation in specific occupational areas and limited participation in at least one of the Services resulted in a final sample that was not as robust as originally planned. Efforts are underway to expand the sample.

Comments/Recommendations
DAC-MPT finds the presentation to be well-done and informative. DAC-MPT notes that the dialogue had nicely encompassed many perspectives around the costs and benefits of calculator use. Understanding the perceived need for a calculator is appreciated, even in cases where the data show it is not an actual need.

If this area of study is pursued further, DAC-MPT recommends:

1. The Department consider if/how beneficial it would be to create a dedicated/special purpose test requiring the use of a calculator for a relatively small number of occupations.
2. Continue the examination on the consistency between (a) what is taught in school and how it is assessed on the ASVAB and (b) what is required in training/on the job. That would include focusing on courses taken by students who are more inclined to join the military.
3. Practical considerations include, but are not limited to, the need to have calculators that are readily available, fully functioning, and clean.

Refinement of the Joint Service-TAPAS Instrument – Dr. Dan Putka (HumRRO)

Dr. Putka provided a review of the Joint-Service TAPAS (JS-TAPAS) instrument. The underlying concept of the JS-TAPAS is a modular approach, consisting of a set of core joint facets and several Service-specific facets. Joint facets are selected to assess military compatibility and enlisted eligibility. Focusing on the Military Compatibility Composite, the goal is the prediction of misconduct. For the Enlistment Eligibility composite, the analysis was focused on first-term enlisted job performance. Services also have the flexibility to use Service specific facets for their individual objectives.

A phased development approach has been implemented, with Phase 0 implemented at MEPS in September 2024, and Phase 1 underway, currently targeted for implementation in FY27. The work in Phase 1 includes both content and psychometric efforts, to update TAPAS statement

pools, calibrate the pools, and develop provisional norms. Following that is Phase 2, with evaluation and refinement of Phase 1, which includes amassing an evidentiary basis for the TAPAS scores.

The research carried out suggested that a total of 17 facets can be assessed in the JS-TAPAS, representing a balance between JS facets and service-specific facets, as well as balancing between having more items (to get higher reliability) vs. fewer items (to shorten administration times). A comprehensive review revealed that 12 facets should be included for use in the JS composites, and another 5 slots in the TAPAS administration should be reserved for Service-specific facets. Future research includes examining practice/coaching effects, AI for development purposes, and the feasibility of TAPAS using machine learning to predict attrition.

Comments/Recommendations
The DAC-MPT was asked about the extent to which JS-TAPAS and cognitive test scores used for high-stakes decisions should be held to different reliability standards. In response, the overall recommendation is to focus on the purpose and use. Specifically, DAC-MPT advises to focus on the reliability at the more specific points or range where selection decisions tend to be made, and not worry about large errors on the extremes.

The DAC-MPT offered additional comments for consideration:

- Because the purpose of the test is selection, reliability statistics might not be as useful as the standard error of measurement (SEM) and the range of the scores.
- Whether operational settings will recommend for or against the move toward composites.
- Ways to identify people who are cheating, or people who have been coached, or people who are disseminating items.

Adverse Impact – Dr. Nick Howald (HumRRO)

Dr. Nicholas Howald (HumRRO) presented an update on the assessment of adverse impact from the FY2023 applicant sample. Adverse impact analyses were performed on the ASVAB AFQT (IIIA+ and IIIB+), ASVAB subtests, Cyber Test, and Coding Speed. Findings were consistent with prior years, varying from negligible to fairly large (in some cases) across tests and across groups being compared, in line with other national testing programs like the SAT and NAEP. It was also pointed out that people self-selecting to enlist in the Armed Services may tend to differ from college-bound SAT test-takers in terms of motivation, personality, knowledge, and other individual differences. With respect to the special tests (Cyber Tests and Coding Speed), small-to-moderate subgroup mean differences were found and were typically smaller than the effects of adverse impact for the ASVAB tests.

(continued)

Email foswald@rice.edu · Office 713-412-2147 · Rice University Department of Psychological Sciences-MS 25
6100 Main St. · Houston, TX 77005 · workforce.rice.edu

Comments/Recommendations
DAC-MPT compliments the charts presented as being easy to understand. Furthermore, DAC-MPT advises on a possible alternative way for looking at the adverse impact of composites by looking at relative sizes of groups in the general population.

DAC-MPT believes the presented analyses were solid, suggesting no further additions to this line of work.

Curriculum Alignment Study – Dr. Rodney McCloy (HumRRO)

Dr. McCloy reviewed the results of the current study regarding how ASVAB subtests align with course content taught in high schools. The presentation summarized previous high school curriculum and high school assessment alignment studies with ASVAB content, provided mappings between the ASVAB and other tests, highlighted the National Assessment of Educational Progress (NAEP) transcript studies, and identified differences between the courses taken by military applicants and the general high school population.

This effort revealed:

- The ASVAB content is largely addressed in relevant testing frameworks (e.g., ACT, NAEP), although some suggestions for additions to test blueprints were noted.
- Addressing some skills would require expansion of item types, which is problematic given the time constraints on the ASVAB.
- ASVAB academic content areas (e.g., GS, AR/MK) are typically addressed in high school courses, but the technical content coverage of the ASVAB is uneven.
- There is some indication of differences in high school courses being taken between propensed and non-propensed youth, with the latter generally taking higher-level courses.
- There are also some indications that propensed youth more likely to take part in extracurricular activities relevant to ASVAB (e.g., automotive, construction).

Comments/Recommendations
DAC-MPT provides a series of questions for open consideration, which include: (a) What changes to the ASVAB might make it more effective at assessing knowledge gained through modern educational trends (e.g., integrated or flipped instruction)? (b) How do ROTC students differ from high school students in their course patterns? (c) Should the ASVAB include other more open-ended or project-based question formats, despite the challenges in scoring and implementation? (d) Would you consider involving educators, recruiters, and SMEs in workshops to refine crosswalks and ensure content reflects both school and military needs? (e) Would you consider conducting follow-up studies on test-takers to evaluate how changes influence career trajectories and military readiness?

The DAC-MPT recommends further research in the value and use of new item formats/types. Specifically, can new item types change the nature of what is assessed, even if the items remained multiple-choice?

ASVAB CEP

a. General – Dr. Irina Rader (OPA/DTAC)

Dr. Irina Rader comprehensively updated the ASVAB Career Exploration Program (CEP) usage metrics, reflecting year-to-date participation and key performance indicators. As of the latest reporting period, 13,105 schools participated in the program, 619,926 students completed the ASVAB assessment, and of those, 339,463 leads were generated for military Services.

A notable trend over the past five years is the transition from paper and pencil assessments to CEP iCAT assessments. This underscores the program's shift toward more modern, technology-driven methods. The recent CEP Jamboree was a three-day strategic planning session involving stakeholders from the following organizations: DTAC, AP, Office of People Analytics (OPA), DTAC, and U.S. Military Entrance Processing Command (USMEPCOM). This event focused on reviewing the previous year's accomplishments, identifying areas for improvement, and setting strategic priorities for School Year (SY) 24–25 and beyond.

Business Strategy Goals for SY24/25 focused on the following:

- Technology: By August 2025, migrate the ASVAB Program and Careers in the Military (CITM) websites into the DPACS boundary to enhance security and functionality.
- Research & Innovation: Advance research efforts to strengthen occupational crosswalks and integrate innovative practices in continuing to address stakeholder needs.
- Occupational Data & Content: Establish a transparent Occupational Crosswalk Process and explore using artificial intelligence (AI) to enhance data collection and analysis.
- Promotion & Engagement: Implement the SY24/25 social media Strategic Plan to increase program visibility and expand ASVAB CEP's digital outreach.
- Workforce Multiplier—Continue to expand the PTI training program, including updates to the training content and tracking; leverage strategic partnerships with U.S. Army Recruiting and Retention College leaders, JROTC, and MEPS Battalion Commanders.
- State Legislative Activities—Continue tracking state and federal legislation and development of interactive mapping and visualization tooling.
- Underserved Populations—Create a pilot program to increase private and homeschool testing as well as post-secondary institution participation.

This multifaceted strategic approach modernizes program operations, supports student career exploration, and strengthens collaboration with key stakeholders.

(continued)

Comments/Recommendations
The DAC-MPT appreciates the impressive and wide-ranging efforts and impact of the ASVAB CEP team and program, and the recognition they are receiving as a result. No additional recommendations for this section are provided.

b. Find Your Interest – Dr. Rod McCloy (HumRRO)

Dr. Rod McCloy provided an update to the DAC-MPT on the development of the new form of the Find Your Interest (FYI) Inventory. Three potential forms were considered, and the analysis endorsed the final version based on several key factors:

- Internal consistency reliability: Estimates for the final version were at high and reasonable levels.
- Subgroup differences: The final version showed minor subgroup differences, despite items not explicitly chosen with this criterion in mind.
- Comprehensive coverage: The final version fully represented fundamental interest areas.

Next steps include (a) dimensionality analyses (item-level exploratory and confirmatory factor analysis, including standard models, as well as circumplex models found in the vocational interests literature), (b) field testing and analysis (administer the new form and analyze the results), and (c) development of norms (to support its practical application). These steps ensure the new FYI form is psychometrically sound and effectively aligned with student career exploration needs.

Comments/Recommendations
DAC-MPT supports the move to basic interests, noting that it merges with the basic interests research of Dr. Rong Su, along with other research that aligns basic interests measurement with O*NET. This move retains the traditional RIASEC dimensionality of vocational interests yet allows more refined better career exploration. The more information about jobs that the respondent has, the more differentiated their basic interests can become on the measure. Furthermore, the basic interests approach can be a good sell to States, because they are often interested in their workforces, and a basic interests test is more aligned with these goals (vs. a more general RIASEC test). DAC-MPT does not have concerns with proceeding with Version 3.

DoD was interested in the DAC-MPT feedback on the need for sex-based norms, given that the inventory will not be used for selection or classification purposes; that is, it will not be used to compare individuals against others, but only to provide an opportunity for self-exploration. DAC-MPT recommends a full norm-based tool that allows users to select their subgroup characteristics to be the best way forward, rather than providing un-normed results.

(continued)

Email foswald@rice.edu · Office 713-412-2147 · Rice University Department of Psychological Sciences-MS 25
6100 Main St. · Houston, TX 77005 · workforce.rice.edu

c. Work Values – Dr. Maura Burke (HumRRO)

Dr. Maura Burke briefed the DAC-MPT and began with the team's initiative for creating a scenario-based Work Values Situational Judgment Activity (WV SJA) to introduce students to work values and provide a work values tool for exploration that would empower students and counselors in their career journey.

The project sought to differentiate the ASVAB CEP program from other career exploration programs by offering unique resources for students and counselors, including the work values inventory to help students reflect on what they value in a job. They have added resources that both students and counselors could use on their website. The WV SJA would be positioned as a tool for student exploration: to learn about work values, to understand where one stands in terms of work values, and to discuss work values further with a counselor. By design, the WV SJA is not judgmental or prescriptive about what a student should value.

Three different measurement formats were investigated.

- Ipsative inventory: Students rank their values and do not compare themselves with others, thus avoiding social desirability bias.
- Policy capture approach: Used regression-based scoring to highlight how students prioritize work values.
- Multiple-choice format: Designed for simplicity, with additive scoring for ease of classroom use.

Given time constraints and classroom needs, the multiple-choice format was chosen. The choice emphasizes accessibility, ensuring paper-and-pencil options, quick completion, and easy scoring.

The WV SJA is a situational judgment test that covers six work values across workplace and school scenarios. The assessment uses 16 scenarios overall, eight regarding work settings and eight regarding school settings. Each scenario offers six response options, and the respondent chooses the one they prefer most. Each of the six options is tied to one of the six work values that the measure assesses.

Comments/Recommendations

DAC-MPT suggests introducing more variability in contexts provided (one example given was volunteer opportunities). DAC-MPT suggested it was important to consider varied student background experiences and the subjectivity that comes with that. Additionally, the committee DAC-MPT recommends considering inclusion of post-survey questions on whether individuals had a difficult time answering the questions if they had not had a job. This may aid in future enhancements to the assessment.

Future Topics – Dr. Tia Fechter, Acting Director, Defense Testing and Assessment Center (DTAC)

Dr. Fechter facilitated the DAC-MPT discussion on what high-priority research areas they would suggest for the future, also mentioning two topics that might be considered: the development of an SJT for military compatibility and FSPC research by the Services.

Comments/Suggestions
DAC-MPT is interested in learning more about:

- TAPAS, high school curriculum work, and ASVAB CEP. These topics can help contribute to informing new learning pedagogies.
- Effectiveness of FSPC.
- Update on the FYI and the normative information that is collected.
- Progress on the WV SJA situational judgment test.

In terms of presentation flow, DAC-MPT is interested in four cross-cutting themes that could guide future DAC-MPT meetings:

- First, have presenters design presentations with their stakeholders in mind, making sure they can access the information that they need to see.
- Second, orient presentations on the effects of tests on operational decision making.
- Third, present a one-page roadmap or crosswalk that bridges the presentations (reflected in the R&D brief) to broader operational and strategic goals (reflected in the AP brief).
- Fourth, include any developments in AI that impact measurement, particularly for the ASVAB and TAPAS.

Summary

This DAC-MPT meeting covered a wide range of research projects, anchored to the core mission of continuously improving the strength of US military talent across all the Services. DAC-MPT members find that rigorous science and research efforts are evident across all DoD-sponsored projects that were presented. The projects seek not only to maintain the high quality of the current ASVAB used operationally, but also to adapt and extend it further for the military's future needs. Projects make use of cutting-edge scientific findings in the testing literature while incorporating responsible and innovative test practices.

We appreciate the efforts of AP, DTAC, HumRRO consultants, MEPCOM, the research staff of each of the Services, and so many unseen personnel who use these instruments daily, to keep us informed on the progress of tests and testing through these meetings. In particular, the Committee appreciates the depth of expertise, experience, and planning that Dr. Velgach brings into each meeting.

Email foswald@rice.edu · Office 713-412-2147 · Rice University Department of Psychological Sciences-MS 25
6100 Main St. · Houston, TX 77005 · workforce.rice.edu

RICE

Department of Psychological Sciences
Fred Oswald
Professor and Herbert S. Autrey Chair in Social Science

As always, the members of the DAC-MPT support these research efforts and continue to believe that the ASVAB/accession testing program is a critical component of military recruitment that strengthens the effectiveness of the military forces. We look forward to our next meeting.

Sincerely,

Fred Oswald
Professor and Herbert S. Autrey Chair in Social Sciences
Department of Psychological Sciences
Rice University

Tab E

Military Personnel Policy (Accession Policy)

SLIDES ONLY

NO SCRIPT PROVIDED

Dr. Katherine Helland
Director, Accession Policy

January 22, 2025

As of: 16 Dec 24 (v1)

THEN AND NOW

Where we were – 2023

- The most difficult year since the inception of the All-Volunteer-Force and the first time since 1979 that three active components failed their recruiting goals.
- Only Marine Corps and Space Force met their recruiting accession missions.

Fiscal Year 2024	Active Component 2023 Recruiting/Accession Data			
	Annual Goal	Fiscal Year Achieved	Fiscal Year Percent of Goal	
Army	65,500	50,181	76.61	R
Navy	37,700	30,236	80.20	R
Marine Corps	28,900	28,921	100.07	G
Air Force	26,977	24,100	89.34	R
Space Force	492	537	109.15	G
Total	159,569	133,975	83.96	

Where we are - 2024

- All components, except for the Navy Active Duty, Army Reserve and Navy Reserve met their 2024 recruiting accession missions.
- Navy made its contracting goals yet fell short of shipping all 40,600 due to basic training capacity limitations.
- Services FY25 DEP was 10% higher than FY24 start

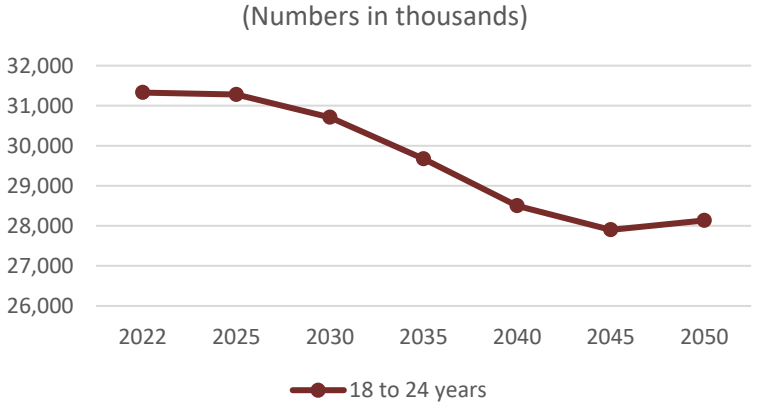
Fiscal Year 2024	Active Component 2024 Recruiting/Accession Data			
	Annual Goal	Fiscal Year Achieved	Fiscal Year Percent of Goal	
Army	55,000	55,150	100.27	G
Navy	40,600	35,804	88.19	R
Marine Corps	27,500	27,500	100.00	G
Air Force	27,200	27,303	100.38	G
Space Force	704	716	101.70	G
Total	151,004	146,473	97.00	

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

CURRENT AND FUTURE MARKET DYNAMICS

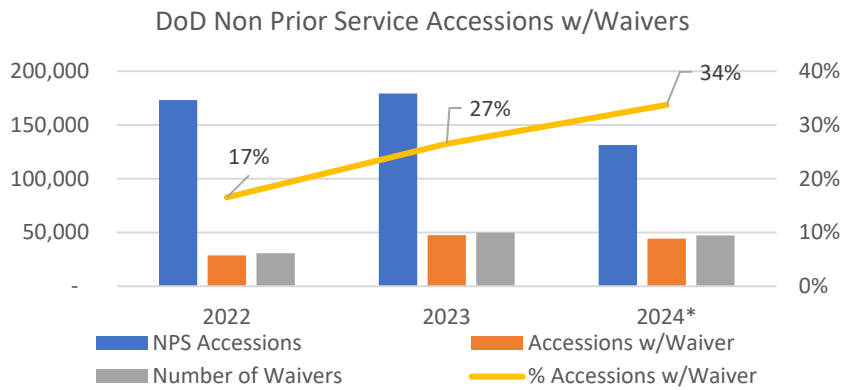
Projected Youth Population

Due to decreasing birth rates starting in 2008, the 18-24 youth population will begin declining in 2026



(Source: U.S. Census Bureau, Population Division)

Waivers

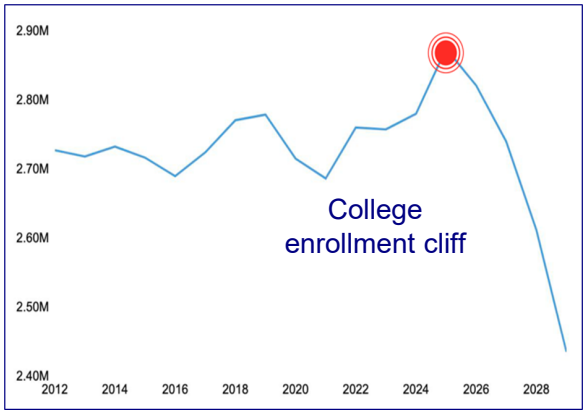


FY24 data as of end of 3rd Qtr

(Source: OSD Accession Policy FY22-24 Waiver Report)

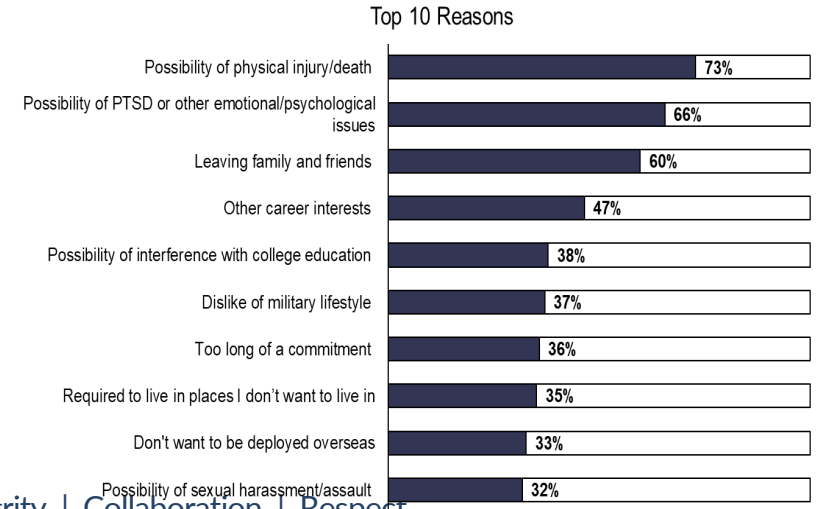
Decreased College Attendance

Corresponding with the youth population decline, college enrollment is also expected to peak in 2025 followed by a sharp



(Source: Nathan Grawe's Higher Education Demand Index)

Barriers to Service: Reasons Not to Join



(Source: DoD JAMRS Youth Poll (Fall 2023)

STRATEGIC MITIGATION EFFORTS

LINES OF EFFORT TO IMPROVE THE ACCESSION PIPELINE

Growing Propensity

Objective: Increase awareness, consideration and motivation to serve

Initiatives

1. Launch of JAMRS adult influencer media campaign and youth digital media campaign with several TV/streaming commercials airing from 30 September-November 10, 2024. Adult influencers who see at least one JAMRS ad are 47% are likely to recommend military service.

2. Developing a standardized methodology to provide states with military affiliation data to include military readiness into their education accountability plans. This will incentivize school officials to promote benefits of military service.

3. Continuing to work legislative proposals to improve quality access.

4. Coordinating and collaborating with industry, academia, non-profits, the military, and across government to operationalize permeability and grow interest in public service.

Expanding Eligibility

Objective: Expand the aperture for those interested in serving

Initiatives

1. Medical Accessions Records Pilot (MARP) expanded from 38 to 51 conditions. Recently added: Asthma in the last 4 years, ADD/ADHD time adjustments, and learning disorders within one year.

2. Exploring the feasibility of alternative medical accessions standards frameworks based upon updated information, medical advances, and a range of possible assumptions.

3. Developed Joint Enlistment Composite for current non-cognitive personality test (TAPAS). As a next phase, the Joint Enlistment Composite will be leveraged to redefine applicant quality and expand the pool of eligible applicants by adding personality into the definition of quality.

4. Developed ASVAB special purpose test: New assessment of fluid intelligence called Complex Reasoning is now available to the Services. Complex reasoning is less reliant on traditional academic knowledge and proficiency of the English language. Future objective: evaluation for inclusion into AFQT.

Discussion/Questions

Questions?

Tab F

OPA

OFFICE OF PEOPLE ANALYTICS

Major ASVAB R&D Efforts

Milestones and Project Schedules

Mary Pommerich

Defense Testing and Assessment Center

Briefing presented to the DACMPT

January 22, 2025

Projects

- **ASVAB Development**
  - Item Development Efforts
  - New CAT-ASVAB Item Pools
  - New P&P-ASVAB Forms*
  - Evaluations of CAT-ASVAB and Form Development Methodologies*
  - Implementation of Calculators*
- **ASVAB and Enlistment Testing Program (ETP) Revision**
  - Next Generation ASVAB/ASVAB Evaluations
    - Adverse Impact Analyses*
    - Differential Prediction Analyses
    - Training Relevance Survey
    - High School Curriculum Study*
    - Focus Groups
    - Validity Frameworks
    - Norming Investigations

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

Projects

- **ASVAB and ETP Revision**
  - Evaluating New Cognitive Tests/Composites for ASVAB
    - Complex Reasoning*
    - Computational Thinking*
    - Cyber Test
    - Mental Counters
  - Adding Non-Cognitive Measures to Selection and/or Classification
    - TAPAS Validity Framework
    - Joint-Service TAPAS*
- **Career Exploration Program***
- **Military Compatibility Assessment**
- **Expanding Test Availability**
  - Web/Cloud Delivery of ASVAB and Special Tests
  - Device Expansion

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

Item Development Efforts

- **Objective**
  - Develop new items for the ASVAB
- **Projected Completion**
  - Ongoing
- **Subtasks**
  - Develop items and graphics for GS, AR, WK, PC, MK, EI, AI, SI, and MC subtests using the ASVAB item bank platform (ongoing)
  - Conduct copy edits (ongoing)
  - Conduct sensitivity reviews, content review, and content edits (ongoing)
  - Develop training materials for the ASVAB item bank (ongoing)
  - Convert AO item graphics for alternate device compatibility (PiCAT, APT, and special tests complete; others in progress)

Item Development Efforts

- **Predecessors**
  - Development of ASVAB item bank
- **Successors**
  - Item tryouts on CAT-ASVAB platform
  - CAT-ASVAB form development

New CAT-ASVAB Item Pools

- **Objective**
  - Develop CAT-ASVAB item pools from new items
- **Projected Completion**
  - CAT-ASVAB Forms 16–20 implementation: TBD
- **Subtasks**
  - Conduct item tryouts ✓
  - Conduct calibration and scaling ✓
  - Conduct item analyses and screening (in progress)
  - Assemble item pools and prepare for online administration (TBD)
  - Collect Initial Operational Test & Evaluation (IOT&E) data (TBD)
  - Conduct final equating analyses and evaluations (TBD)
  - Update software for operational implementation (TBD)
  - Implement Forms 16–20 operationally (TBD)

New CAT-ASVAB Item Pools (cont.)

- **Predecessors**
  - ASVAB item development
  - CAT-ASVAB Forms 11–15 implementation (Feb 2024) ✓
- **Successors**
  - Use of CAT-ASVAB Forms 5–9 in CEP, AFCT, PiCAT, overseas ✓
  - Documentation of CAT-ASVAB Forms 11–15 development in a technical bulletin (May 2024) ✓

New P&P-ASVAB Forms*

- **Objective**
  - Develop P&P-ASVAB Forms 29F/G, 30F/G, 31F/G, 32F/G from new items
- **Projected Completion**
  - New form implementation: TBD
- **Subtasks**
  - Conduct item tryouts ✓
  - Conduct scaling ✓
  - Assemble test forms ✓
  - Resolve issues with PC, AS, and AO (in progress)
  - Assemble test booklets (TBD)
  - Print test booklets (TBD)
  - Update scanning and scoring systems (TBD)
  - Prepare for operational implementation (TBD)
  - Implement P&P-ASVAB Forms 29–32 operationally (TBD)

New P&P-ASVAB Forms* (cont.)

- **Predecessors**
  - Development of CAT-ASVAB Forms 11–15
- **Successors**
  - Implementation of new P&P-ASVAB forms in the ETP
  - Implementation of new P&P-ASVAB forms in the CEP
  - Documentation of P&P-ASVAB Forms 29–32 development in a technical bulletin

Evaluation and Implementation of Calculators*

- **Objective**
  - Move forward with incorporating calculator use on the ASVAB
- **Projected Completion**
  - TBD
- **Subtasks**
  - Identify calculator-sensitive items via SME review* (Feb 2024) ✓
  - Conduct data collection and evaluate calculator impact* (Sep 2024) ✓
  - Conduct needs assessment with Services* (Feb 2025)
  - Determine next steps† (TBD)
- **Predecessors**
  - Development of MK and AR items and CAT-ASVAB pools

† Next steps will depend on findings in the initial phases and on receipt of funding.

Evaluation and Implementation of Calculators* (cont.)

- **Successors**
  - Possible equating (if no multidimensionality or other measurement issues are found in impact analysis)
  - Possible new MK, AR, or special purpose test (if multidimensionality or other issues are identified in impact analysis)

Evaluation of CAT-ASVAB and Form Development Methodologies*

- **Objective**
  - Evaluate CAT-ASVAB methodologies and ways to streamline form development efforts
- **Projected Completion**
  - TBD
- **Subtasks**
  - Evaluate modernized options for CAT-ASVAB and technical approaches to test administration and scoring (Mar 2023) ✓
  - Explore the efficacy of using machine learning methods to streamline the form assembly process (Sep 2023) ✓
  - Develop and carry out a Bayesian item calibration sample size reduction study (June 2024) ✓
  - Conduct evaluation of Differential Item Functioning approaches (Apr 2025)
- **Predecessors**
  - Transition form development responsibilities from the government to the contractor

Evaluation of CAT-ASVAB and Form Development Methodologies* (cont.)

- **Successors**
  - Possible adjustments to item seeding practices
  - Possible adjustments to item calibration practices
  - Possible adjustments to item analysis practices
  - Possible adjustments to form assembly practices
  - Possible adjustments to form equating practices

Next Generation ASVAB/ASVAB Analyses*

- **Objective**
  - Evaluate state of the ASVAB and prepare for the next generation of ASVAB and special purpose tests to be administered on the ASVAB platform in the ETP
- **Projected Completion**
  - Ongoing
- **Subtasks**
  - Conduct training relevance survey (June 2022) ✓
  - Conduct differential prediction analyses for the AFQT (Mar 2023) ✓
  - Conduct focus groups with stakeholders (Mar 2023) ✓
  - Refine validity argument for the AFQT (Sep 2023) ✓
  - Develop and implement a plan for annually evaluating the need for re-norming the ASVAB (Sep 2023) ✓
  - Conduct a norming needs assessment (Mar 2024) ✓

Next Generation ASVAB/ASVAB Analyses* (cont.)

- **Subtasks** (continued)
  - Conduct differential prediction analyses for Service-specific classification composites (Sep 2024) ✓
  - Conduct adverse impact analyses for the ASVAB and special tests on FY 2023 applicants* (Oct 2024) ✓
  - Conduct Next Generation ASVAB workshop (Nov 2024)
  - Conduct high school curriculum study* (Dec 2024)
  - Develop a roadmap for next generation ASVAB (Apr 2025)
  - Refine validity argument for the ASVAB (Sep 2025)
- **Predecessors**
  - Prior revisions to ASVAB contents
  - Evaluations of special tests of interest
- **Successors**
  - Possible revisions to ASVAB contents

Evaluating New Cognitive Tests: Complex Reasoning*

- **Objective**
  - Develop a non-verbal reasoning assessment and evaluate for possible inclusion in the ASVAB
  - Develop an item generator for Complex Reasoning
- **Projected Completion**
  - Oct 1, 2024, to meet NDAA Computational Thinking requirement ✓
  - Ongoing evaluations after NDAA milestone is met
- **Subtasks**
  - Evaluate existing, non-proprietary item generator ✓
  - Generate items and conduct pilot study ✓
  - Finalize Complex Reasoning Item Generation Tool ✓
  - Generate items for a follow-up pilot study and evaluation of refined Complex Reasoning capability ✓
  - Conduct follow-up pilot study (Sep 2023) ✓
  - Provide test forms, tryout items, and programming requirements for implementing on ASVAB platform (Sep 2023) ✓

Evaluating New Cognitive Tests: Complex Reasoning* (cont.)

- **Subtasks** (continued)
  - Provide research and development/maintenance plan (Mar 2024) ✓
  - Program Complex Reasoning application for administration on the ASVAB test delivery platform and conduct QA (Aug 2024) ✓
  - Implement Complex Reasoning test operationally (Aug 2024) ✓
  - Develop and pilot new items (Jan 2025)
  - Develop CAT pools and conventional forms (Jun 2025)
  - Identify refinements for blueprints, item generation, and form assembly procedures (TBD)
  - Design studies to evaluate Complex Reasoning scores (TBD)
- **Predecessors**
  - Evaluation of Abstract Reasoning Test
- **Successors**
  - Evaluate operational Complex Reasoning scores
  - Possible operational implementation of a CAT version of Complex Reasoning test

Evaluating New Cognitive Tests: Computational Thinking*

- **Objective**
  - Develop a Computational Thinking composite score to meet National Defense Authorization Act (NDAA) requirement to address Computational Thinking skills
- **Projected Completion**
  - Oct 1, 2024, to meet NDAA Computational Thinking requirement ✓
  - Ongoing evaluations after NDAA milestone is met
- **Subtasks**
  - Conduct alignment study to establish Computational Thinking composite(s) from existing ASVAB and military tests (Dec 2023) ✓
  - Program necessary modifications to Complex Reasoning application to compute Computational Thinking composite scores and QA (Aug 2024) ✓
  - Implement Computational Thinking composite scores operationally (Aug 2024) ✓
  - Conduct empirical validation study of Computational Thinking composite scores using Shippers (Sep 2024) ✓
  - Design studies to evaluate Computational Thinking scores (TBD)

Evaluating New Cognitive Tests: Computational Thinking* (cont.)

- **Predecessors**
  - Development and evaluation of Cyber Test
  - Development and evaluation of Complex Reasoning test
- **Successors**
  - Evaluation of operational Computational Thinking scores
  - Possible inclusion of Computational Thinking composite scores into selection and classification decisions

Non-Cognitive Measures for Selection & Classification: Joint-Service TAPAS*

- **Objective**
  - Integrate the use of non-cognitive measures in the military selection and classification process to open the aperture and widen diversity of military service eligible applicants
- **Projected Completion**
  - Ongoing
- **Subtasks**
  - Establish a validity argument for TAPAS ✓
  - Refine the validity argument for TAPAS (ongoing)
  - Develop a joint-Service version of TAPAS (JS-TAPAS)
    - Explore TAPAS effect on adverse impact (May 2023) ✓
    - Develop research plan for a criterion data collection (Summer 2023) ✓
    - Make recommendations for an interim joint-Service composite (Fall 2023) ✓
    - Add joint-Service facets to Service-specific TAPAS versions (where needed) and program interim (Phase 0) joint-Service composite (Sep 2024) ✓
    - Implement JS-TAPAS and Phase 0 composite operationally (Sep 2024) ✓
    - Evaluate initial composite (TBD)

Non-Cognitive Measures for Selection & Classification: Joint-Service TAPAS* (cont.)

- **Subtasks** (continued)
  - Select Service-specific facets for inclusion in JS-TAPAS (Service task)
    - Army (Fall 2024) ✓
    - Air Force/Space Force (Fall 2024) ✓
    - Marine Corps (Fall 2024) ✓
    - Navy (Fall 2024) ✓
- **Predecessors**
  - Program TAPAS for administration on the cloud platform
- **Successors**
  - Phase 1 JS-TAPAS instrument refinement
  - Phase 1 JS-TAPAS enlistment composite refinement

Non-Cognitive Measures for Selection & Classification: Military Compatibility

- **Objective**
  - Integrate the use of non-cognitive measures in the military selection and classification process to ensure military compatibility among enlisted and officer populations
- **Projected Completion**
  - Ongoing
- **Subtasks**
  - Identify TAPAS-based compatibility composite for initial operational test and evaluation (IOT&E) with applicants (Sep 2023) ✓
  - Identify possible alternate compatibility composites (Sep 2023) ✓
  - Make software changes to TAPAS/update facets and implement on cloud platform (Sep 2024) ✓
  - Program interim (Phase 0) military compatibility composite for enlisted personnel and implement on cloud platform (Sep 2024) ✓
  - Begin data collection for enlisted population (Sep 2024) ✓
  - Develop plans for evaluating applicable tests and feasibility of clinical/holistic evaluation for enlisted accessions (Sep 2024) ✓

Non-Cognitive Measures for Selection & Classification: Military Compatibility (cont.)

- **Subtasks** (continued)
  - Review avenues for non-cognitive assessment in the officer population (ongoing)
  - Evaluate IT options for implementation with officers (ongoing)
  - Conduct research on non-cognitive assessment methodologies (ongoing)
- **Predecessors**
  - Develop TAPAS Conduct composite (Army)
  - Data collections on Service-specific versions of TAPAS (Army, Air Force, Marine Corps)
- **Successors**
  - Develop and implement a roadmap for officer implementation and evaluation
  - Phase 1 JS-TAPAS instrument refinement
  - Phase 1 JS-TAPAS military compatibility composite refinement

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**
  - Develop CEP briefings and materials for external sources, as needed (ongoing)
  - Revise program materials as suggested by expert panel and evaluation efforts (ongoing)
  - Refresh the Find Your Interests (FYI) inventory items (in progress)
  - Monitor state usage of ASVAB and ASVAB CEP as related to legislative changes (ongoing)
  - Monitor CEP *i*CAT usage in schools (ongoing)
  - Update standard operating procedures for collecting and analyzing military occupational data (in progress)
  - Assess incorporating AI into occupational analysis (in progress)
  - Increase touch points with stakeholders and formalize program lead tracking (ongoing)

Career Exploration Program* (cont.)

- **Subtasks** (continued)
  - Develop, enhance, and deliver training (SY 2024–2025)
    - Website enhancements
    - Post-Test Interpretation recertification
    - Update Education Services Specialist training
    - Coordinate with Services to include ASVAB CEP in Recruiting School curriculum
    - Develop ASVAB CEP MEPS Commanders training
  - Launch new research initiative to investigate potential future program enhancement (in progress)
- **Predecessors**
  - Seeded new items for the FYI
  - Collated authoritative data sources for the OCCU-Find
  - Conducted Post-Test Interpretation training and expanded to Europe
  - Developed enhancements to ASVAB CEP website
- **Successors**
  - Expand ASVAB CEP to the Pacific

Expanding Test Availability: Device Expansion

- **Objective**
  - Expand *i*CAT test delivery application to run on additional operating systems and browsers for desktops/laptops
  - Expand PiCAT and APT to run on tablets and smartphones
- **Projected Completion**
  - Summer 2025
- **Subtasks**
  - Initiate development of implementation plan (Jun 2023) ✓
  - Update *i*CAT to run on additional browsers and operating systems for desktops and laptops (Fall 2023) ✓
  - Expand PiCAT/APT to run on 3 tablets with touchscreen capabilities (May 2024) ✓
  - Monitor operational performance across desktops/laptops and tablets (ongoing)
  - Expand PiCAT/APT to run on select smartphones (Summer 2025)
  - Implement new interface for all *i*CAT applications (Summer 2025)

Expanding Test Availability: Device Expansion (cont.)

- **Predecessors**
  - Evaluation of ASVAB performance across different devices
- **Successors**
  - Monitor operational performance across desktops/laptops, tablets, and smartphones
  - Evaluate need and feasibility of expanding delivery of special tests to additional devices

Appendix
List of Acronyms

List of Acronyms (cont.)

AFCT	Armed Forces Classification Test
AFQT	Air Force Qualification Test
AI	Auto Information
APT	AFQT Predictor Test
AR	Arithmetic Reasoning
ASVAB	Armed Services Vocational Aptitude Battery
ATO	Authority to Operate
CAT-ASVAB	Computerized Adaptive Testing version of the ASVAB
CEP	Career Exploration Program
CS	Coding Speed
CT	Cyber Test
DMDC	Defense Manpower Data Center
EI	Electronics Information
ETP	Enlistment Testing Program
FYI	Find Your Interests inventory
GS	General Science
*i*CAT	Internet-based CAT-ASVAB
*i*CAT-A&R	*i*CAT Authorization and Registration
IOT&E	Initial Operational Test and Evaluation

List of Acronyms (cont.)

MEPS	Military Entrance Processing Stations
MET	Military Entrance Test site
MC	Mechanical Comprehension
MCt	Mental Counters test
MEPCOM	Military Entrance Processing Command
MEPS	Military Entrance Processing Stations
MET	Military Entrance Test site
MK	Math Knowledge
NDAA	National Defense Authorization Act
OCCU-Find	Occupational Finder
PC	Paragraph Comprehension
P&P	Paper and Pencil
P*i*CAT	Pending Internet CAT-ASVAB
PV-ETP	Post-VTest ASVAB
QA	Quality Assurance
R&D	Research and Development
TAPAS	Tailored Adaptive Personality Assessment System
TBD	To Be Determined
VTest	Verification Test
WinCAT	Windows-based CAT-ASVAB
WK	Word Knowledge

Tab G

OPA

OFFICE OF PEOPLE ANALYTICS

Update on Committee Recommendations

Mary Pommerich

Defense Testing & Assessment Center

Briefing presented to the DACMPT

January 22, 2025

Accession Policy Briefing

DAC Recommendations (12/22)

1. Because of so many new members to the DACMPT, the Committee found the overview particularly helpful and would like to be regularly updated on Accession Policy's activities, including the challenges it faces in accomplishing its mission.

AP Response

1. Concur. Accession Policy (AP) provides a routine briefing to the DACMPT members, updating them on the current challenges and efforts to overcome challenges and continue process improvements, modernization, and innovation.

New Member Briefing

DAC Recommendations (12/22)

1. The DACMPT appreciated the detailed information and wishes to be updated on changes to the testing programs as well as the results of the research efforts being conducted.

DTAC Response

1. DTAC will work with AP to continue to keep the DACMPT apprised of relevant changes and research efforts.

Major ASVAB R&D Efforts: Milestones and Project Schedules

DAC Recommendations (12/22, 08/23)

1. [12/22] The DACMPT appreciated the scope of research on the ASVAB and other cognitive and non-cognitive measures and the efforts to improve the Career Exploration Program and delivery of tests. The Committee supported the ongoing review of current high school course content, curriculum, standards, and instructional methods to ensure that the next-generation ASVAB is aligned to high school content, particularly with respect to courses likely to be taken by individuals inclined to join the Services. The Committee requests a curated list of technical reports (and access to them as appropriate) and updates regarding progress on this research.

2. [08/23] The DACMPT appreciated the detailed information Dr. Pommerich provided and wishes to be updated on the results of the research efforts being conducted and the plans for new research. The DACMPT also recommends that DTAC monitor developments in GAI to determine if it will be a useful tool at some point in the future. DTAC should also stay up to date on innovations in virtual proctoring and continue to research other countries' positions to determine what input to give to policymakers who will make decisions regarding the use of virtual proctoring.

DTAC Response

1. DTAC believes the best resources for a "curated list of technical reports" for the DACMPT are the ASVAB, AFQT, and TAPAS validity frameworks. DTAC can work with AP (as allowed per FACA guidelines) to provide the most current documentation, and updates to the validity frameworks will be provided as they are completed (anticipated to be on a biennial basis).

2. Agree. DTAC will continue to keep the DACMPT apprised of research efforts. DTAC has recently begun a new effort to review AI, generative AI, and technology capabilities for testing and will plan to brief the DACMPT on the effort at a future meeting. DTAC continues to monitor trends in virtual proctoring and investigate new virtual proctoring technologies as they arise.

Major ASVAB R&D Efforts: Milestones and Project Schedules (cont.)

DAC Recommendations (06/24)

3. [06/24] The DACMPT remains impressed by both the number of projects OPA/DTAC manages and the quality of the research produced. The Committee voiced a potential concern about the high workload of this group. Dr. Pommerich pointed out that her team intends to create standard operating procedures so that they can move more quickly to deliver new test items in the future. The DACMPT applauds the careful development of items and encourages procedures that will facilitate that process.

DTAC Response

3. Thank you. The entire DTAC team (civilians and contractors alike, working on all aspects of our R&D efforts) is dedicated to maintaining the highest quality testing program. We are continually looking to investigate and refine our products and practices, standardize procedures, and introduce efficiencies, so we can alleviate workloads for our small but mighty team!

ASVAB/AFQT Validity Framework

DAC Recommendations (12/22)

1. The Committee agreed that Theory of Action [TOA] was applied very successfully in the AFQT selection context presented in developing, justifying, and empirically supporting the claims that were tested. Committee members appreciated how TOA-based validation efforts can usefully evolve over time. No validity evidence is static, and the TOA approach allows the body of validation work to be revised as the literature changes, and in light of different stakeholder purposes. ASVAB for classification may be more useful when average scores are higher because scores are less correlated (Legree, et al., 1961). The DACMPT recommends continued use of the TOA as an organizing framework for validity.

DTAC Response

1. Agree. DTAC has continued to use the Theory of Action as an organizing framework for validity. DTAC is continually updating its AFQT, ASVAB, and TAPAS validity arguments based on their respective TOAs.

Device Expansion Plans

DAC Recommendations (12/22)

1. The DACMPT asked about research on the interaction of item features and device variability to determine if different performance was observed for different items and tests when delivered on different devices, taking into account interactions among familiarity with the device, the task to be performed, response action, and device. Another question was raised about mode comparability research and the studies that were done or planned to ensure comparability of results across devices, operating systems, and browsers.

DTAC Response

1. Agree. DTAC did take into account the interaction of item features and device variability and determined that these were not drivers of performance and response time differences. Familiarity of device was the only significant factor that sometimes (depending on device and subtest) resulted in significant response time and performance differences. Likewise, the past device evaluation efforts did address various device, operating system, and browser conditions. Again, familiarity was the only factor with any significant interactions.

Device Expansion Plans (cont.)

DAC Recommendations (12/22)

2. The DACMPT made several recommendations regarding future research into alternate devices and their effects on test scores. Continuing research in this area should focus on differential analyses, as well as interaction effects that may impact dropping items from tests and/or evolving technologies (hardware and software). Data on the nature of the task, information on how the content is displayed, and the test taker's knowledge of moving around the screen should be collected and incorporated into the research.

DTAC Response

2. Agree. DTAC has developed a device expansion maintenance plan. This plan includes the collection of data from examinees regarding their test-taking experience, including how familiar they are with the device used. Examinees are encouraged to use a device they are familiar with before beginning the APT or PiCAT. DTAC plans to continue to research the impact of device expansion on performance differences, especially for new subtests added to the ASVAB battery.

ASVAB Adverse Impact

DAC Recommendations (12/22)

1. The DACMPT recommends regular analyses of adverse impact and exploration of potential reasons for differences in test performance to aid in promoting diverse accessions into the Military Services. Future assessments of adverse impact should also consider whether English is the examinee's first language.

DTAC Response

1. Agree. DTAC is developing a standardized analytic tool to evaluate adverse impact on an annual basis. DTAC does not currently have access to a standardized demographic question on language proficiency or English as the applicant's first language but can explore potential proxy variables.

AFQT Differential Prediction Study

DAC Recommendations (12/22)

1. Dr. Putka requested input from the DACMPT in three areas: the modified Cleary approach to assess differential prediction, other factors that may explain overprediction and underprediction, and approaches for dealing with limited power for analyses involving occupations with small sample sizes. Committee members noted that overprediction was expected and asked questions regarding combinations of outcome measures, the effect of the scores of individuals who did not make it into the study, the use of multilevel modeling for these multi-group analyses, other ways to probe differential prediction, (e.g., using the Johnson-Neyman regions of significance approach; Preacher, Curran & Bauer, 2006), and the use of multilevel modeling to address selection artifacts and comparisons involving technical and non-technical occupations.

DTAC Response

1. DTAC appreciates the input received from the DACMPT.

AFQT Differential Prediction Study (cont.)

DAC Recommendations (12/22)

2. After discussion of the approach taken and the available data for such analyses, the DACMPT made several suggestions regarding modifications to this research that might be considered: using performance measures that are broader and more direct than job knowledge tests, clustering related jobs or sorting jobs into technical and non-technical positions, using multilevel modeling as an analytic approach be considered going forward, and evaluating the effect of the test taker's native language. Despite these suggestions, the DACMPT is aware that the data needed for these initiatives may not exist at all, may not be reliably collected, or may not be available for a sufficient sample of test takers.

DTAC Response

2. Agree. The use of broader and more direct job performance measures rather than job knowledge tests is being looked into by the Services, particularly the Army in terms of military fitness and suitability. However, for criterion measures intended to be predicted by outcomes appropriate for ASVAB and other cognitive ability tests, it will require extensive planning and execution that would take a lengthy amount of time to run through the course of development. Clustering related jobs or sorting jobs into technical and non-technical positions is something that could and should be done. We are looking into this as a possible extension on previous studies. Using multilevel modeling as an analytical approach is something that could be explored and utilized in the next study, such as differential prediction. It would be interesting to know which multilevel techniques (e.g., HLM) the DACMPT has in mind, and DTAC would appreciate further elaboration. Evaluating the effect of a test-taker's native language would be an interesting application for DLI Foreign Language Center students or English Language Center students. As of yet, this has not gone past the conceptualization stage. Also, it could be a challenge gaining cooperation with DLI as these students are engaged in rigorous courses of study in language acquisition involving full-immersion learning.

DTAC appreciates the DACMPT's acknowledgment of limitations to their recommendations:
1) Data may not exist
2) Data may not be reliably collectible
3) Data may not be available for a sufficient sample of test takers

Non-Native English Speakers Analysis

DAC Recommendations (08/23)

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

AP Response

1. Concur. Military training and operations are conducted in English. DoD supports programs such as Foreign Language Recruiting Initiative (FLRI) for non-native English speakers (NNES) to improve their English skills. To ensure all requirements are considered and to provide for the maximum ability to affiliate with the military, work on NextGen ASVAB will take into account the needs of the NNES within the constraints of the training and operational requirements. Furthermore, when developing classification standards, Military Services take into account training and job requirements to include minimum level of English proficiency required for all servicemembers, to include both NNES and Native English Speakers. Finally, the Department has developed additional non-verbal assessment of cognitive ability, which should aid with identifying individuals who have the potential to benefit from immersive English proficiency training provided by the DoD. DTAC/AP will share this recommendation with the MAPWG Service representatives for consideration by their respective Military Services when designing enlistment programs and developing classification standards.

Complex Reasoning

DAC Recommendations (12/22)

1. The DACMPT valued the development of a complex reasoning measure because such a measure is lacking in the ASVAB, and virtually all jobs in the military require complex reasoning. Complex reasoning measures require very little verbal ability and therefore may be fairer to applicants, so long as they are familiar with this type of test. The DACMPT suggested that future research consider including non-English speakers in the pilot study to increase the potential to validate the test for those populations.

DTAC Response

1. DoD policy currently requires applicants to speak, read, and write English fluently. Military training and operations are conducted in English. Communication is a core requirement for training and job performance. Non-verbal assessment of cognitive ability should aid with identifying individuals who have the potential to benefit from immersive English proficiency training provided by the DoD. Recruiting non-English speakers for pilot studies poses some exceptional challenges as general information about the studies and instructions are presented in English. Nevertheless, DTAC has included demographic questions about English proficiency in subsequent pilot studies in an attempt to address this recommendation. Very few (less than 1%) of participants report that they do not speak English well or not at all, which limits analysis. DTAC will continue to work to increase representation of non-English speakers in research and development efforts but must acknowledge logistical obstacles.

Complex Reasoning (cont.)

DAC Recommendations (08/23)

2. Measure development: Determine why CR [Complex Reasoning] scores were "spiked" at a score of 11 across the three forms (this is unlikely to be coincidence). Continue expanding the item bank: Given that only 24 items were developed here, the item content might be leaked to examinees who then cheat. Fortunately, this can be remedied, because the quick generation of thousands of items is a virtue of the item format.
3. Nomological net: Correlate CR with ASVAB subtests to understand the nature of CR, where shared and unique sources of variance occur between the measures.
4. Validation: Support the CR measure further with validity evidence drawn from sources such as past military studies involving similar CR measures, or research literature when the results are generalizable to the military setting, as well as from new studies with the current CR measure.

DTAC Response

2. Agree. Histograms presented at the August 2023 DACMPT were based on incomplete results. This "spike" at raw score of 11 appears to have smoothed out somewhat in the final sample that is twice as large as what was included in the DACMPT presentation. Follow-on work includes additional item development efforts to expand the item bank.
3. Agree. These analyses will be presented at the January 2025 DACMPT.
4. Agree. DTAC has task orders in place for continued development and validation of CR and Computational Thinking composites to include plans for construct validation and criterion-related validation work.

Complex Reasoning (cont.)

DAC Recommendations (08/23)

5. Locate existing military data with CR-related data, in addition to conducting new validation work on the current CR measure (both selection- and classification-oriented validation). Although some military tests involving CR have not demonstrated incremental validity (see Besetsny et al., 1993), there is clearly more work to be done under a broader research framework. To this end, job analyses, O*NET data, and other resources may speak clearly to the need for an agenda for CR research across a wide range of MOS's.
6. Profile-driven analyses: Future research might consider how CR might work in tandem with a recruit or enlistee's profile of ASVAB scores. For example, specific ability tests are known to be more correlated (less differentiated) for those with lower general cognitive ability (see Detterman & Daniel, 1989), and those with higher cognitive ability may be more trainable for MOSs that do not fit their ASVAB subtest profile. These points have implications for classification that considers each enlistees' current interests and future goals alongside broader recruiting and labor demands.

DTAC Response

5. Agree. Criterion related validity evidence is typically the purview of the Services. DTAC will provide support with proposed research designs to facilitate cross-Service comparisons.
6. Agree. Classification composites are the purview of the Services. DTAC will assist as needed with composite or profile development efforts.

Complex Reasoning (cont.)

DAC Recommendations (06/24)

7. Members of the DACMPT commented on several aspects of the results of this work, including the difficulty of single-layer CR items, double-layer CR items, and items that are based on the diagonal of the matrix instead of the horizontal or vertical. The DACMPT agrees with the research team that appropriate methods of evaluating difficulty should be evaluated. The DACMPT also voiced concern about the need for practice items for test takers who are not experienced with this item type. Aware of the time limitations for any individual test, the DACMPT recommends careful consideration of the impact of practice on the difficulty of the items.

DTAC Response

7. Agree. DTAC is evaluating the impact of practice in the context of item presentation order and potential impacts on a Computerized Adaptive Test (CAT) version of CR. CR items are traditionally presented in order of increasing difficulty, which provides additional opportunity for experience and learning with these novel stimuli. This may necessitate a constrained CAT algorithm to accommodate for such impacts.

Computational Thinking

DAC Recommendations (12/22)

1. The DACMPT supports the development of the Computational Thinking [CT] measure via a composite and the plans for doing so. Many jobs in the military have increased requirements to develop, engage in, and solve technological problems. Consequently, the development and implementation of a computational thinking measure will likely improve military classification. More specifically, the Committee suggested increasing the representation of non-English speakers in the pilot study sample and reviewing the work of Zach Hambrick, who has developed a similar measure.

DTAC Response

1. DoD policy currently requires applicants to speak, read, and write English fluently. Military training and operations are conducted in English. Communication is a core requirement for training and job performance. Non-verbal assessment of cognitive ability should aid with identifying individuals who have the potential to benefit from immersive English proficiency training provided by the DoD. Recruiting non-English speakers for pilot studies poses some exceptional challenges as general information about the studies and instructions are presented in English. Nevertheless, DTAC has included demographic questions about English proficiency in subsequent pilot studies in an attempt to address this recommendation. Very few (less than 1%) of participants report that they do not speak English well or not at all, which limits analysis. DTAC will continue to work to increase representation of non-English speakers in research and development efforts but must acknowledge logistical obstacles.

Computational Thinking (cont.)

DAC Recommendations (08/23)

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

DTAC Response

2. Agree. Construct validation analysis results will be presented at the January 2025 DACMPT meeting. These will not include MOS-specific results. Nevertheless, DTAC will incorporate similar strategies in research design templates developed to assist the Services in further validation work.
3. Agree. Fairness evaluation is part of planned analyses.
4. Agree. EDPT is not part of future DTAC research plans.

Computational Thinking (cont.)

DAC Recommendations (06/24)	DTAC Response
5. The Committee appreciated the time-urgent need for developing the CT test and recommended that additional work should investigate subgroup differences and other fairness issues and conduct further validation research.	5. Agree. Updates on subgroup differences and construct validation plans will be presented at the January 2025 DACMPT meeting.

ASVAB Item Development Process—Item Analysis

DAC Recommendations (08/23)

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

DTAC Response

1. Agree. DTAC uses item-total correlations to evaluate item characteristics and quality. Tryout items administered under the planned missingness design do not contribute to operational scores.

CAT-ASVAB Pool and P&P—ASVAB Form Development

DAC Recommendations (12/22)

1. The DACMPT inquired about the transformation steps taken in terms of equating to understand better the processes used and to ensure that variability was not being introduced as a consequence of methodology. More information regarding these steps and the results is requested. Additional information on the efforts to detect and manage multidimensionality in data from CAT-ASVAB forms is also requested. The DACMPT also requests more information about the nature of the PC Test stimuli (length, content focus on informational vs. literary reading), given the research to meet operational constraints and ensure comparability between P&P and CAT.

DTAC Response

1. Agree. A comprehensive briefing of CAT-ASVAB equating methodology and rationale was presented to the DACMPT on August 16, 2023 (Reeder; 2023a). The equipercentile objective of producing equivalent composite distributions across alternate forms was discussed. The August 2023 briefing included a comparison between relying solely on IRT invariance property vs. application of the standard score post-equating methodology to illustrate impact of the equipercentile objective on qualification rates. A briefing specifically targeted toward addressing DACMPT concerns over potential of the equating procedure to produce biased or more variable scores at the individual level was presented on June 12, 2024 (Dahlke, 2024). Simulation analyses suggest the equating procedure is responsible for a very small proportion of observed-score variance and does not systematically bias estimated scores. Analysis results presented in both the 2023 and 2024 briefings indicate that the equating process serves its intended purpose without detrimental impacts on examinees' scores. The DACMPT was briefed on analytic methods for evaluating and managing multidimensionality in CAT-ASVAB tests on August 16, 2023 (Reeder, 2023b). Further investigation into dimensionality of the Assembling Objects test will be briefed at a future DACMPT. A briefing on the comparability of P&P-ASVAB to CAT-ASVAB is planned for the January 2025 DACMPT.

Form Equating Methodology

DAC Recommendations (08/23)

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

DTAC Response

1. Agree. A briefing specifically targeted toward addressing DACMPT concerns over potential of the equating procedure to produce biased or more variable scores at the individual level was presented on June 12, 2024 (Dahlke, 2024). Simulation analyses suggest the equating procedure is responsible for a very small proportion of observed-score variance and does not systematically bias estimated scores. Analysis results presented in both the 2023 and 2024 briefings indicate that the equating process serves its intended purpose without detrimental impacts on examinees' scores. DTAC does not understand the questions regarding consistency of scoring methods and believe those questions to be a misunderstanding of the materials presented. DTAC uses Bayes modal estimation consistently in scoring.

Form Equating Simulation Study

DAC Recommendations (06/24)

1. The DACMPT praised the thoroughness of the simulation study, viewing it as a valuable confirmation that the two-stage equating process works effectively at both the group and individual levels. The Committee recommended examining whether the results without the second stage produced similar outcomes. If the procedures with and without the second stage yielded comparable results, the possibility of simplifying the entire equating process in the future, if desired, could be contemplated.

DTAC Response

1. DTAC has previously presented results indicating that relying solely on the IRT invariance property (i.e., without the second stage) does not produce similar outcomes with respect to the equipercentile objective of qualification rates (Reeder; 2023a). Follow-up analyses will be presented at the January 2025 DACMPT to illustrate these impacts within the same simulation framework as the June 2024 presentation.

Form Development Methodology: Calibration Sample Size

DAC Recommendations (06/24)

1. The DACMPT acknowledged the outstanding work and recognized the importance of examining alternative calibration methods with smaller sample sizes. The differences in calibration results between flexMIRT and BILOG-MG were generally small, suggesting that the calibration program could be suitably replaced. The Committee raised a question about whether these differences could be further minimized by aligning the calibration settings of the two programs as closely as possible. In addition, the Committee recommended that DTAC consider the implications of switching the calibration program, including the need for recalibration of the current pools with the new program.
2. Regarding the use of a smaller sample size, the study showed that the psychometric properties, particularly reliability, did not change substantially across different sample sizes ranging from 700 to 1,200, supporting the use of a smaller sample size in the future. The practical benefit is clear, in the sense that a calibration sample size of about 970 would reduce the current data collection period by 8.3%. However, the Committee believes it is prudent to examine the impact of a smaller sample size on other aspects of the test, such as examinees' scores, DIF analysis, and more.

DTAC Response

1. Agree. Although there is not an immediate need to replace the current operational calibration procedure, DTAC is poised to replace BILOG-MG if and when circumstances dictate it is necessary. DTAC does not believe recalibration of the current pools is necessary given current robust scaling and equating procedures.
2. Agree. DTAC is currently engaged in research to evaluate impacts of smaller calibration sample sizes for DIF and other item-level analyses that are part of the pool development process.

Form Development Methodology: Use of Machine Learning and Natural Language Processing

DAC Recommendations (06/24)

1. Following the overview, the DACMPT praised the proposed system's use of modern technology and its potential to streamline ASVAB form development. There were no specific recommendations from the Committee on this topic. However, Committee members inquired whether generative artificial intelligence had been considered or used in this process. Dr. Pommerich responded that it is being considered as a multi-year project.

DTAC Response

1. Agree. DTAC is currently evaluating the security-related implications of incorporating generative models into this process but believes they can add value if content and process security can be assured.

Norming Requirements/Plans

DAC Recommendations (12/22)

1. One Committee member asked for a plot of trend results for AFQT scores.
2. The Committee discussed the possible effects of COVID on test scores, noting that some groups were more affected than others. The DACMPT recommends that efforts to re-norm should be deferred until the effects of COVID on propensity to serve have abated.
3. The DACMPT recommended that DTAC be sensitive to changes resulting from more vulnerable groups being differentially affected and wait until more time has elapsed before initiating a major re-norming effort. In addition, the methodology used for re-norming the ACT and SAT should be considered as plans to re-norm the ASVAB are developed.
4. The Committee also explored the development of norms based on the applicant pool instead of the customary approach of using the entire population. The DACMPT recommends that the DTAC consider the relative advantages and disadvantages of each approach before deciding which approach to use.

DTAC Response

1. Agree. Select AFQT trends were presented during the June 2024 DACMPT (McCloy, 2024). DTAC has developed a template analysis to monitor AFQT and other ASVAB score trends over time.
2. Agree. The technical working group (TWG) noted post-pandemic drops in student scores on NAEP, MAP, and other standardized tests. They noted the effects of school closures and remote learning could take a decade or more to rectify as most K–12 students were affected.
3. Agree. DTAC presented a summary of re-norming options and contingencies during the June DACMPT (McCloy, 2024) that include considerations for (a) the disruption to schooling that took place during the COVID pandemic, (b) differential impact of disruption to schooling, and (c) multiple methodological approaches to potential re-norming. DTAC agrees that waiting for the full impact of schooling disruptions is understood.
4. Agree. The TWG considered five options for renorming the ASVAB, including applicant-based norms. DTAC will consider the arguments for and against each approach as summarized in the June 2024 DACMPT (McCloy, 2024) briefing.

Norming Efforts

DAC Recommendations (06/24)

1. The DACMPT agrees with the presented results and does not believe that the age of the scale alone is a reason to renorm, noting that there may be public resistance to changing the long-standing interpretations of the scale. The DACMPT felt that the TWG had carefully considered a number of different advantages and disadvantages and had no suggestions for further work to inform the decision regarding renorming. The costs and common interpretations of scores further limit interest in renorming.
2. The DACMPT agrees with the TWG that renorming is not needed at this time; however, the Committee recommends continued monitoring of ability and demographic changes in the population.

DTAC Response

1. Agree. DTAC is aligned with the DACMPT and TWG in believing there are few if any substantive reasons to renorm at this time.
2. Agree. DTAC is working with a data monitoring/visualization tool to assist in evaluating NAEP, SAT, ACT and Census data trends in relation to ASVAB/AFQT scores and demographics.

Use of Calculators on the ASVAB

DAC Recommendations (12/23)

1. Continue with the planned research approach presented by DTAC. Research and subsequent transition plan should incorporate:
    - Clear articulation of the problem
    - Planned needs analysis
    - Impact on psychometric properties
    - Thoroughly designed transition including potential need for training of test administrators and applicants on calculator use and standardized roll out across the Military Services
    - Continuous program monitoring
    - Carefully defining and collecting appropriate outcome data

DTAC Response

1. Agree. Substantive updates on the research plan, including empirical impact analyses and needs analysis, will be presented at the January 2025 DACMPT meeting.

Use of Calculators on the ASVAB (cont.)

DAC Recommendations (06/24)

2. Committee members and other participants asked a number of questions, including concerns about adverse impact and individual differences when a calculator was used, the responsibility for bringing calculators to the test administration session, the need for training on the use of a calculator, the process of equating all applicable forms, and the potential need to examine calculator use and score differences by MEPS location. Committee members also raised questions about the relationship between the nature of Arithmetic Reasoning (AR) items and the effects of calculator use, the introduction of test anxiety when calculators are allowed, alternative analytic approaches (e.g., correlational studies), and the impact of calculators when the ASVAB is administered on tablets.
3. Overall, the research presented was well done and informative. The DACMPT looks forward to seeing the full result from Study 2 and Study 3. Given the study results and logistical concerns, the DACMPT does not find value in allowing the use of calculators on the ASVAB and does not anticipate that this effort would increase the number of qualified applicants.

DTAC Response

2. Agree. DTAC shares these concerns and will present further detail at the January 2025 DACMPT meeting.
3. Agree. More comprehensive findings from the empirical impact study and needs analysis will be presented at the January 2025 DACMPT meeting to address many of the DACMPT's concerns, which are shared by DTAC. Given the ambiguity of the problem definition, arbitrary timeline, administrative barriers, and potential scope of the impact, DTAC's capacity to address emerging issues revealed by these studies may be limited.

Next Generation ASVAB/Testing—Evaluation Plan

DAC Recommendations (12/22)

1. The DACMPT asked how DTAC defined improvements in selection (e.g., increases in validity or satisfaction). The answer will require another look at the philosophy or purpose of the ASVAB. The DACMPT recommends careful consideration of the criteria for "improvement."
2. Committee members recognized the diversity of needs among stakeholders. For example, military trainers are generally pleased with the current tests because new recruits succeed during training. At the same time, recruiters want a test that will qualify more people and allow them to meet their recruiting missions. Although a completely shared vision for the ASVAB is likely impossible, there are no major complaints, and DTAC is hoping to meet most of the stakeholders' goals. The DACMPT encourages continued efforts to evaluate stakeholder perceptions and to educate them on the compromises that must be made.

DTAC Response

1. DTAC agrees that careful consideration of the criteria for improvement in selection is needed. DTAC has actively been considering the criteria and process for making changes to the ASVAB since 2011. A detailed plan for NextGen ASVAB was presented to the DACMPT in 2020. Regarding the philosophy of the ASVAB question, DTAC completed a thorough review in 2023 of all the ASVAB philosophy discussions that took place over the past several decades and concluded that the DACMPT's 2011 recommendation to articulate the ASVAB philosophy might have unintentionally led to an impasse between DTAC and the Services regarding ASVAB content decisions due to competing philosophies. As such, current thinking is to remove references to a specific philosophy and reframe ASVAB content discussions to focus on guidelines and evaluation processes that have been mapped out. DTAC continues to solicit input from stakeholders to ensure that the different purposes for which they use the ASVAB continue to be met.
2. Agree. DTAC continues to communicate with stakeholders to learn their differing needs and perspectives, build a shared understanding, and help identify a way forward for Next Generation Testing. Most recently, DTAC held a 3-day workshop with a variety of ASVAB stakeholders in November 2024, as well as conducted interviews with additional stakeholders not participating in the workshop.

Next Generation ASVAB/Testing—Evaluation Plan (cont.)

DAC Recommendations (12/22)

3. The primary concern about testing time does not come from applicants but comes from MEPCOM, which prefers to complete all testing in a single day to avoid overnight stays. Committee members discussed the issue of the length of the tests and briefly explored alternatives such as changing the CAT stop rules, moving item seeding requirements from proctored testing to VTest administrations, employing psychometric refinements, using a multidimensional approach (e.g., multidimensional IRT), and initiating a taxonomy content review to identify redundancies. Although the tests are already short, the DACMPT recommends that DTAC continue to explore various ways to shorten the length of time required for administering the ASVAB and special tests.
4. The Committee also discussed applicant perceptions of the ASVAB. The available data were collected from individuals who had taken the ASVAB but had not yet completed training and did not include high school students taking the CEP or applicants who were not accepted. The DACMPT encourages efforts to understand a broader range of applicant reactions to the ASVAB.

DTAC Response

3. Agree. DTAC continues to consider avenues to reducing testing time to alleviate the burden on MEPCOM resources. Testing time was a focus of one of the exercises conducted at the November 2024 ASVAB stakeholder workshop.
4. DTAC agrees that it would be useful to get the perspectives of CEP participants and applicants that do not qualify for entry into the military, but also notes that these are difficult populations to get access to. In focus groups that were conducted with qualifying applicants, a number discussed taking the ASVAB via the CEP. If there are future focus group efforts, we will make every effort to speak with as broad of a swath of the test-taking population as is practically feasible.

Next Generation Testing

DAC Recommendations (06/24)

1. The DACMPT supports the systematic approach to considering future changes to the ASVAB and has no substantive comments to make, other than that the focus groups of panelists should consider the needs of the Services in the future, as they make their judgments on which tests should be included.

DTAC Response

1. Duly noted. DTAC has recently conducted a Next Generation ASVAB workshop with various stakeholder groups, including technical representatives, policy representatives, recruiters, classifiers, and trainers from the Services. DTAC plans to keep the Services involved as Next Generation ASVAB efforts unfold.

Next Generation Testing Stakeholder Focus Group Study

DAC Recommendations (12/22)

1. The DACMPT asked about the representation of the study participants relative to the populations. Demographic information about the participants was limited. Consequently, the sample did not meet strict sampling conditions. The DACMPT recommends that future focus groups ensure adequate representation of all critical groups.

DTAC Response

1. Agree. If there are future focus group efforts, we will make every effort to speak as broad a representation of relevant subgroups as is practically feasible.

High School Curriculum Study

DAC Recommendations (08/23)

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

DTAC Response

1. Agree. For clarification, DTAC is using the common high school curricula study as a separate source of information to support the NextGen ASVAB work. That is, we are not looking for the high school curricula study to support the inclusion of Complex Reasoning and Computational Thinking. While the data collection plan has already been established and implemented, DTAC will take a multilevel approach, to the extent possible, with the existing data to explore the hypothesis that schools with more resources provide more courses. Likewise, DTAC will consider an extension of the work to assessing schools from a workforce development perspective but would like to hear more from the DACMPT on what they envision and how this work could improve the composition of the ASVAB for selection and classification purposes. Another follow-up effort DTAC will consider is collaborating with other DPAC teams to determine whether schools with better resources that offer curricula aligned with ASVAB subtests offer better recruiting environments and thereby also produce more eligible students with a propensity for military service.

ASVAB CEP

DAC Recommendations (12/22)

1. The DACMPT suggested that the “Bring ASVAB CEP to your school” program be looked at closely to determine if the scheduling forum has pushed people away since 2019 and if the demographic questions on the forum should be revised.
2. The DACMPT also suggested that students be assigned an identification code (e.g., pseudo name or number) to reduce the concerns about Military Service.
3. Other suggestions included using social media to facilitate a culture of interest in schools, emphasizing the focus on exploring jobs and work as opposed to college and stressing the “whole-person” nature of the assessment.
4. The Committee also felt that strong student testimonies placed on the homepage might engage more users and should be considered.
5. Other efforts to engage more users include working jointly with programs like Upward Bound and offering the program to undeclared freshmen in college and those in the TRIO program.
6. YouTube videos that are aligned with the topics in the “Student Articles” would also be helpful.
7. Understanding other programs in high school that compete with the ASVAB-CEP could help direct marketing efforts, and the use of social media tools such as Kahoot could enable better connections among educators, students, and the military.

DTAC Response

1. Agree—form was revised to allow the user to input only critical information to allow for ESS follow-up.
2. Agree—collaboration with USMEPCOM is required to modify the score sheet.
3. Agree—launched social listening activities and social campaigns tailored to educator sharing and promotion among the educator community.
4. Agree—this information is gathered when possible (challenge: multiple layers of approval required to contact students but ESSs can and do encourage student self-posting).
5. Agree—a new business strategy was activated in 2023 to engage underserved populations and broaden efforts in community colleges and other organizations with relevant populations.
6. Agree—relevant videos have been created and are being developed that align with this suggestion.
7. Agree—an in-depth Competitor Analysis is underway. A white paper was provided to Accession Policy that outlined specific comparisons between ASVAB CEP and SchoolLinks.

ASVAB CEP (cont.)

DAC Recommendations (12/22)

8. No major concerns were uncovered; however, the DACMPT would like to see more information regarding the Army's success in using CEP scores for enlistment. The DACMPT also requests that the following questions be addressed in future meetings:
    - Should ASVAB-CEP be mandatory for high school students, and what will be the ramifications for the military services?
    - What methods will best persuade students to take the ASVAB-CEP and take it seriously?
    - How can the military promote, "Do you know people like you who took the ASVAB-CEP"?
    - How should non-cognitive measures be incorporated into the selection and classification programs?
    - How does the content in high school curricula align with the ASVAB, and what are implication for changes to either or both?
9. Finally, the Committee endorses the idea of the Committee members working through the website to better understand the program.

DTAC Response

8. Agree—Where not yet briefed to the DACMPT, recommend adding to the agenda for future meetings.
9. Agree—A walkthrough was provided at the 08/23 DACMPT meetings, and login credentials have been provided to Committee members.

ASVAB CEP (cont.)

DAC Recommendations (08/23, 06/24)

1. [08/23] Following the overview, the DACMPT complimented the tool and made a recommendation to identify ways to evaluate user engagement that goes beyond merely counts of accessing the website, such as by measuring frequency of return users.
2. [08/23] The Committee also endorsed the idea of better explaining the program, so that more participants take advantage of the Post-Test Interpretation service.
3. [06/24] The DACMPT continues to believe that the ASVAB CEP is an important tool for identifying potential recruits for the Services and provides a public service to youth in America. The biggest shortfall in this program appears to be its limited use. Consequently, the DACMPT encourages continued marketing efforts to inform the public in general and high school leadership specifically.

DTAC Response

1. Agree—return user has been added to the Key Performance Indicators on website analytics. The team is also exploring a pop-up survey to be administered to gather more specific feedback.
2. Agree—this correlates closely with ongoing efforts to standardize training and program delivery, disseminate marketing communications, and introducing the Ambassador Program.
3. Agree—expanded and refined marketing efforts to reach targeted audiences including school board members, community colleges, superintendents, and state- and district-level decision makers.

TAPAS Validity Framework and Joint Enlistment Composite

DAC Recommendations (08/23)

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

DTAC Response

1. Agree. DTAC has a plan in place to explore suitable criteria, which begins with reviewing past work by contractors to establish a common set of criterion measures. The goal is to map any existing measures within the existing framework to military compatibility efforts. Likewise, we are asking the Services to also offer their criterion measures and experiences. Finally, DTAC will propose additional avenues for validating the TAPAS military compatibility composite, including possible synthetic and concurrent validity approaches, as suggested. Included in the validity research will be a review of the facet weighting schemes applied. As TAPAS development evolves, facets will be refined, and new facets will be developed to better support the assessment of military compatibility. Refinement efforts are planned for FY25, and new development will begin in FY27. The Services have the flexibility to introduce new Service-specific facets within the Joint-Service TAPAS.

TAPAS for Military Compatibility

DAC Recommendations (08/23)

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

DTAC Response

1. Agree. DTAC has ongoing plans to explore suitable criteria, which begins with reviewing past work by contractors to establish a common set of criterion measures. The goal is to map any existing measures within the existing framework to military compatibility efforts. Likewise, we are asking the Services to also offer their criterion measures and experiences. There are 10 categories of misconduct that the military compatibility composite will address. It is these 10 categories for which we will focus on finding suitable criterion measures. Unfortunately, research shows that the military core values across Services are not correlated with (or a reverse measure of) the 10 categories of misconduct. DTAC plans to explore multi-level measurement models to address possible issues with stability.

TAPAS for Military Compatibility (cont.)

DAC Recommendations (08/23)

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

DTAC Response

2. Agree. The development of the military compatibility composite based on TAPAS facets is a phased approach. Phase 0 makes it possible to collect data across all Services on the Army Conduct Composite, which is our first military compatibility composite. With this data, we can begin to explore issues related to validity, subgroup differences, and stability. DTAC has developed a targeted definition of military compatibility that focuses on 10 categories of misconduct. DTAC's goal is not to assess adherence to military core values, as we found that these are not correlated with the 10 categories of misconduct defined by the Military Compatibility Research Group (MCRG). The Joint-Service TAPAS Military Compatibility Composite is not planned to be the sole source of evidence for disqualifying candidates from the Services. Instead, it will serve as a flagging tool that would invoke further investigation via a clinical psychological interview by a licensed clinician who would provide a Service eligibility recommendation. This two-stage approach is currently being refined and will undergo various levels of validity studies before implemented operationally. Phase 1 JS-TAPAS development will focus on refining the facet pools and the military compatibility composite. Phase 2 will focus on introducing new facets into the JS-TAPAS that support the increased validity for the military compatibility composite. Research in the area of military compatibility assessment is also ongoing for the Officer population where 13 existing assessments are being evaluated for their appropriateness. Findings from this research will inform the enlistment testing program. Accession Policy and DTAC are open to instruments other than TAPAS. DTAC also plans to develop and pilot its own Situational Judgment Test, intended to address the assessment of military compatibility defined by the 10 categories of misconduct.

TAPAS for Military Compatibility (cont.)

DAC Recommendations (08/23)

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

DTAC Response

3. Agree. The current plan is to structure the military compatibility assessment into two parts: 1) Use TAPAS Military Compatibility Composite (or equivalent composite for officers) to flag individuals at risk for deviant behaviors; and 2) Use the clinical assessment to obtain professional judgment on those flagged by the test in part 1.

Non-Cognitive Updates

DAC Recommendations (06/24)

1. Members of the DACMPT applauded the careful approach to developing these measures and encouraged future research to pay careful attention to the criteria used for deviant behaviors, particularly those that occur less frequently. The literature on honesty and integrity may be a useful source of information.

DTAC Response

1. Agree. DTAC has a plan in place to explore suitable criteria that begin with reviewing past work by contractors to establish a common set of criterion measures. The goal is to map any existing measures within the existing framework to military compatibility efforts. Likewise, we are asking the Services to also offer their criterion measures and experiences. Finally, DTAC will propose additional avenues for validating the TAPAS military compatibility composite, including possible synthetic and concurrent validity approaches, as suggested at the Aug 2023 meeting of the DACMPT. Literature on honesty and integrity is a key resource that DTAC has been reviewing within the Best Practices Project, as that team contains an expert researcher in the area.

Best Practices Project

DAC Recommendations (06/24)

1. Members of the DACMPT voiced similar concerns regarding the weaker prediction of extreme forms of counterproductive work behaviors and advised attention to the criterion used, given the implications of using such an instrument to reject potential officers. A second recommendation is to examine sex differences and race/ethnicity differences in future work, as well as the effects of providing warnings to keep respondents from minimizing or ignoring past misconduct.

DTAC Response

1. Agree. DTAC has a plan in place to explore suitable criteria, which begins with reviewing past work by contractors to establish a common set of criterion measures. The goal is to map any existing measures within the existing framework to military compatibility efforts. Likewise, we are asking the Services to also offer their criterion measures and experiences. Finally, DTAC will propose additional avenues for validating the military compatibility facets/scales administered for the officer population. Presently, DTAC is exploring various scales within 13 existing assessments for their utility with an officer population. Likewise, DTAC will begin development of a Situational Judgment Test aimed at addressing the 10 identified focus areas for military compatibility assessment. Validation research will include an exploration of sex differences and race/ethnicity differences.

Legislation/Policy Review

DAC Recommendations (06/24)

1. The DACMPT has no comments on the law but noted that the legal requirements for minimum scores emphasize the importance of accurate equating of forms.

AP Response

1. Concur. This is the normal practice of the Department and will continue to be followed.

Resource Overview

DAC Recommendations (06/24)

1. The Committee asked about the maintenance of current funding levels, cautioning that the description of levels as “healthy” may result in future reductions or future “leveraging” of funding for other purposes. Dr. Pommerich clarified that current levels of funding would be sufficient if cuts are not imposed, but that cloud costs could become an issue. In addition, Dr. Pommerich also said that if the ASVAB were to be re-normed, additional funding would be needed. Because current funding levels are adequate, no additional funding is needed at this time. This assumes cloud costs remain constant, and DTAC is not tasked with additional norming work or other unforeseen efforts. If major projects like ASVAB re-norming are directed, the DACMPT strongly recommends additional resources be provided to address these issues.

DTAC Response

1. DTAC continues to monitor funding levels and cloud costs, to ensure that an optimal level of funding is maintained or that steps could be taken to secure additional funding, if needed.

Future Topics

DAC Recommendations (12/22)

1. The DACMPT recommends future meetings incorporate briefings on the following topics:
    - Item development and equating methodology
    - Non-cognitive measures
    - The high school curriculum study
    - The TAPAS validity framework
    - Non-native English speakers and test performance
    - CEP website
    - Integrating measures into the master plan for selection and classification

DTAC Response

1. Duly noted. DTAC has and will continue to coordinate with AP to schedule briefings on suggested topics when applicable and feasible.
    - Item development processes briefed at the 8/23 DACMPT meeting
    - Equating methodology briefed at the 08/23, 06/24, and 01/25 DACMPT meetings
    - Non-cognitive measures briefed at the 08/23, 06/24, and 01/25 DACMPT meetings
    - High school curriculum study briefed at the 08/23 and 01/25 DACMPT meetings
    - TAPAS validity framework briefed at the 08/23 DACMPT meeting
    - The non-native English speakers study briefed at the 08/23 DACMPT meeting
    - CEP website demonstrated at the 08/23 DACMPT meeting
    - Next Generation Testing briefed at the 06/24 DACMPT meeting

Future Topics (cont.)

DAC Recommendations (08/23)

1. The DACMPT recommends future meetings incorporate briefings on the following topics:
    - Overview of the various tests that highlights similarities and differences among tests (e.g., Cyber test vs. EDTP)
    - Another review of the equating procedures
    - Overview of Next Generation ASVAB and how the pieces (e.g., Complex Reasoning) fit together
    - Overview of the process for planning that takes into account a rapidly changing testing landscape (especially important given the rapid influx of AI technologies that affect testing)
    - Norming procedures
    - Allowing the use of calculators
    - Reviewing the nature, pros, and cons of super-scoring

DTAC Response

1. Duly noted. DTAC has and will continue to coordinate with AP to schedule briefings on suggested topics when best applicable and feasible.
    - Next Generation Testing briefed at the 06/24 DACMPT meeting
    - Equating methodology briefed at the 06/24 and 01/25 DACMPT meetings
    - Next Generation Testing briefed at the 06/24 DACMPT meeting
    - DTAC is exploring AI/GAI/technology advancements and can report on the status of ASVAB and non-cognitive efforts in this realm at a future meeting
    - Norming efforts briefed at the 06/24 DACMPT meeting
    - Use of calculators on the ASVAB briefed at the 12/23 and 06/24 DACMPT meetings
    - DTAC is prepared to brief the DACMPT on super-scoring whenever it is scheduled

Future Topics (cont.)

DAC Recommendations (06/24)

1. The DACMPT believed that all the suggestions for future research were worthy of attention. The DACMPT recommended future meetings incorporate briefings on the following topics:
   - Adverse impact analyses
   - TAPAS
   - Calculator implementation efforts
   - Complex Reasoning test
   - Interest measures
   - Curriculum studies

DTAC Response

1. Duly noted. DTAC has and will continue to coordinate with AP to schedule briefings on suggested topics when best applicable and feasible.
   - Adverse impact will be briefed at the 01/25 DACMPT meeting
   - TAPAS will be briefed at the 01/25 DACMPT meeting
   - Calculator efforts will be briefed at the 01/25 DACMPT meeting
   - Complex Reasoning efforts will be briefed at the 01/25 DACMPT meeting
   - The Find Your Interests inventory will be briefed at the 01/25 DACMPT meeting
   - The high school curriculum study will be briefed at the 01/25 DACMPT meeting

Tab H

OPA

OFFICE OF PEOPLE ANALYTICS

Research Supporting Alterations to the Specifications for P&P-ASVAB

Jeff Dahlke

Human Resources Research Organization

Briefing presented to the DACMPT

January 22, 2025

Agenda

- Background information
- Overview of necessary adjustments to specifications for the paper-and-pencil ASVAB (P&P-ASVAB)
- P&P-ASVAB research summaries
  - IRT rescaling method for Auto and Shop Information (AS)
  - Length-reduction analyses for Paragraph Comprehension (PC)
  - Length-reduction analyses for Arithmetic Reasoning (AR)
  - Time limit adjustments
- Summary of recommended alterations to P&P-ASVAB specifications for new forms

Background Information

- The P&P-ASVAB is a linear fixed-form version of the ASVAB, administered using physical test booklets and answer sheets
  - Produces standard scores on the same dimensions as CAT-ASVAB
- P&P-ASVAB is administered in the Enlistment Testing Program (ETP) and the Career Exploration Program (CEP)
  - Represents a very small share of the testing volume for ETP but a large share of the testing volume for CEP
- HumRRO has developed new P&P-ASVAB forms for both ETP and CEP to replace the current sets of forms
- Due to P&P-ASVAB being administered in a group setting (as opposed to individually, like CAT-ASVAB), testing time is at a premium
  - Exceeding the current total testing time is not viable for ETP or CEP

Background Information (Continued)

- All items available for P&P-ASVAB forms were developed for and tried out in CAT-ASVAB
- Items for some subtests were not directly compatible with the P&P-ASVAB design:
  - Auto and Shop Information (AS)
    - Whereas AS scores are computed as a composite of Auto Information (AI) and Shop Information (SI) scores for CAT-ASVAB, AI and SI must be administered and scored together as a single AS subtest for P&P-ASVAB
    - Based on dimensionality research that informed the development of CAT-ASVAB, AI and SI items are calibrated, scaled, and administered separately for CAT-ASVAB
    - All CAT-ASVAB AI and SI item parameters are on their respective subtest scales, and the items are tried out (i.e., field tested) with non-overlapping groups of examinees
  - Paragraph Comprehension (PC)
    - Past P&P-ASVAB PC sections used a testlet design (multiple items about each passage)
    - All CAT-ASVAB PC items use a stand-alone passage for each item

High-Level Specifications for Past P&P-ASVAB Forms

Subtest	Item Count	Time Limit (Minutes)
General Science (GS)	25	11
Arithmetic Reasoning (AR)	30	36
Word Knowledge (WK)	35	11
Paragraph Comprehension (PC)[1]	15	13
Mathematics Knowledge (MK)	25	24
Electronics Information (EI)	20	9
Auto and Shop Information (AS)[2]	25	11
Mechanical Comprehension (MC)	25	19
Assembling Objects (AO)[3]	25	15

[1] Past PC item sets were constructed using a testlet design, where multiple items are administered for each passage.
[2] AS is scored as a single subtest for P&P-ASVAB; in CAT-ASVAB, it is scored as a composite of separate AI and SI subtest scores.
[3] AO is administered only in the Enlistment Testing Program, not the Career Exploration Program.

OPA
OFFICE OF PEOPLE ANALYTICS

Overview of Necessary Adjustments to P&P-ASVAB Specifications

1. Estimate transformations that link the AI and SI IRT scales to a single AS scale
   - Define a target AS scale that closely approximates the scores examinees would earn if it were possible to score AS as a composite of AI and SI scores
2. Update the number of items in PC item sets to account for the use of items with stand-alone passages instead of testlets
   - Decrease the item count to reduce the reading load and limit testing time demands while maintaining an acceptable level of score reliability
3. Update the number of items in AR item sets to mitigate speededness
   - Decrease the item count to limit testing time demands while maintaining an acceptable level of score reliability
4. Update time limits
   - Identity potential changes to subtest-level time limits to accommodate an increased time limit for PC

OPA
OFFICE OF PEOPLE ANALYTICS

Note: Due to ongoing research examining the Assembling Objects (AO) subtest, this briefing is focused on the other 8 subtests, all of which are shared between ETP and CEP.

IRT Rescaling Method for Auto and Shop Information (AS)

OPA
OFFICE OF PEOPLE ANALYTICS

Background and Motivation

- All IRT item parameters for available AS items are on separate Automotive Information (AI) and Shop Information (SI) scales
  - The separate scales exist to support the CAT-ASVAB, where AI and SI get scored separately and those scores are combined into an AS composite
  - P&P-ASVAB must administer and score AS as a single subtest
- The AI- and SI-scaled items must be translated to the P&P-ASVAB AS scale before they can be used
  - AI and SI items are tried out with non-overlapping samples, so the data used to calibrate them cannot support a combined AS-scaled calibration
  - We initially planned to collect new data to recalibrate a set of AI and SI items
  - We now plan to use a custom-built rescaling procedure to accomplish this

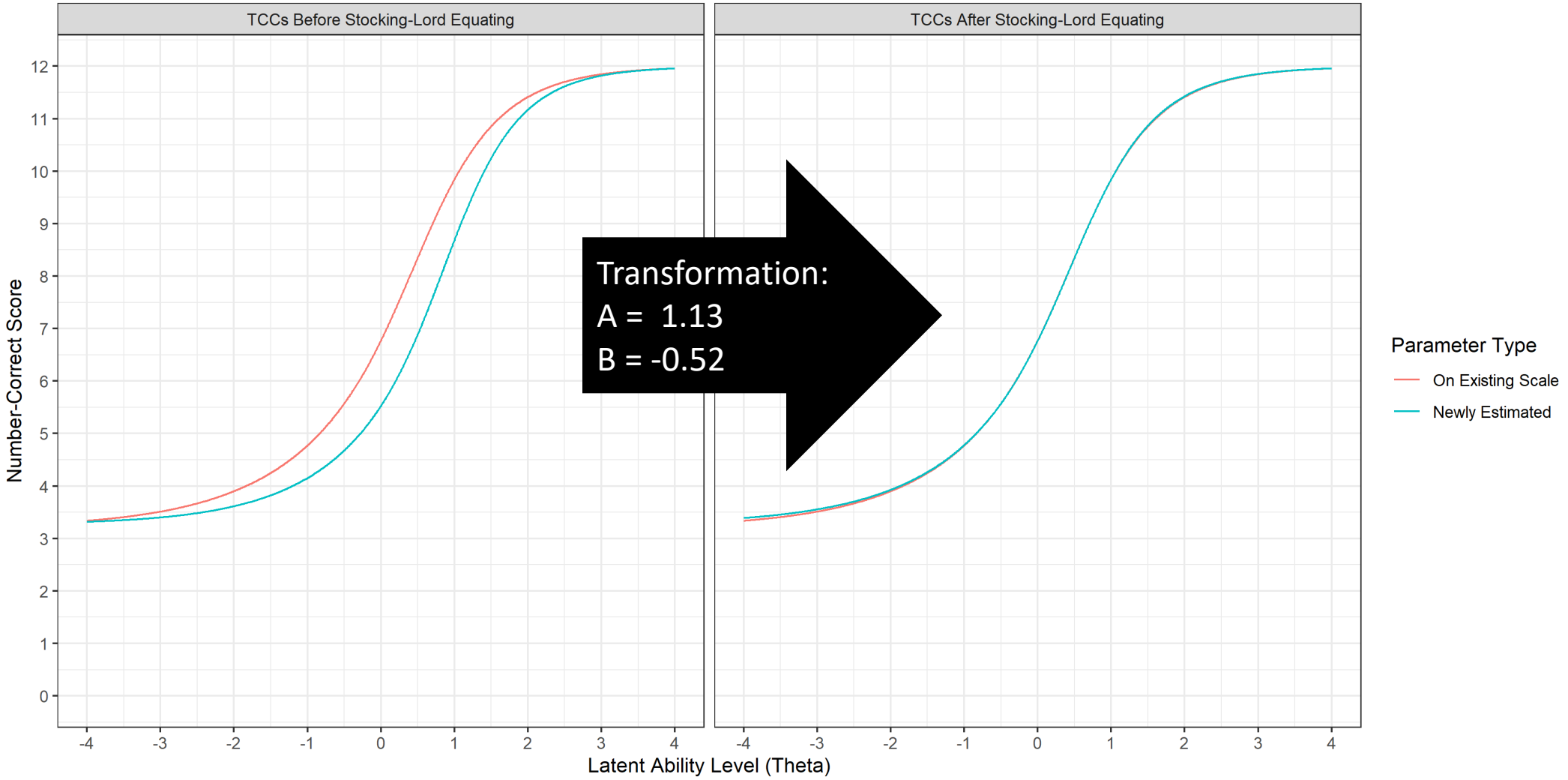
Background and Motivation (Continued)

- Initial plan to get AS-scaled item parameters:
  - Administer to examinees (a) CAT-scaled AI and SI items and (b) anchor items from past P&P-ASVAB AS item sets
  - Calibrate all items together, scaling them on a single dimension
  - Use anchor items' IRT parameters to link the newly estimated parameters to the historical AS scale
  - Rescale all items to the historical AS scale
- Drawbacks to the initial plan:
  - Psychometrically suboptimal
  - Time-consuming
  - Expensive
  - Risky (unclear how it would turn out, given that we would violate an IRT assumption)
- Critical Question: **How can we shift AI and SI item parameters onto the AS scale without collecting new data?**

Inspiration from the Stocking-Lord Equating Procedure

OPA
OFFICE OF PEOPLE ANALYTICS

Example: Stocking-Lord Test Characteristic Curves (TCCs)



OPA
OFFICE OF PEOPLE ANALYTICS

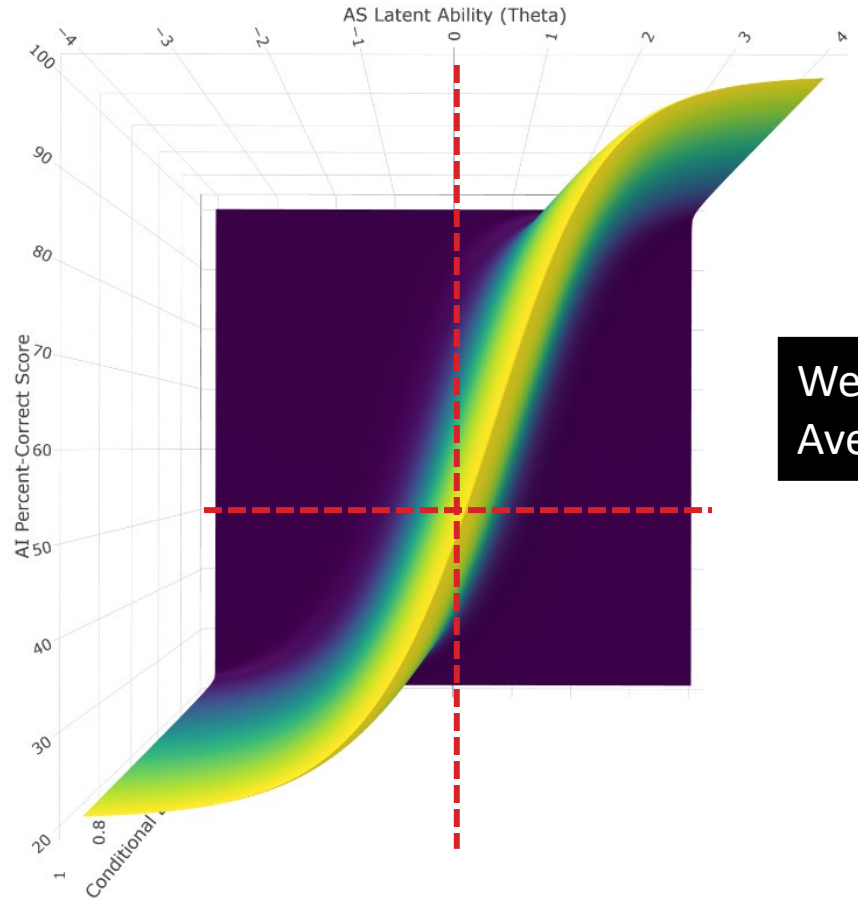
If anchor items can provide the scaling information needed to rescale item parameters, could we use alternative scale-anchoring information to achieve the same effect?

Solution: The Modified Stocking-Lord Procedure (MSLP)

- Instead of using anchor items' parameters to define the scale of a test, we can get relevant scale information from latent ability distributions of person parameters
  - AI, SI, and AS have latent means and SDs from past research on the scaling of P&P-ASVAB and CAT-ASVAB
  - We have an estimate of the correlation between latent AI and SI distributions derived from operational CAT-ASVAB data
  - Using the above, we can construct a complete variance-covariance matrix relating AI and SI to AS, where AS is a composite of AI and SI
  - The variance-covariance matrix and means allow the parameters on one scale to be reflected onto another scale while accounting for their shared variance
- Instead of anchor items, all we really need for rescaling is a relevant target TCC
  - S-L uses a target TCC based on item parameters that are already on the test's scale
  - MSLP constructs a target TCC by reflecting AI and SI TCCs onto a composite scale

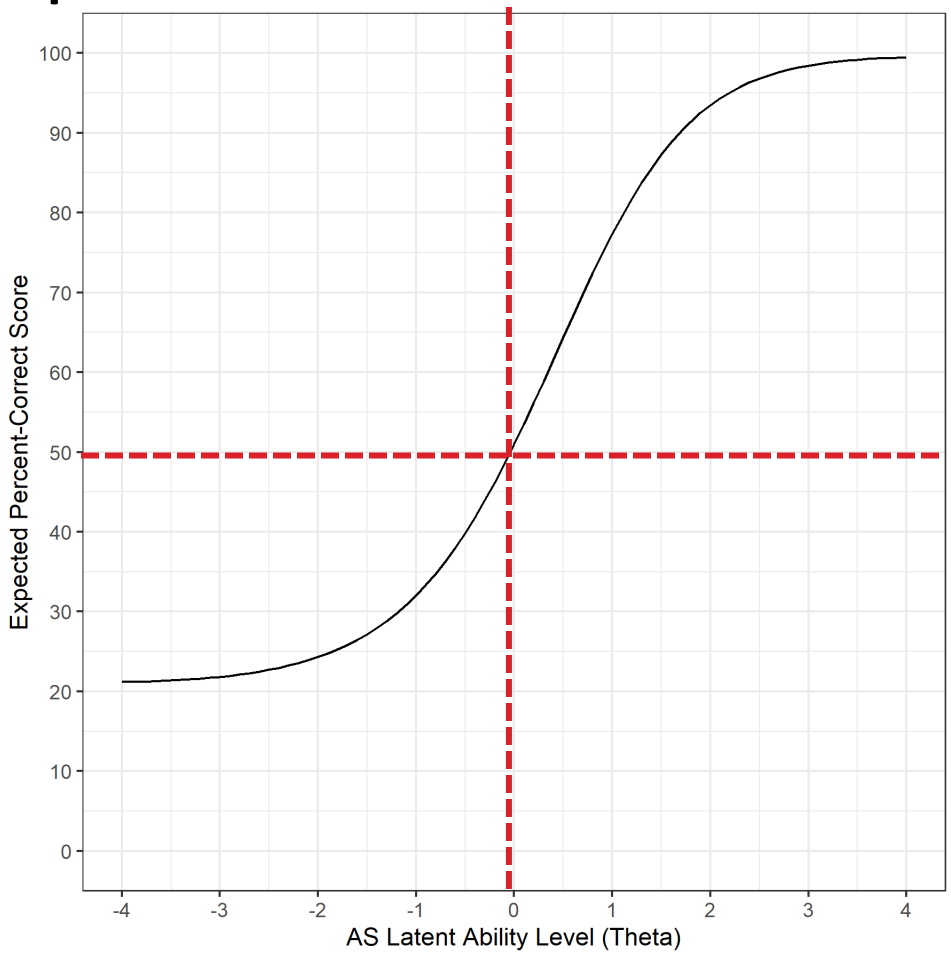
Deriving a Target TCC from Distributional Information About Abilities

AI Score Distribution, Conditioned on AS Ability



Weighted Averaging

Expected TCC for AI Items on the AS Scale



OPA
OFFICE OF PEOPLE ANALYTICS

Estimating Rescaling Coefficients

- After using a multivariate density distribution to estimate the expected TCC for a subtest, that TCC can be used as a target in a rescaling procedure
- From this point onward, the MSLP functions exactly like the traditional Stocking-Lord procedure in how it iteratively estimates coefficients:
  1. Identify a set of provisional linear rescaling coefficients
  2. Use the provisional coefficients to rescale the item parameters
  3. Use the provisionally rescaled item parameters to compute a TCC
  4. Subtract the provisionally rescaled TCC from the target TCC
  5. Compute the density weighted sum of absolute-value differences
  6. Repeat steps 1–5 until Nelder-Mead optimization reaches convergence
     - Relative tolerance criterion for TCC-matching objective function = 1e-8

OPA
OFFICE OF PEOPLE ANALYTICS

Simulation to Evaluate the MSLP

OPA
OFFICE OF PEOPLE ANALYTICS

MSLP Evaluation Simulation

- Purpose: Benchmark the MSLP's performance against relevant comparators
- We evaluated the accuracy of expected TCCs against empirical TCCs
- We compared TCCs from MSLP to other calibration methods:
  - Co-calibration of subtest items with BILOG-MG
    - After calibration, item parameters were rescaled to match the composite AS scale
  - Fixed-theta calibration with MULTILOG
    - This is conceptually the most similar to what the MSLP is meant to accomplish because it allows item parameters to be expressed on the composite AS theta metric without strict dimensionality assumptions

MSLP Evaluation Simulation: Design

- Simulated AI- and SI-like item parameters based on multivariate-normal distributions (*a* and *c* parameters were scaled as logits)
  - AI-like items designated “Test A” and SI-like items designated “Test B”
  - 200 items per test to reflect current item-seeding practices
- Simulated person parameters from bivariate-normal distributions
  - Ability distributions were based on latent means and SDs for AI and SI estimated from recent operational CAT-ASVAB data
  - Varied correlation between Tests A and B from 0.0 to 1.0 in 0.1 increments
  - 16k simulees per correlation condition
    - Resulted in an average of 1,200 responses per item with 15 random items administered to each simulee per test
  - Composite ability was an unweighted average of ability on Tests A and B

OPA
OFFICE OF PEOPLE ANALYTICS

MSLP Evaluation Simulation: Design (Continued)

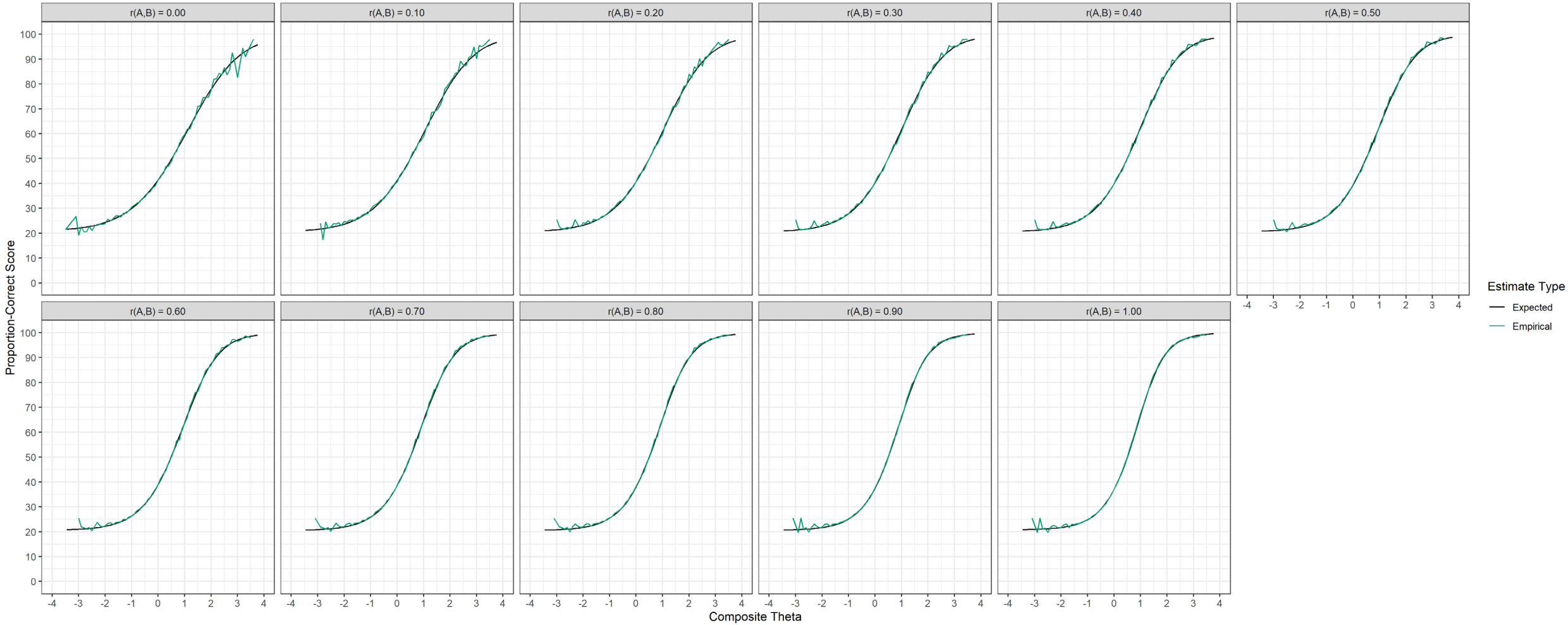
- Simulated item responses using person and item parameters
  - For each simulee-item combination, the simulee's true theta and the item's true IRT parameters were used to estimate the probability of a correct response
  - To introduce measurement error, simulee's probabilities of correct responses were compared to randomly generated values from a [0,1] uniform distribution
    - A simulee got an item correct if their probability of a correct response was greater than or equal to the random value
- Calibrated items from each test
  - BILOG-MG parameter estimates were rescaled using latent means and SDs
- 100 replications
  - The results were highly consistent across replications; we will focus on one of them

OPA
OFFICE OF PEOPLE ANALYTICS

Accuracy of Expected TCCs

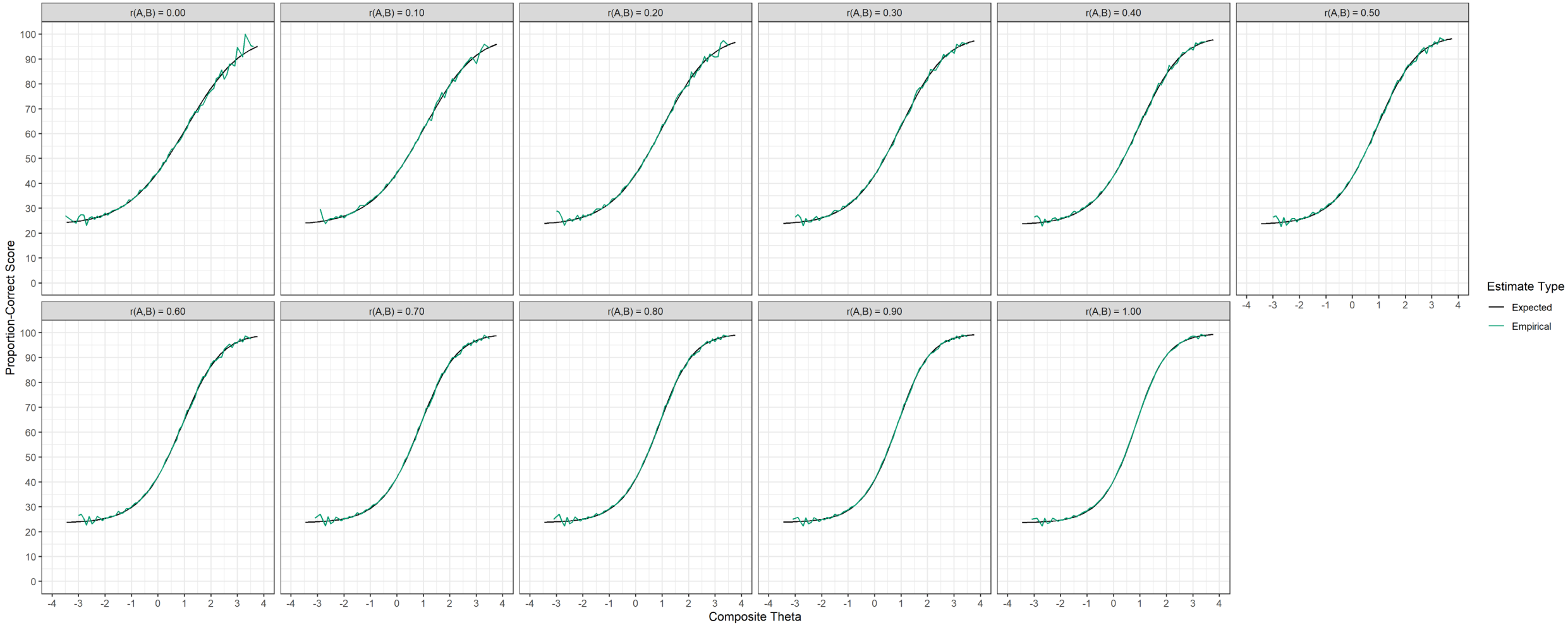
OPA
OFFICE OF PEOPLE ANALYTICS

Expected and Empirical TCC Alignment: Test A



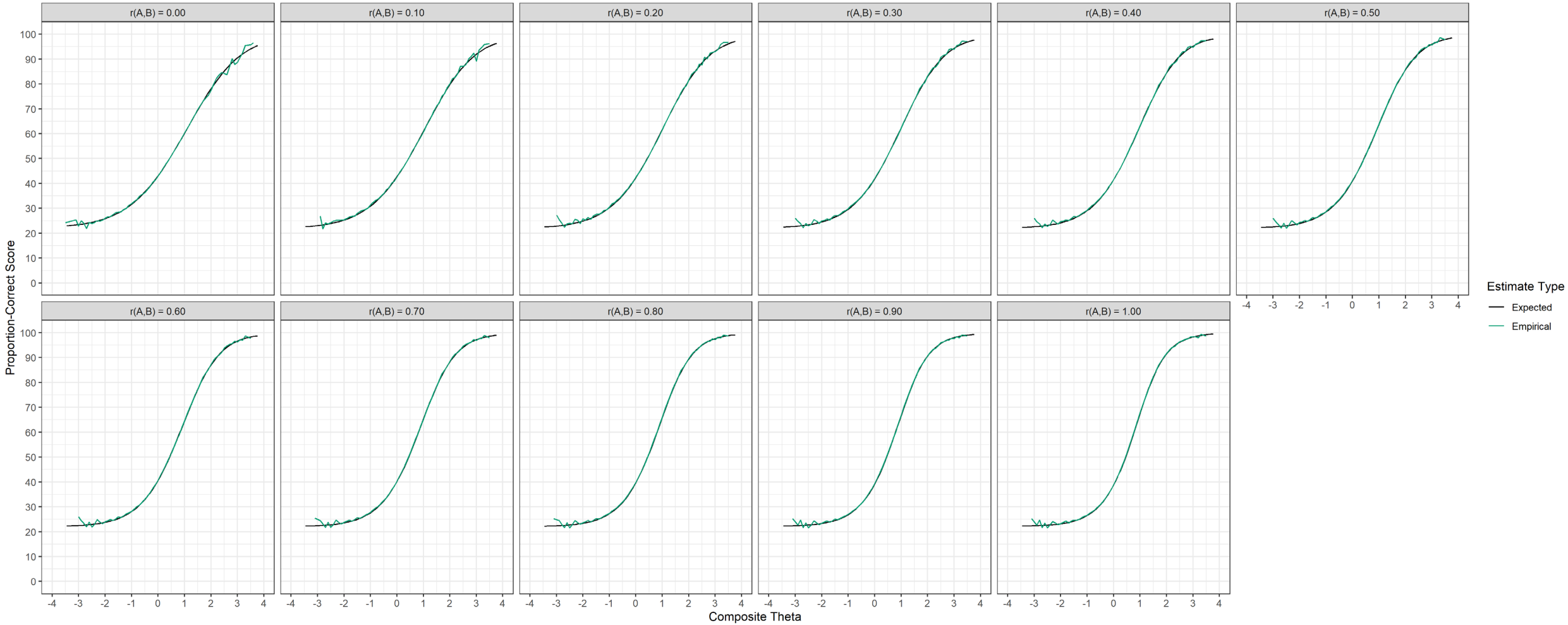
OPA
OFFICE OF PEOPLE ANALYTICS

Expected and Empirical TCC Alignment: Test B



OPA
OFFICE OF PEOPLE ANALYTICS

Expected and Empirical TCC Alignment: Combined Test

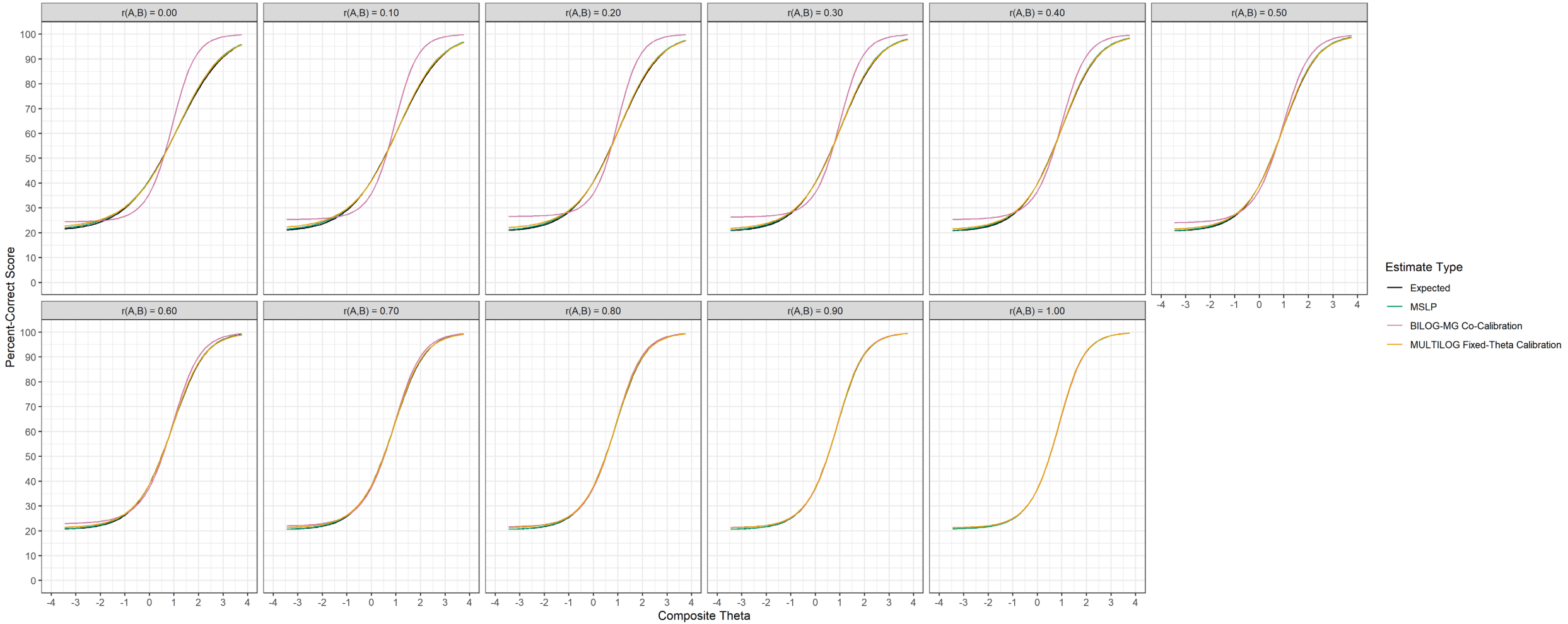


OPA
OFFICE OF PEOPLE ANALYTICS

TCC Comparisons for Rescaling/Calibration Methods

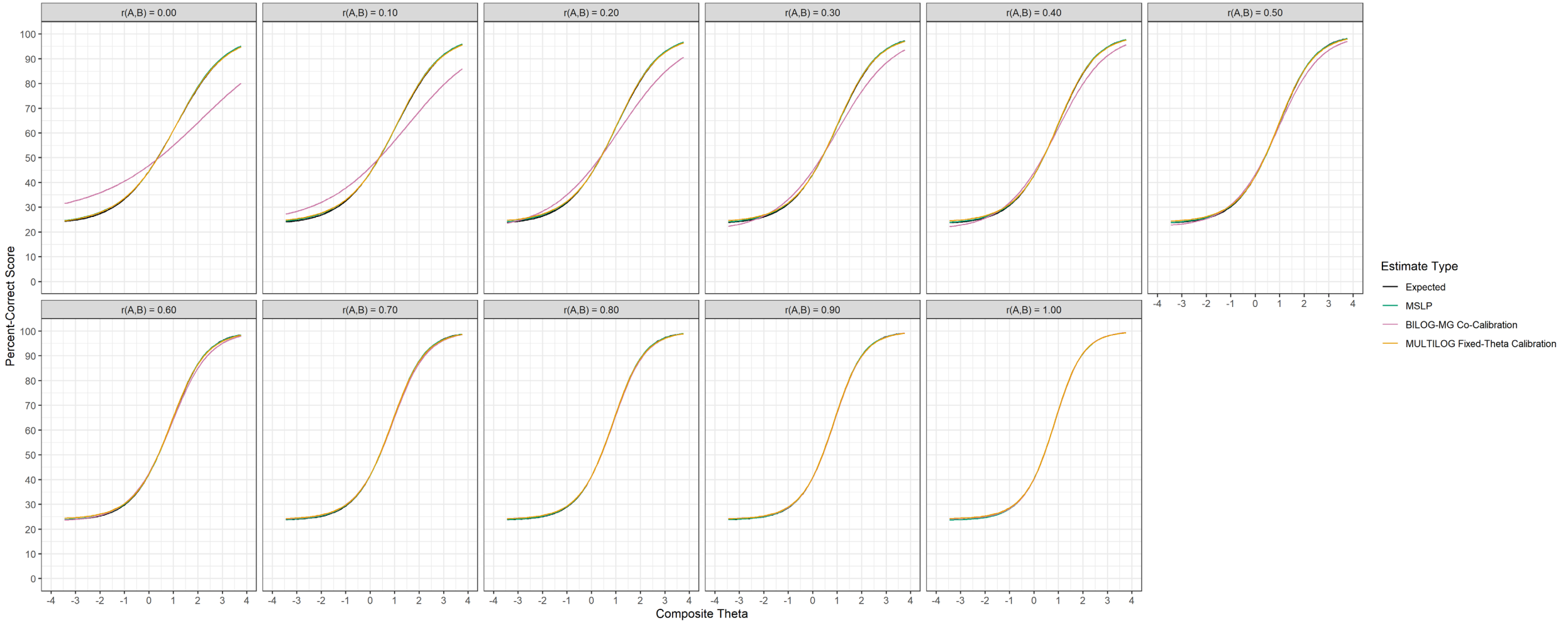
OPA
OFFICE OF PEOPLE ANALYTICS

TCC Comparisons for Rescaling/Calibration Methods: Test A



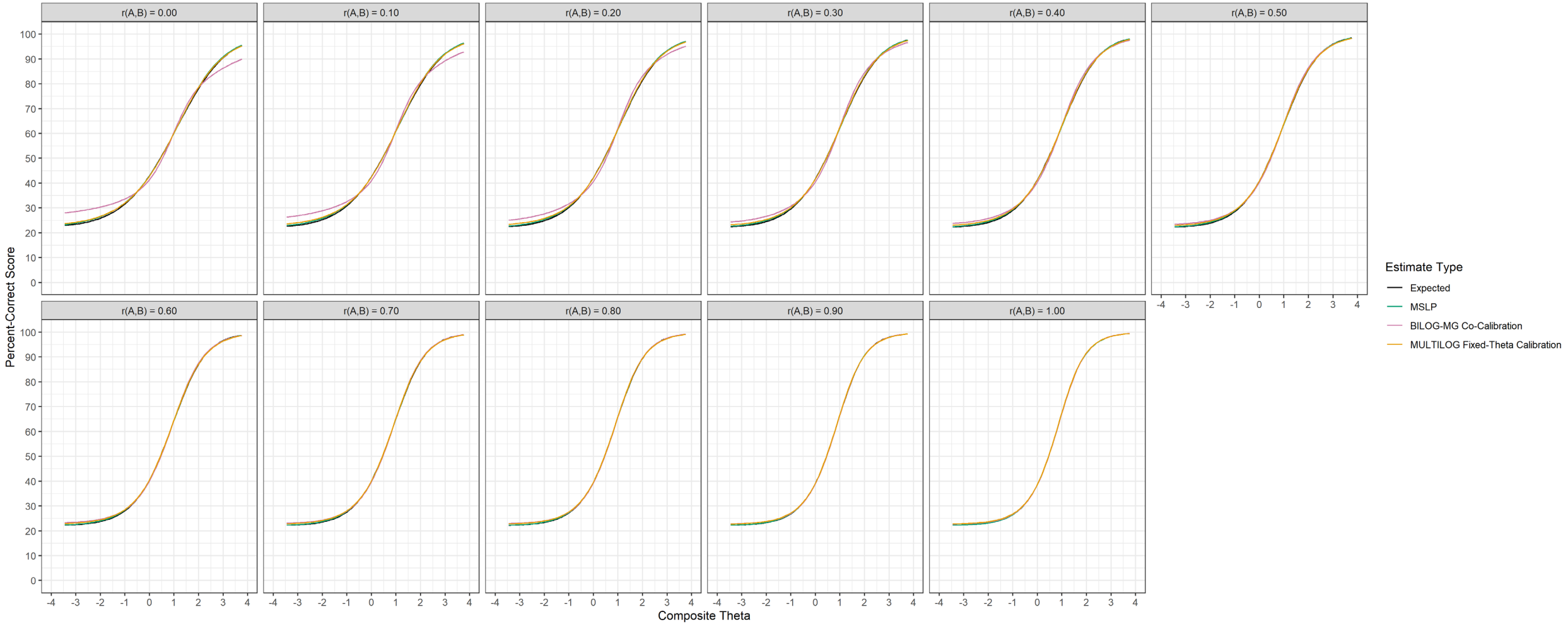
OPA
OFFICE OF PEOPLE ANALYTICS

TCC Comparisons for Rescaling/Calibration Methods: Test B



OPA
OFFICE OF PEOPLE ANALYTICS

TCC Comparisons for Rescaling/Calibration Methods: Combined Test



OPA
OFFICE OF PEOPLE ANALYTICS

Simulation Summary

- The MSLP's expected TCCs were closely aligned with the empirical TCCs associated with the composite theta dimension
  - This supports their use as targets in the rescaling procedure
- The MSLP performed well, even when the dimensions contributing to the composite scale were uncorrelated
  - MSLP-rescaled item parameters produced TCCs that were closely aligned with the expected composite-scaled TCCs
  - MSLP solutions were quite similar to the results from fixed-theta calibrations
  - MSLP solutions were better at recovering expected TCCs than were co-calibrations with BILOG-MG (especially when abilities were correlated < .7)
- The MSLP appears well-suited for this use case

OPA
OFFICE OF PEOPLE ANALYTICS

MSLP Applied to Items Assigned to New P&P-ASVAB Forms

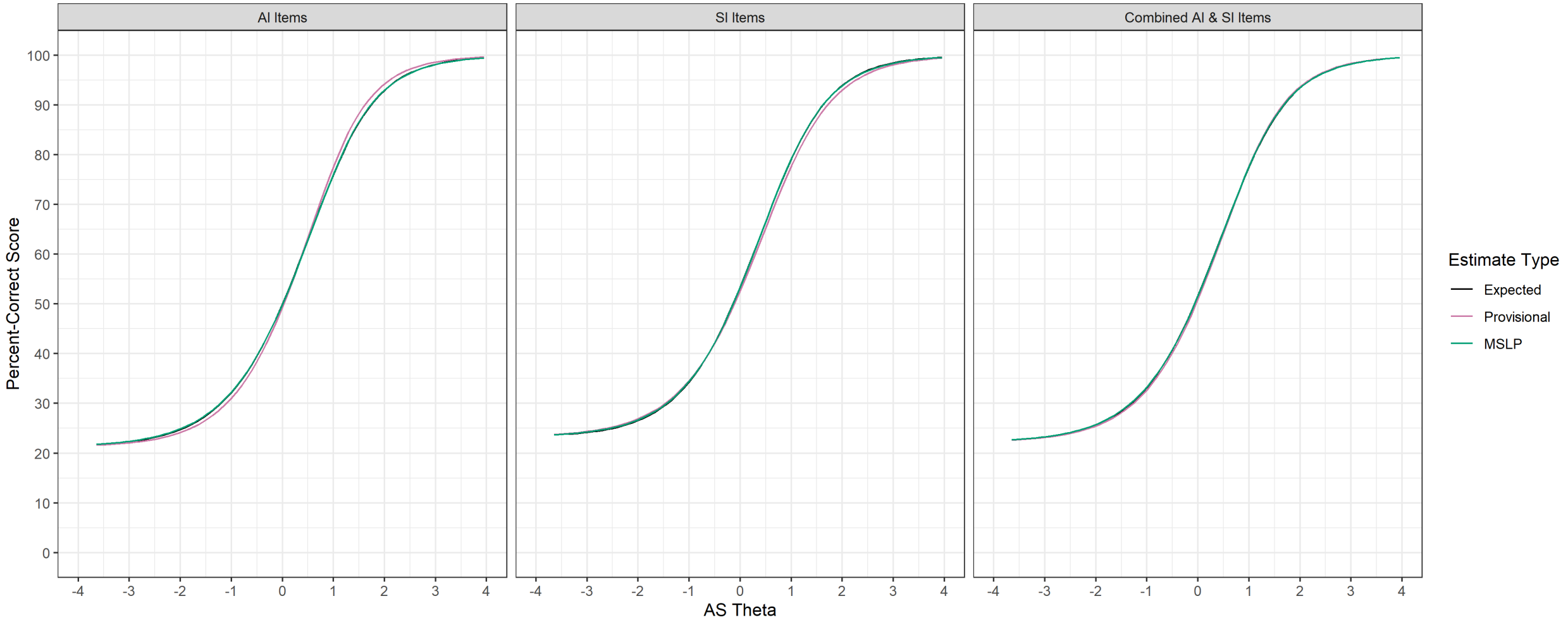
OPA
OFFICE OF PEOPLE ANALYTICS

MSLP Applied to Items Assigned to New P&P-ASVAB Forms

- HumRRO has assembled separate AI and SI item sets that will be administered in the AS sections of the new P&P-ASVAB forms
- The IRT parameters for the items assigned to the AI and SI solutions require rescaling before they can be combined into usable AS sections
- To ensure that item parameters (and resulting theta estimates) are scaled consistently across forms, we applied a single MSLP rescaling to the complete sets of items instead of rescaling each form separately
- We have plotted the rescaled TCCs against the expected TCCs for these item sets
  - As a point of comparison, we have also plotted “provisional” TCCs that ignore the differences in scaling and naively presume that the AI, SI, and AS scales are equivalent

OPA
OFFICE OF PEOPLE ANALYTICS

Expected, Provisional, and MSLP-Rescaled TCCs



OPA
OFFICE OF PEOPLE ANALYTICS

Note: The "provisional" scale represents a naive comingling of item parameters on the AI and SI scales, presuming scale equivalence.

Conclusion

- The MSLP is our recommended approach for obtaining AS-scaled item parameters
- The MSLP's target scale can be defined as a composite scale
  - Scores produced using MSLP-rescaled item parameters represent the expected scores examinees would receive if it were feasible to score AI and SI separately and combine them into a composite
  - Will increase the alignment of AS scaling between P&P-ASVAB and CAT-ASVAB
- The MLSP is effective at mapping item parameters onto a target IRT scale
  - It is more accommodating of multidimensionality than co-calibration of items
  - It does not require item-level data as would be the case with fixed-theta calibrations

Length-Reduction Analyses for Paragraph Comprehension (PC)

OPA
OFFICE OF PEOPLE ANALYTICS

Length-Reduction for New P&P-ASVAB PC Sections

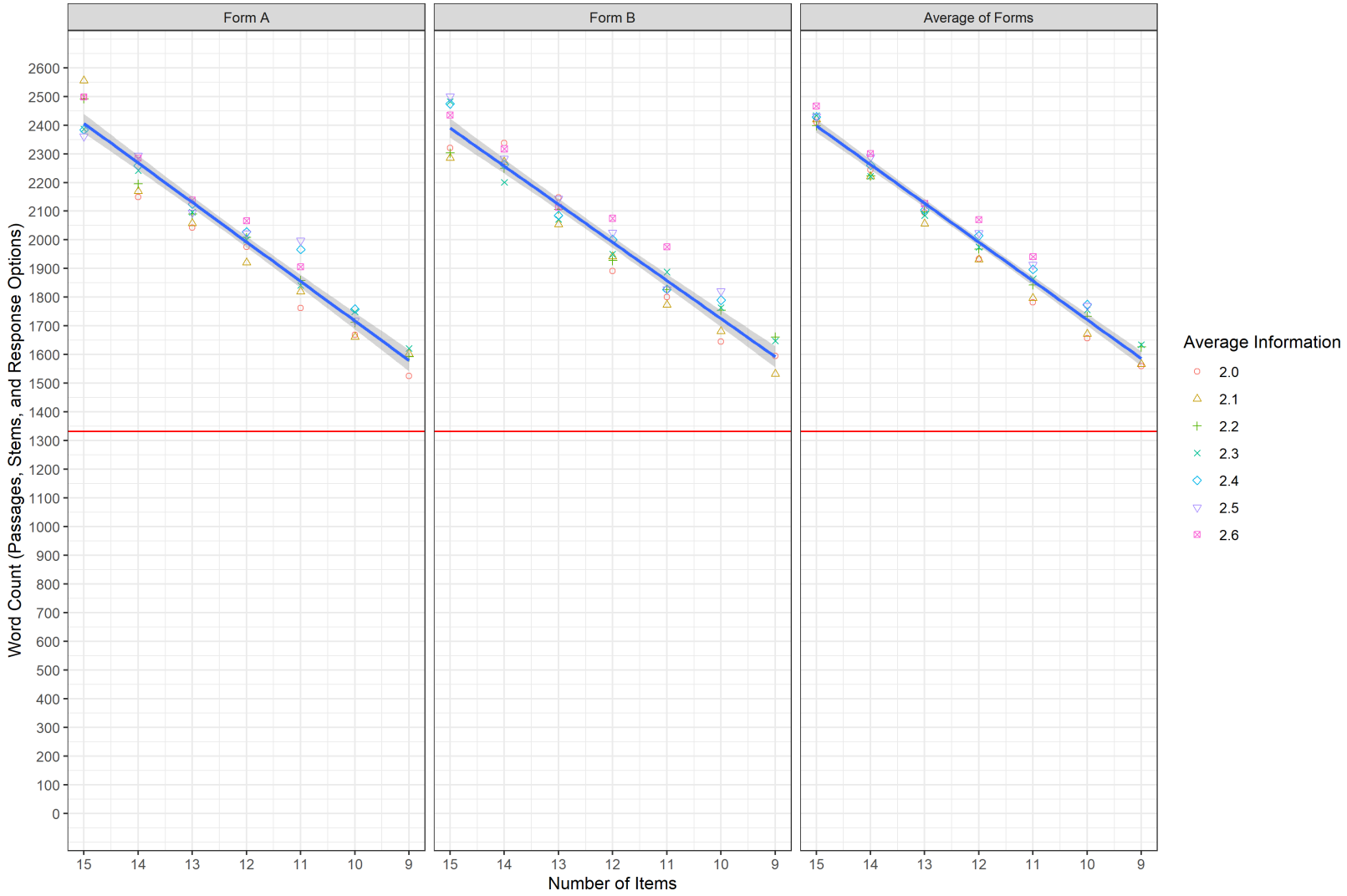
- Compared to past testlet-based PC sections, constructing new PC sections from items with stand-alone reading passages requires reducing the number of items administered to control the reading load
- When shortening a test, there are two primary objectives to satisfy:
  - Maintain an acceptable level of score reliability
  - Maintain adequate coverage of the construct to support score validity
- In addition to these goals, we also aimed to minimize total word count

Impact of P&P-ASVAB PC Section Length on Score Reliability

- To evaluate the effect of form length on reliability, we ran the P&P-ASVAB automated test assembly (ATA) procedure using varied PC specifications:
  - Form length: 9, 10, 11, 12, 13, 14, and 15 items

    fully crossed with
  - Quadrature-Weighted Average IRT information: 2.0, 2.1, 2.2, 2.3, 2.4, 2.5, and 2.6
- Not all combinations of length and information were possible due to the impact of length on test information
  - Forms with 9 items could not achieve average information greater than 2.3
  - Forms with 10 items could not achieve average information greater than 2.5
- We estimated simulated test-retest reliability coefficients for PC scores (BME theta estimates) and composite scores that include PC
  - 10k simulees with abilities based on latent means and SDs

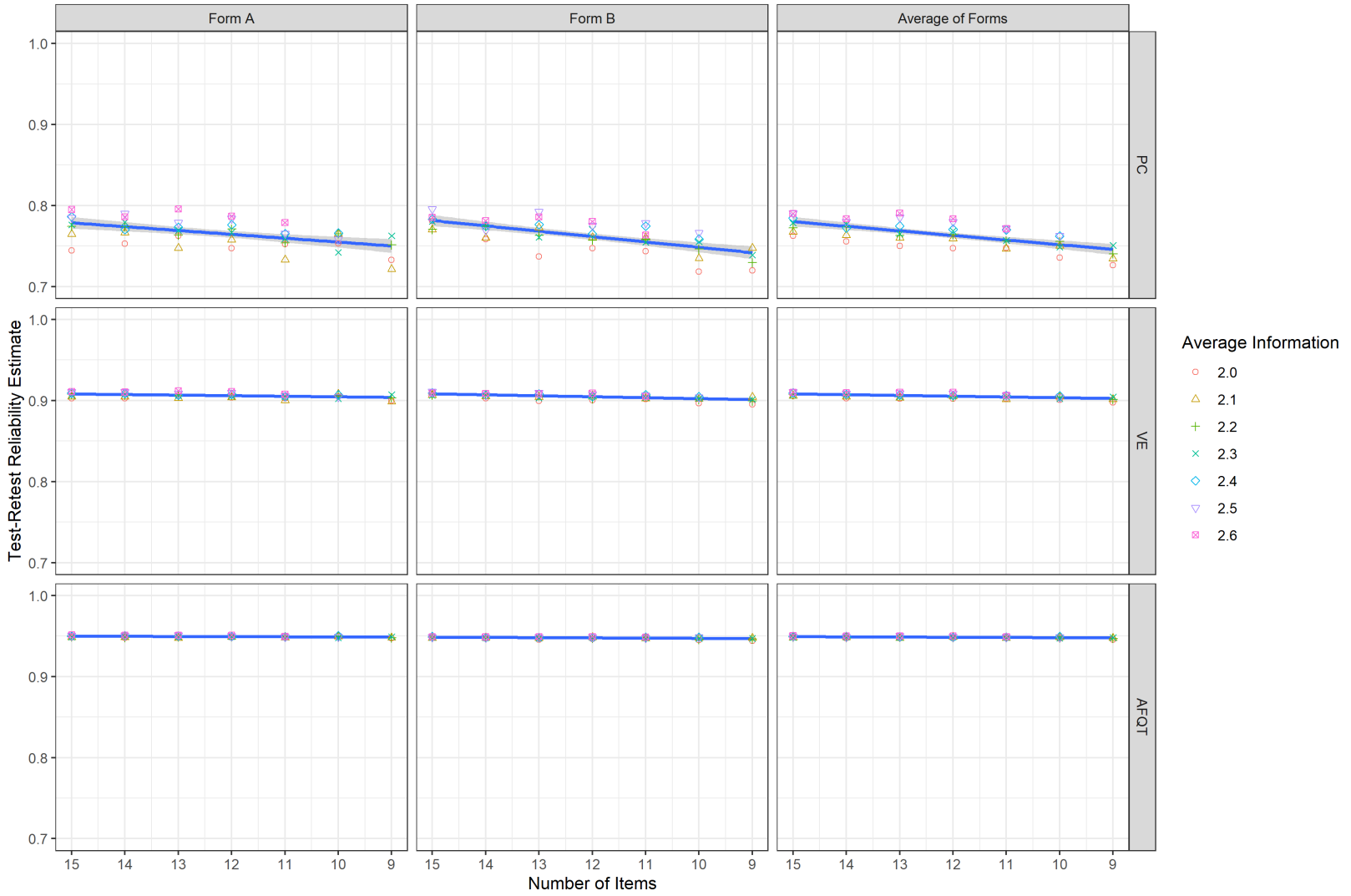
Word Counts for PC Item Sets with Varied Information



OPA
OFFICE OF PEOPLE ANALYTICS

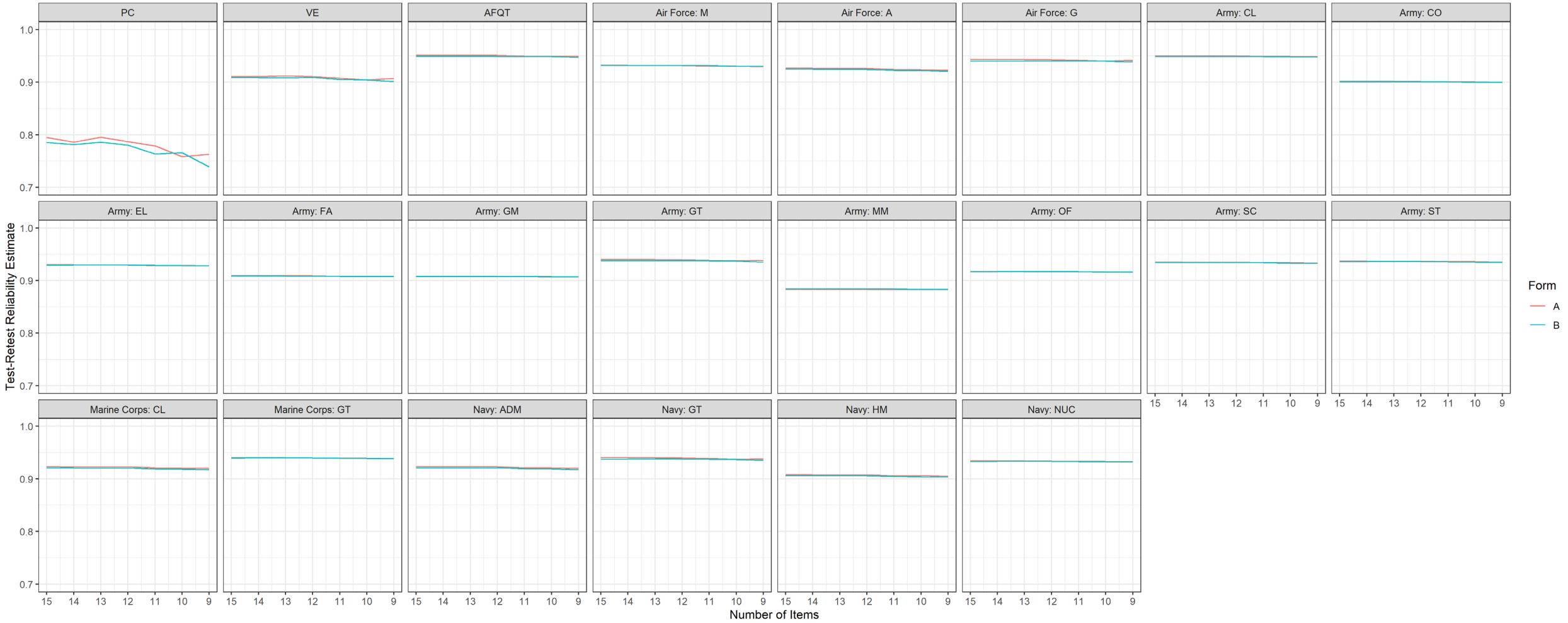
Note: Red horizontal line = average word count of PC subtests from P&P-ASVAB Forms 23 – 26

Test-Retest Reliability for PC Item Sets with Varied Information



OPA
OFFICE OF PEOPLE ANALYTICS

Test-Retest Reliability Estimates for Highest-Information PC Item Sets



OPA
OFFICE OF PEOPLE ANALYTICS

P&P-ASVAB PC Length Recommendation

- **Recommendation:** Reduce P&P-ASVAB PC item sets to 10 stand-alone items and target the highest average information during form development
- PC is already the shortest P&P-ASVAB subtest, and administering 10 items still allows PC to cover its blueprint categories
- 10-item solutions offer competitive levels of reliability compared to other form lengths with a substantially lower reading load
- Using forms with the highest average information corresponds closely to maximizing reliability
  - Reducing the length of the PC subtest (regardless of information) had a trivial impact on the reliability of composites that include PC scores
  - PC scores are never used in isolation for selection or classification into the military, so the impact on composite reliability is more important than PC's stand-alone reliability

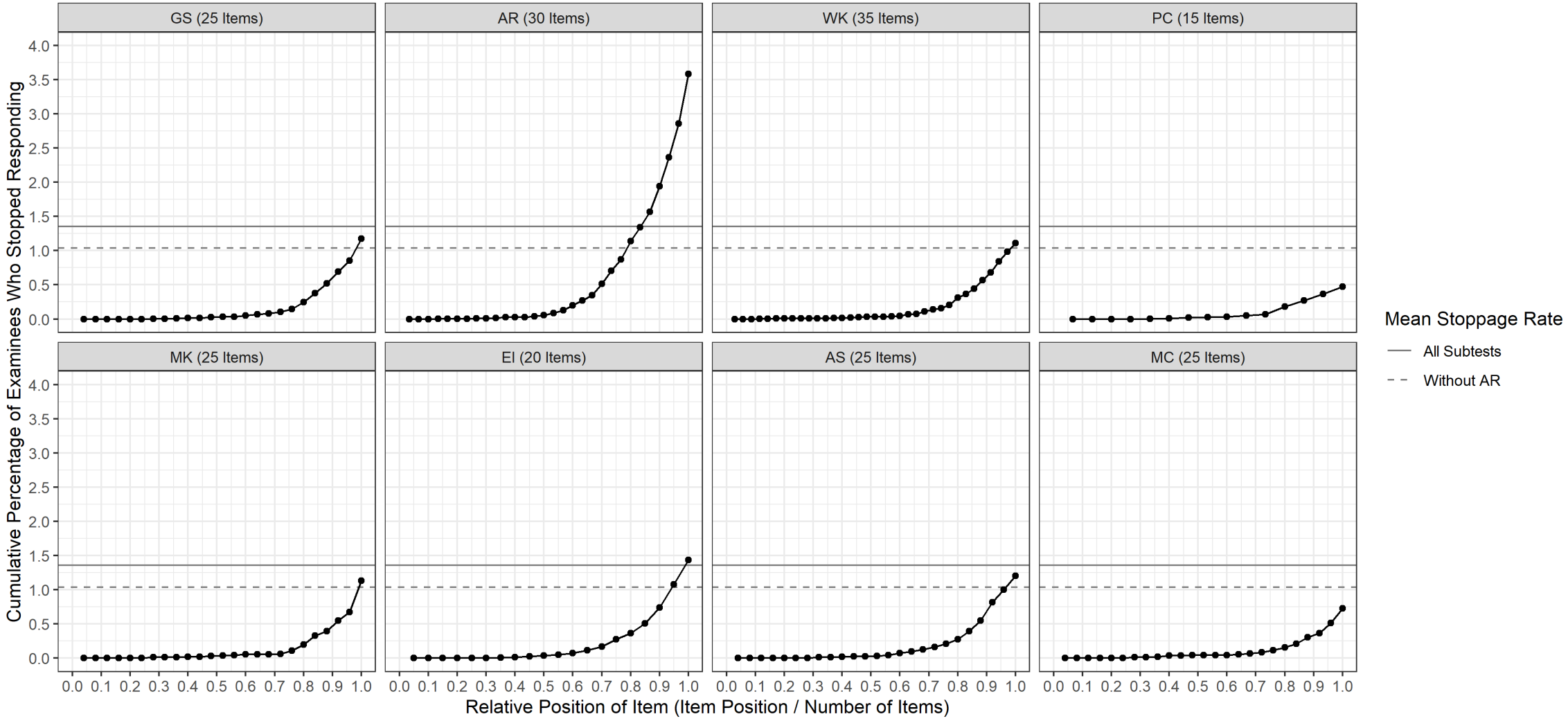
Length-Reduction Analyses for Arithmetic Reasoning (AR)

OPA
OFFICE OF PEOPLE ANALYTICS

Length-Reduction for New P&P-ASVAB AR Sections

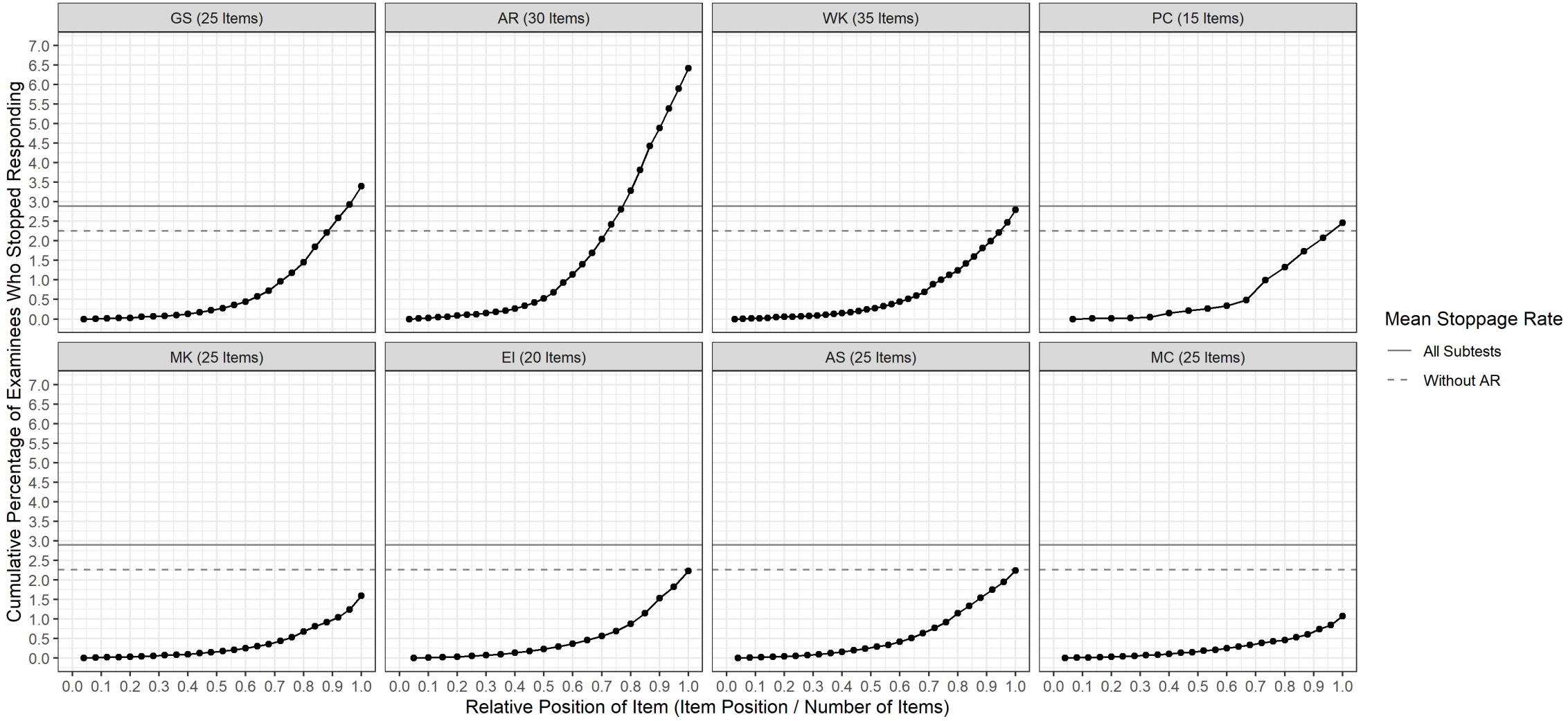
- As we explored the impact of the recommended changes to PC on time limits, we benchmarked whether past P&P-ASVAB PC sections appeared to have sufficient time limits
- We examined trends from all P&P-ASVAB subtests to provide context for the PC trends
- We found that AR appeared to be much more speeded than the other subtests
  - 3.5% of ETP P&P-ASVAB examinees failed to complete the AR section, while only an average of only 1% failed to complete each of the other subtests
  - This trend generalized to the CEP P&P-ASVAB, but with higher overall non-completion rates (likely due to a less-motivated examinee population)
    - 6.5% non-completion rate for AR
    - 2.75% average non-completion rate for other subtests
- For reference, the CAT-ASVAB time limits are designed to target a 99% completion rate (Gao, Pommerich, & Segall, 2019)

Speededness Evaluations for Current Operational ETP P&P-ASVAB Forms



OPA
OFFICE OF PEOPLE ANALYTICS

Speededness Evaluations for Current Operational CEP P&P-ASVAB Forms

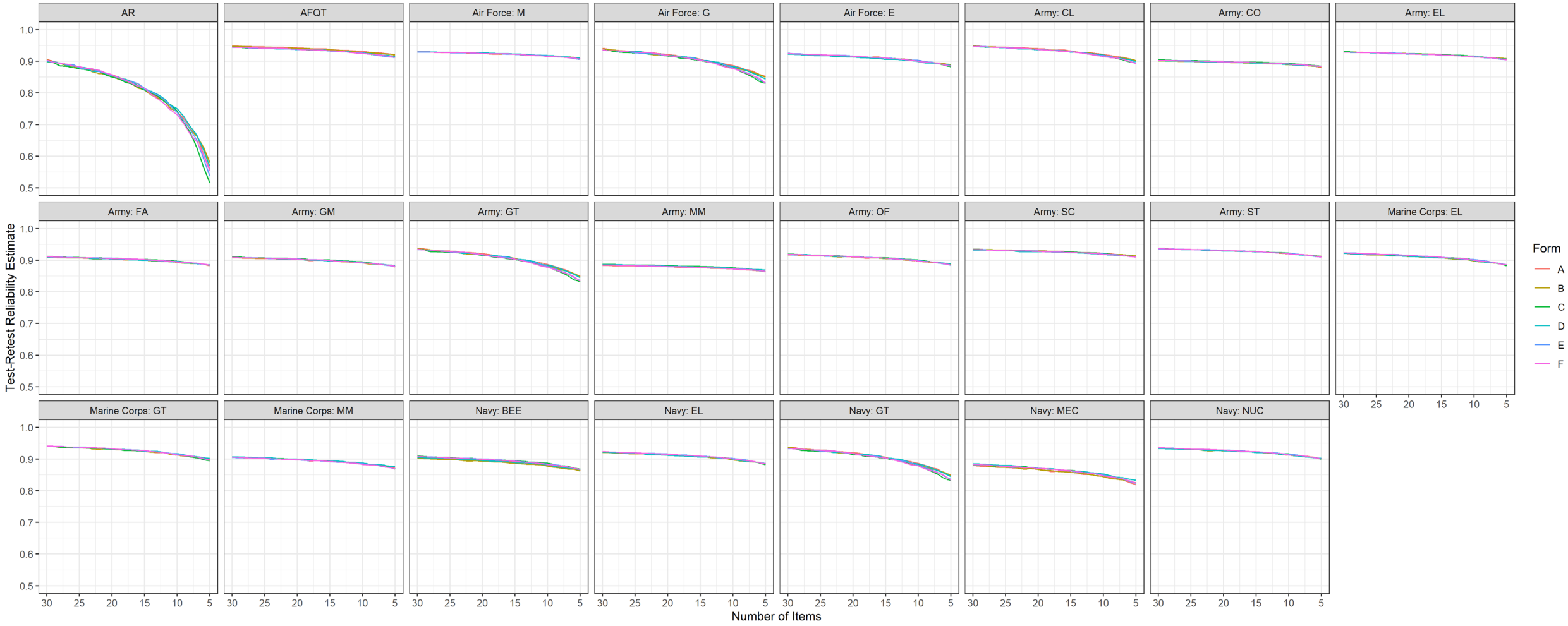


OPA
OFFICE OF PEOPLE ANALYTICS

Impact of P&P-ASVAB AR Section Length on Score Reliability

- Before we began evaluating the impact of reducing the number of items in P&P-ASVAB AR sections, we had assembled six new 30-item sections
  - The sections had gone through all necessary reviews
  - They were free of enemy items and passed all other content checks
- Rather than start over and repeat a painstaking form assembly/review process, we used these existing sections as the basis for reduced-length sections
  - We explored the impact on simulated score reliability when the least reliable item was iteratively removed from each form, examining solutions with between 5 and 30 items
- We used the shortened item sets to simulate test-retest reliability coefficients for AR scores (BME theta estimates) and composite scores that include AR
  - 10k simulees with abilities based on latent means and SDs

Test-Retest Reliability Estimates AR Item Sets

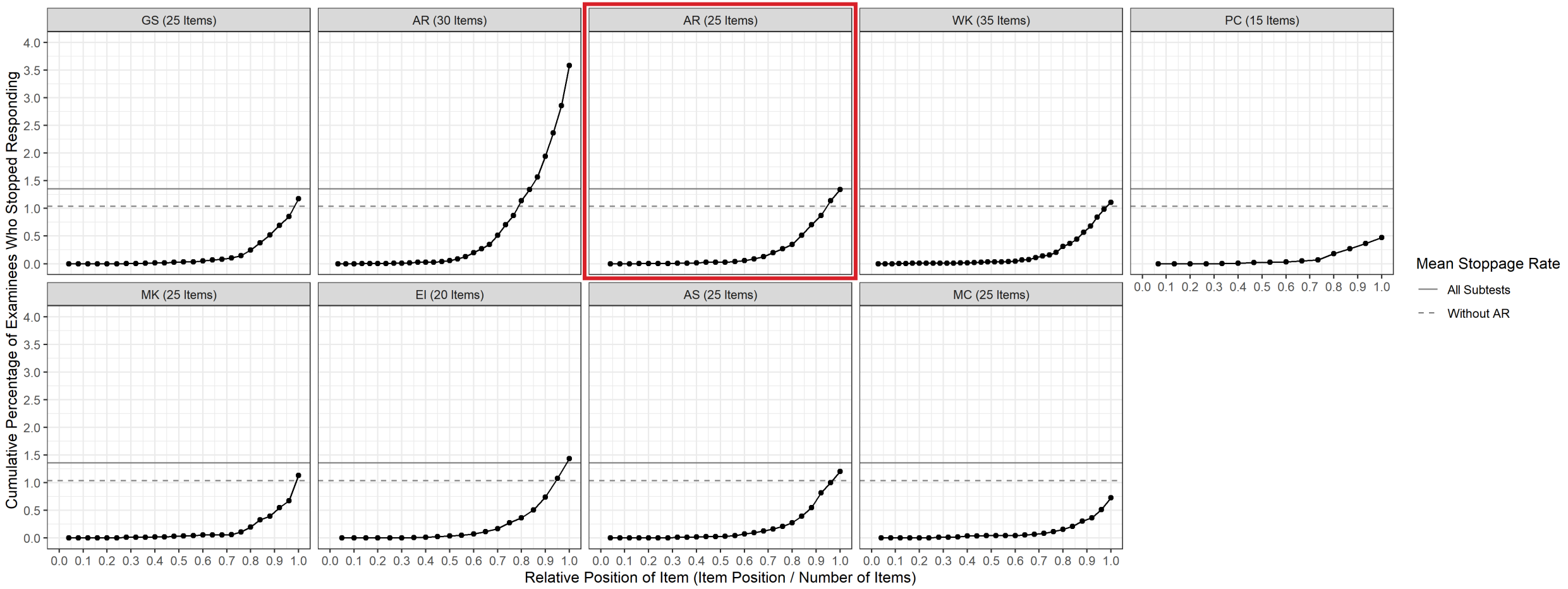


OPA
OFFICE OF PEOPLE ANALYTICS

Speededness Evaluations for Past ETP P&P-ASVAB Forms (Again)

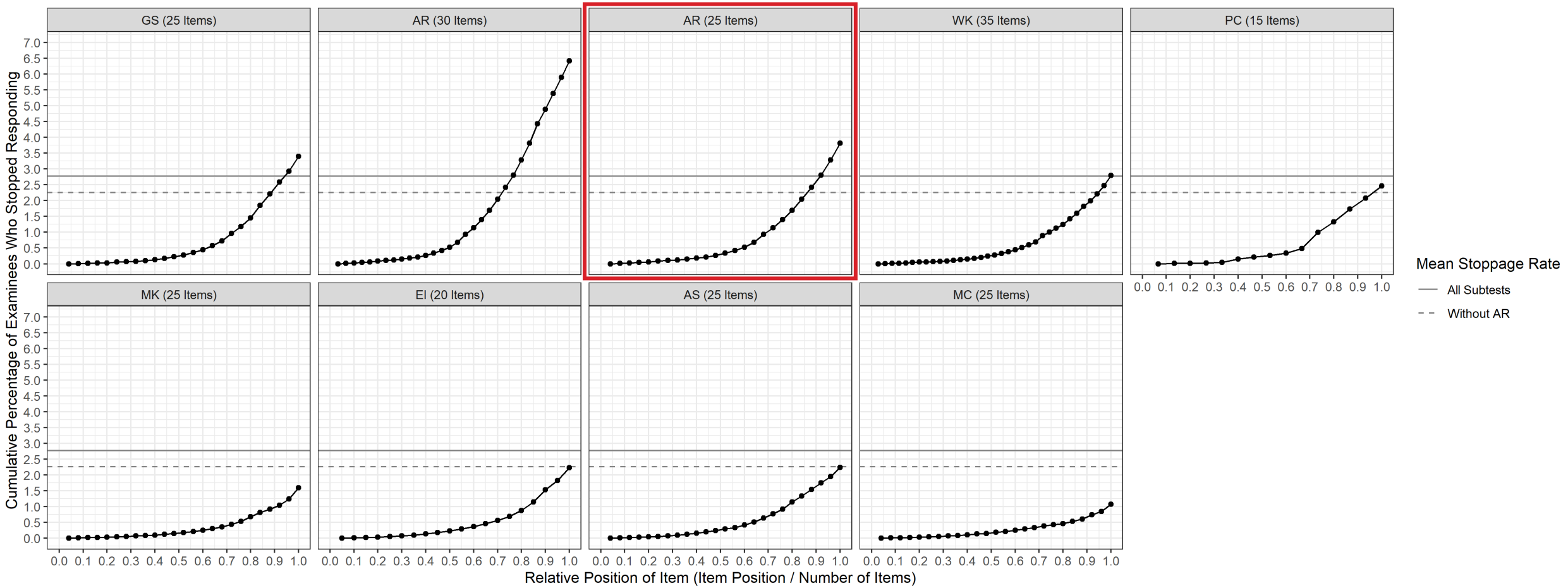
- Based on simulated reliability estimates, 25 items appear to preserve reliability across all scores we evaluated
  - This length also works well for covering all test blueprint categories
  - As with PC, AR scores are never used in isolation for selection or classification into the military, so the impact on composite reliability is more important than AR's stand-alone reliability
- With this length in mind, we re-examined the speededness trends for past P&P-ASVAB forms, omitting the last 5 AR items from the analyses
  - These analyses can give a sense of whether shifting from 30 to 25 items is enough of a reduction to mitigate the speededness we observed
  - Not a perfect approach: The last 5 items are also among the most difficult, so we must consider that these results are slightly optimistic
    - Especially true for CEP, where examinees have lower motivation

Speededness Evaluations for Past ETP P&P-ASVAB Forms (with Truncated AR Section)



OPA
OFFICE OF PEOPLE ANALYTICS

Speededness Evaluations for Past CEP P&P-ASVAB Forms (with Truncated AR Section)



OPA
OFFICE OF PEOPLE ANALYTICS

P&P-ASVAB AR Length Recommendations

- Reduce AR item sets from 30 items to 25
  - 25-item sections allow scores to retain high levels of reliability
  - Based on evaluations of response data from past P&P-ASVAB forms, 25 items seems to be a sufficient length to mitigate the speededness concerns that motivated this research
  - Using 25 items allows good coverage of all test blueprint categories
- Use the 30-item AR sets that have already been built and reviewed as the basis for the reduced-length forms, and remove 5 items from each
  - Remove items based on their contributions to reliability, and balance removals across content areas

Time Limit Adjustments

OPA
OFFICE OF PEOPLE ANALYTICS

Time Limit Adjustments

- Even after reducing the number of items in the new P&P-ASVAB PC sections, the 10-item sets had higher word counts than past PC sections
  - The greater reading demands of the new sections requires allocating more time to PC to avoid introducing speededness
- We examined the reading demands of the new PC sections and the PC sections from past forms to estimate the necessary time limit adjustment
- We also considered whether any other subtests could be donors of this additional time, to avoid increasing the total total battery-wide testing time

OPA
OFFICE OF PEOPLE ANALYTICS

Reading Load Analyses for P&P-ASVAB PC Sections

- Evaluated PC sections on five common reading metrics:
  - Word Count
  - Flesch-Kincaid Age
  - Flesch-Kincaid Grade Level
  - Flesch Reading Ease
- Because estimates of the Flesch-Kincaid and Flesch metrics can vary across programs, we used two programs to compute them:
  - TreeTagger (a part-of-speech tagger and lemmatization program; Schmid, 1994)
  - Microsoft Word

Summaries and Comparisons of Readability Metrics for the Previous and New Generations of P&P-ASVAB Forms

Readability Metric	Previous Generation of P&P-ASVAB Forms		New Generation of P&P-ASVAB Forms		Mean Difference (New – Previous)	Percentage Increase Relative to Previous Generation's Mean
	Mean	SD	Mean	SD		
Word Count	1332.17	91.64	1704.33	45.62	372.17	27.94
Flesch-Kincaid Age (TreeTagger)	13.68	0.78	15.13	0.66	1.45	10.60
Flesch-Kincaid Grade Level (TreeTagger)	8.67	0.79	10.14	0.66	1.47	16.98
Flesch-Kincaid Grade Level (MS Word)	9.72	0.52	11.08	0.89	1.37	14.07
Flesch Reading Ease (TreeTagger)	62.35	4.54	55.85	3.96	-6.49	-10.41
Flesch Reading Ease (MS Word)	54.98	2.64	49.32	4.68	-5.67	-10.31

OPA
OFFICE OF PEOPLE ANALYTICS

Recommended PC Time Limit Adjustment Based on Reading Demands

- The current P&P-ASVAB time limit for PC is 13 minutes
  - Because P&P-ASVAB is timed for groups of examinees rather than individuals, the time limit should allow most examinees to finish
  - However, the time limit should not be set too high or examinees who complete the section more quickly will have to wait longer for others to finish
- We considered both word count and overall reading complexity:
  - Based on word count alone, a time limit of 17 minutes would be appropriate
    - 13 x 1.2794 = 16.6322 minutes
  - However, our reading complexity metrics suggested a roughly 10% change in the reading ease compared to the past forms
  - Based on both word count and complexity, **18 minutes** should be appropriate
    - 13 x 1.2794 x 1.10 = 18.295 minutes

Estimating Time Required to Complete P&P-ASVAB Forms

- We used response latency data from CAT-ASVAB test records to estimate how much time examinees would likely need to complete the new P&P-ASVAB forms
- This evaluation was meant to indicate which subtests could most likely be administered with shorter time limits to make up for the five-minute increase required for PC since increasing the battery-wide time limit is not feasible
- We used a five-step process to estimate the amount of time examinees would need in order to respond to all items on a new form

OPA
OFFICE OF PEOPLE ANALYTICS

Estimating Time Required to Complete P&P-ASVAB Forms (Continued)

1. We computed a response latency score for each examinee on each subtest based on their responses to tryout (i.e., unscored) items
   - Computed the mean and standard deviation of response latencies for each item
   - Used the means and SDs for response latencies to convert all examinees' item-level response latencies to Z scores
   - Averaged each examinee's item-level response latency Z scores across items within each subtest to get their composite response latency score for that subtest
   - Converted examinees' composite response latency estimates to percentiles within each subtest, then organized them into twenty equally sized ordinal categories, each of which spanned a range of five percentiles (e.g., the slowest response category included examinees who were at or above the 95th percentile)
2. For each tryout item, we computed the mean amount of time examinees from each response latency percentile category spent answering the item

Estimating Time Required to Complete P&P-ASVAB Forms (Continued)

3. Some items assigned to the new P&P-ASVAB forms predated the CAT-ASVAB data that we processed in Steps 1 and 2, so we used linear regression analyses to impute missing item-level response latencies for each response latency percentile category
   - These imputation models based their predictions on items' 3PL IRT item parameters (difficulty, pseudo-guessing, and discrimination) and—for PC only—word counts
4. We merged our complete database of item-level response latencies with assembled forms' item lists and computed the sum of item-level latencies for each response latency percentile category within each form
5. For each subtest, we computed the mean estimated test time across forms for each response latency percentile category to arrive at an overall summary of how much time examinees in each percentile category would require to complete an average form

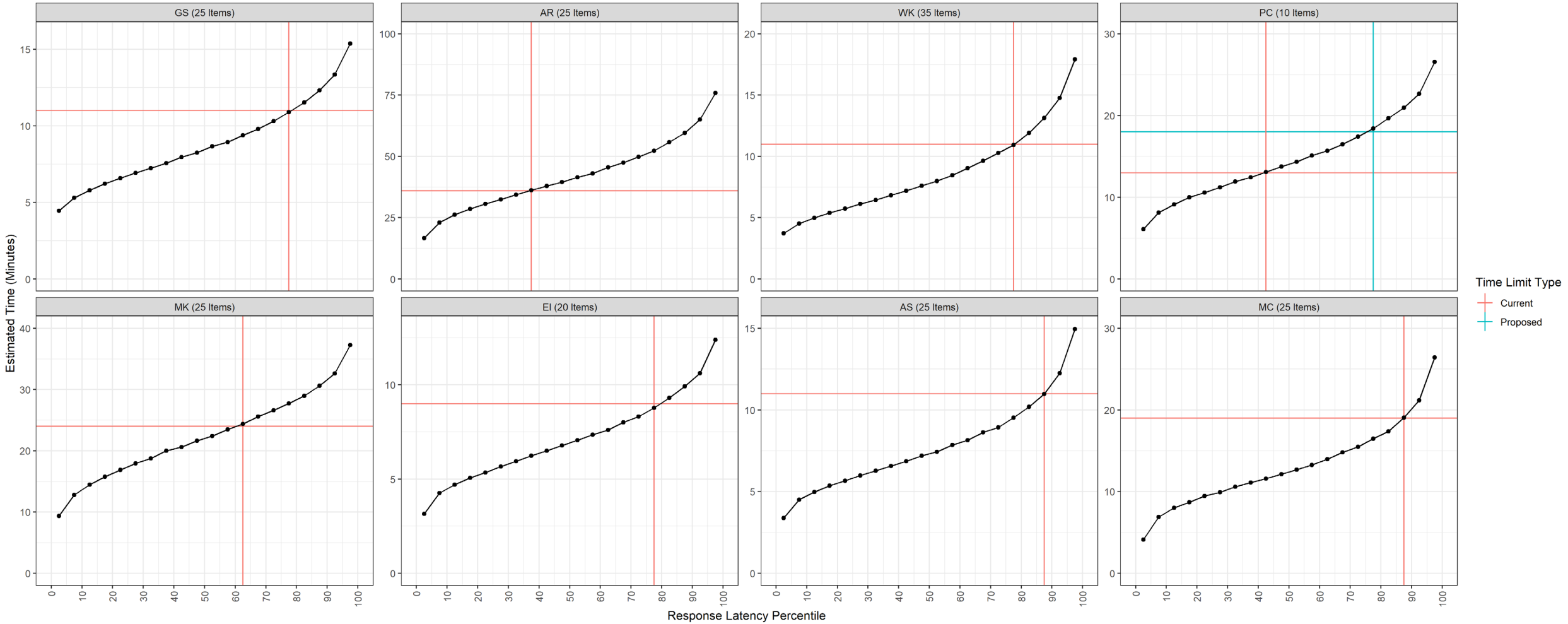
Context for Interpreting the Projected Time Requirements

- CAT-ASVAB and P&P-ASVAB have different item-level time allowances (esp. for AR), so we must generalize from CAT-ASVAB to P&P-ASVAB with care
  - Examinees likely use their time differently when time allowances differ

Subtest	Minutes per Item			CAT / P&P Minutes-per-Item Ratio		
	CAT-ASVAB		P&P-ASVAB	CAT W/O Tryout	CAT W/ Tryout	Average
	W/O Tryout	W/ Tryout				
GS	0.80	0.83	0.44	1.82	1.89	**1.86**
AR[1]	3.67	3.77	1.20	2.55	2.62	**2.58**
WK	0.60	0.60	0.31	1.91	1.91	**1.91**
PC[2]	2.70	3.00	1.80	1.50	1.67	**1.58**
MK	2.07	2.17	0.96	2.15	2.26	**2.20**
EI	0.67	0.70	0.45	1.48	1.56	**1.52**
AS[3]	0.65	0.70	0.44	1.48	1.59	**1.53**
MC	1.47	1.40	0.76	1.93	1.84	**1.89**

[1] P&P-ASVAB values for AR are based on the recommended 25-item and 36-minute configuration.
[2] P&P-ASVAB values for PC are based on the recommended 10-item and 18-minute configuration.
[3] CAT-ASVAB AS values are based on AI and SI combined; the "W/ Tryout" estimates for AS are approximate because AI and SI items are tried out with non-overlapping samples of examinees.

Relations Between Response Latency Percentiles and Projected Testing Time



OPA
OFFICE OF PEOPLE ANALYTICS

Summary of Recommendations

OPA
OFFICE OF PEOPLE ANALYTICS

Recommended Alterations to P&P-ASVAB Specifications for New Forms

- Use the newly developed MSLP rescaling technique to translate IRT item parameters for AI and SI items onto an AS scale
- Reduce the number of PC items from 15 to 10
  - Administering fewer PC items offsets the increased text in the passages caused by shifting from a testlet design to the use of stand-alone items
  - Using 10 items is sufficient to maintain acceptable reliability for composite scores
- Reduce the number of AR items from 30 to 25
  - Previous P&P-ASVAB forms showed evidence of speededness
  - Using 25 items is sufficient to maintain acceptable reliability for composite scores while mitigating speededness effects
- Adjust time limits to account for increased PC reading load
  - Even after reducing the number of PC items, the reading load of new PC sections will be greater than past PC sections
  - To offset this, we recommend increasing the PC time limit from 13 to 18 minutes

Suggested Options for Time Limit Adjustments

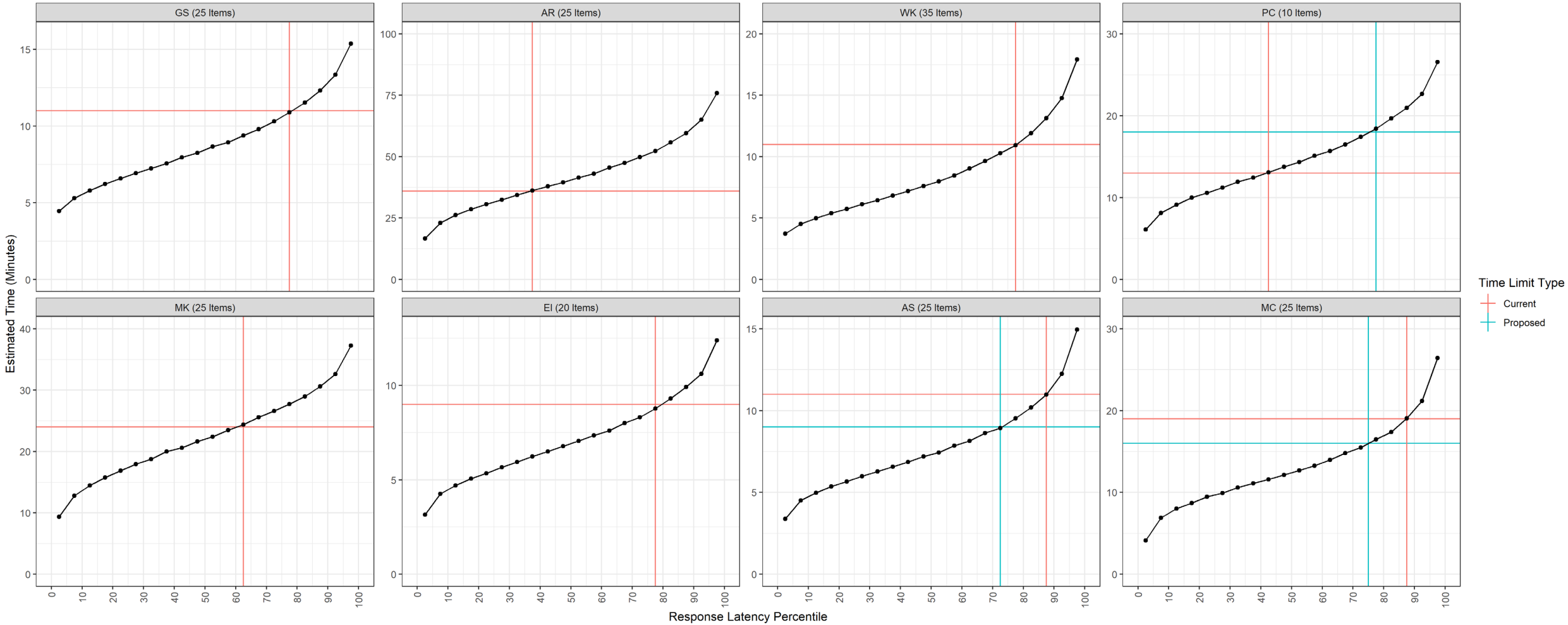
- **Option A**
  - Increase time limit for PC without altering other subtests' time limits
- **Option B (recommended)**
  - Offset the increased PC time limit by reducing time limits for AS and MC

Subtest	Item Count	Previous Time Limit (Minutes)	Recommended Time Limit (Minutes)			
			Option A		Option B	
			Limit	Δ	Limit	Δ
GS	25	11	11	0	11	0
AR	25	36	36	0	36	0
WK	35	11	11	0	11	0
PC	10	13	18	+5	18	+5
MK	25	24	24	0	24	0
EI	20	9	9	0	9	0
AS	25	11	11	0	9	-2
MC	25	19	19	0	16	-3
Total	**195**	**134**	**139**	**+5**	**134**	**0**

Note: The ETP P&P-ASVAB also includes a 25-item, 15-minute AO section.

OPA
OFFICE OF PEOPLE ANALYTICS

Relations Between Response Latency Percentiles and Projected Testing Time



OPA
OFFICE OF PEOPLE ANALYTICS

Questions for the DAC

OPA
OFFICE OF PEOPLE ANALYTICS

Questions for the DAC

- Does the DAC concur with our use of the Modified Stocking-Lord Procedure (MSLP) to resolve the AS scaling problem for P&P-ASVAB?
- Does the DAC concur with the recommended lengths for the PC (10 items) and AR (25 items) P&P-ASVAB sections?
- Does the DAC concur with our recommended P&P-ASVAB time limit adjustments to account for the new PC sections' increased time requirements?

OPA
OFFICE OF PEOPLE ANALYTICS

Thank You!

For more information,
please contact:

Jeff Dahlke
jdahlke@humrro.org
jeffrey.a.dahlke.ctr@mail.mil

OPA
OFFICE OF PEOPLE ANALYTICS

Supplemental Slides

Paragraph Comprehension (PC) Reading Load Analyses by Form

OPA
OFFICE OF PEOPLE ANALYTICS

Word Counts for the Previous and New Generations of P&P-ASVAB Forms

P&P-ASVAB Generation	Form/ Item Set	Testing Program	Word Count	Percentage Increase Relative to Previous Generation's Mean
Previous (15 items per form; 5 items for each of three passages)	23A/B	CEP	1,329	
	24A/B	CEP	1,167	
	25A	ETP	1,370	
	25B	ETP	1,418	
	26A	ETP	1,305	
	26B	ETP	1,404	
	Average	---	1,332	
New (10 items per form; 1 item per passage)	A	CEP	1,708	**28.21**
	B	CEP	1,692	**27.01**
	C	ETP	1,755	**31.74**
	D	ETP	1,692	**27.01**
	E	ETP	1,749	**31.29**
	E	ETP	1,630	**22.36**
	Average	---	1,704	**27.94**

OPA
OFFICE OF PEOPLE ANALYTICS

Form-Level Readability Metrics for the Previous and New Generations of P&P-ASVAB Forms

P&P-ASVAB Generation	Program	Form/ Item Set	Word Count	Flesch-Kincaid Age (TreeTagger)	Flesch-Kincaid Grade Level		Flesch Reading Ease	
					TreeTagger	MS Word	TreeTagger	MS Word
Previous	CEP	23A/B	1,329	13.90	8.93	10.10	63.35	54.50
	CEP	24A/B	1,167	13.40	8.40	9.30	63.01	57.90
	ETP	25A	1,370	13.90	8.87	9.80	61.40	54.10
	ETP	25B	1,418	14.80	9.76	10.50	54.90	50.60
	ETP	26A	1,305	13.70	8.69	9.50	62.33	55.40
	ETP	26B	1,404	12.40	7.35	9.10	69.08	57.40
	---	Mean	1,332	13.68	8.67	9.72	62.35	54.98
New	CEP	A	1,708	15.50	10.55	11.90	52.01	43.60
	CEP	B	1,692	14.30	9.35	9.90	60.70	55.70
	ETP	C	1,755	14.90	9.90	11.10	58.32	50.30
	ETP	D	1,692	14.50	9.46	10.10	58.49	53.20
	ETP	E	1,749	15.70	10.68	11.60	54.85	48.10
	ETP	F	1,630	15.90	10.89	11.90	50.75	45.00
	---	Mean	1,704	15.13	10.14	11.08	55.85	49.32

OPA
OFFICE OF PEOPLE ANALYTICS

Tab I

Recommended Updates to the CAT-AVAB Equating Design

Jeff Dahlke

Human Resources Research Organization

Briefing presented to the DACMPT
January 22, 2025

Agenda

- Background: Overview of the current CAT-ASVAB equating design
- Follow-up analyses requested by the DACMPT in June 2024
  - Simulated bias from using provisional transformation constants (no equating)
- Equating design evaluations using equating study data from Forms 11–15
  - Sample size per form
  - Allocation of the sample across equating phases
- Summary of recommended alterations to the CAT-ASVAB equating design

Background: Overview of the Current CAT-ASVAB Equating Design

OPA
OFFICE OF PEOPLE ANALYTICS

Overview of CAT-ASVAB Scale Maintenance Procedures

- The consistency of scaling for newly developed CAT-ASVAB forms is maintained via a two-stage process:
  1. Item Response Theory (IRT) Rescaling
     - Maintains the scale for IRT item parameter and person parameter estimates
     - After new items are calibrated, their IRT parameters are rescaled to match the scaling of parameters for existing operational items
  2. Standard Score Equating
     - Maintains the scale of standard scores (the reporting metric for scores) to ensure they are linked to relevant norms (currently, 1997 Profile of American Youth [PAY97] norms)
     - New forms are administered with a reference form in an equating study to derive linear transformation constants (TCs) for converting IRT theta-metric scores to standard scores
       - Equating ensures the means and standard deviations of standard scores for the new forms equal those of the reference form

CAT-ASVAB Equating: Design Overview

- Linear equating methods are used to derive TCs to transform IRT-based theta scores ($\hat{\theta}$) on new forms to match the scale of the reference form in a phased approach
  - Done for each subtest and for the Auto & Shop Information (AS) and Verbal (VE) composites
- Random-groups design
  - Each applicant is assigned to a single form with equal assignment probability
    - The reference form (administered only during equating studies)
    - An operational form (a form from the previous set of CAT-ASVAB forms)
    - A new form
  - New forms initially inherit the TCs from the reference form
    - New forms' TCs are progressively adjusted over three phases as their sample sizes increase
      - Final sample size goal = 10k per form
    - TCs for the reference form and operational form do *not* undergo adjustment during this process
- Objective: Arrive at a final set of TCs for each new form that will produce standard score distributions with the same mean and SD as the reference form

CAT-ASVAB Equating: Mechanics of the Process

- A set of pre-established reference form TCs exists for each standard score
  - A set of TCs consists of intercept and slope coefficients
    - One slope for determining standard scores for individual subtests, two slopes for composites (AS and VE)
  - These serve as the starting point for establishing new forms' TCs
- When new forms are administered during equating, we collect distributions of theta estimates for the new forms and the reference form
  - These distributions inform adjustments to the reference form's TCs to fit the new forms
  - For individual subtests, reference form TCs (α = intercept; β = slope) are adjusted to fit a new form as follows:
    - $\alpha_{Equated} = \alpha_{Reference} + \beta_{Reference}\left(\mu_{\hat{\theta}_{Reference}} - \frac{\sigma_{\hat{\theta}_{Reference}}}{\sigma_{\hat{\theta}_{New}}}\mu_{\hat{\theta}_{New}}\right)$
    - $\beta_{Equated} = \beta_{Reference}\frac{\sigma_{\hat{\theta}_{Reference}}}{\sigma_{\hat{\theta}_{New}}}$
  - This is identical to the process one would use to adjust regression coefficients to account for a change to the scaling of predictors/features used in a model
  - Process for AS and VE is similar, but also accounts for contributing subtest scores' covariance

CAT-ASVAB Equating: Refinement of Transformations over Three Phases

- Equating is implemented in three phases of operational administration of new forms to military applicants
  - Each phase uses a progressively larger sample size (final goal = 10k per form)
  - Phase sample sizes are cumulative such that they include all individuals from the previous phase
  - The phased design is meant to maximize accuracy of reported operational scores
    - In the initial period of data collection, standard scores for examinees assigned to the new forms are computed using the reference form's TCs (relies on IRT's invariance properties)
    - In the first two phases of TC estimation, data are pooled across the new forms to estimate one set of TCs that is shared by all the new forms
    - The final phase computes a separate set of TCs for each form

OPA
OFFICE OF PEOPLE ANALYTICS

CAT-ASVAB Equating: Sample Size Targets

Form	Assignment Probability	Phase 1 Target	Phase 2 Target	Phase 3 Target
Reference	1/7	500	1,500	10,000
Operational	1/7	500	1,500	10,000
New Form A	1/7	500	1,500	10,000
New Form B	1/7	500	1,500	10,000
New Form C	1/7	500	1,500	10,000
New Form D	1/7	500	1,500	10,000
New Form E	1/7	500	1,500	10,000
Total	—	**3,500**	**10,500**	**70,000**

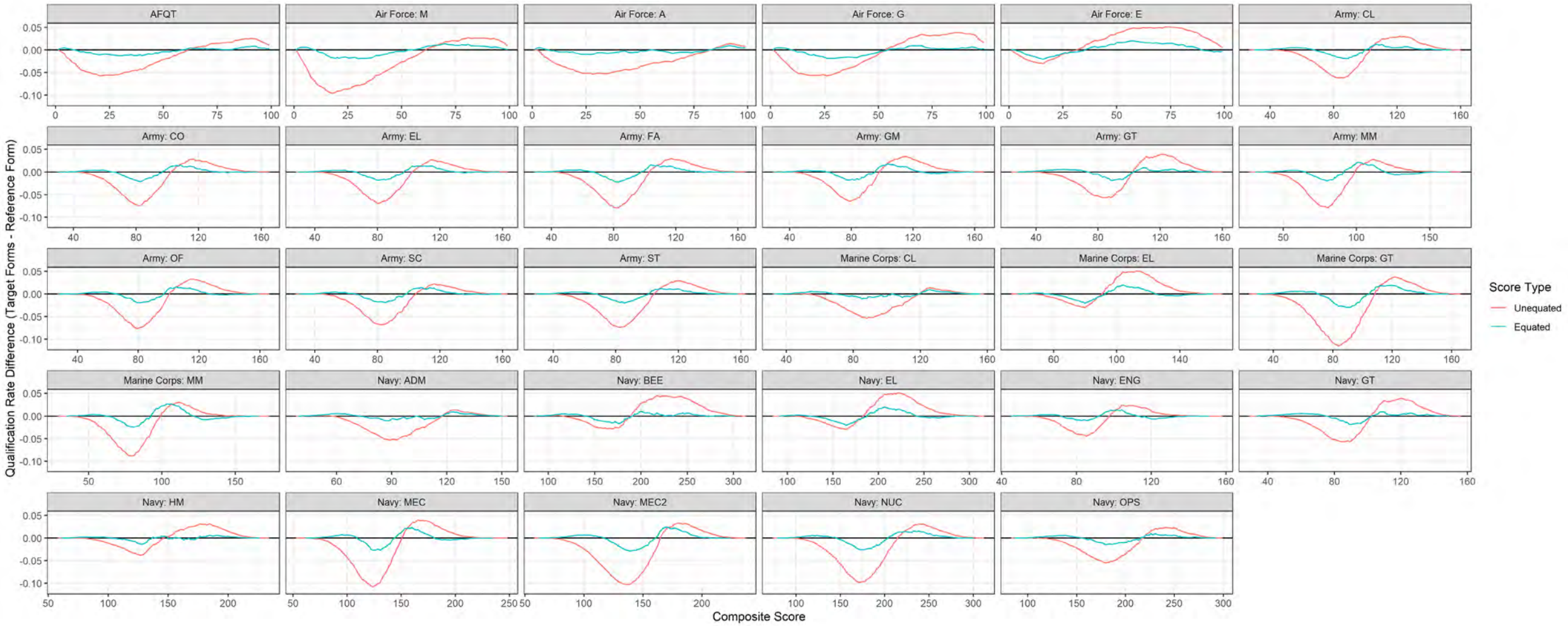
Gradual scoring refinements for new forms:

- During Phase 1, examinees' standard scores are computed using the reference form's TCs
- During Phase 2, the TCs estimated using the Phase 1 sample are put into use
  - Examinees early in this phase are scored using reference form TCs due to a delay for Phase 1 analyses and TC updates
- During Phase 3, the TCs estimated using the Phase 2 sample are put into use
  - Examinees early in this phase are scored using TCs estimated based on Phase 1 due to a delay for Phase 2 analyses and TC updates

OPA
OFFICE OF PEOPLE ANALYTICS

Note. Sample sizes across phases are cumulative. For example, the 1,500 examinees targeted for the reference form in Phase 2 include the 500 examinees targeted in Phase 1.

Unequated vs. Equated Qualification Rate Differences for CAT-ASVAB Forms 11–15 Compared to the Reference Form (from Equating Study for Forms 11–15)



OPA
OFFICE OF PEOPLE ANALYTICS

Research Questions

- Would the use of unequated standard scores from new CAT-ASVAB forms result in biased scores relative to the scores examinees would get if they took the reference form?
  - The equating briefing from the June 2024 meeting of the DACMPT already showed that equated scores are not biased (Dahlke, 2024)
- Could the sample size for an equating study be reduced from 10k per form to a smaller sample size target while achieving functionally equivalent equating results?
- Could the current equating design be updated to change the allocation of the sample across phases, the use of pooled vs. form-level equating analyses in early phases, or the use of three phases vs. two phases?

OPA
OFFICE OF PEOPLE ANALYTICS

Simulation-Based Evaluation of Unequated Scores

OPA
OFFICE OF PEOPLE ANALYTICS

Simulation Infrastructure and Scope

- Follow-up analyses requested by the DACMPT at the June 2024 meeting
  - *Would the use of unequated standard scores from new CAT-ASVAB forms result in biased scores relative to the scores examinees would get if they took the reference form?*
- Used the simulation pipeline infrastructure described in the June 2024 meeting of the DACMPT ("An Evaluation of Calibration Method and Sample Size on the Reliability of New CAT-ASVAB Forms;" Heinrich-Wallace, 2024)
  - The scores evaluated here came from the same simulation briefed by Dahlke (2024)

- Simulated 9* out of the 10 CAT-ASVAB subtests

General Science (GS)	Electronics Information (EI)
Word Knowledge (WK)	Paragraph Comprehension (PC)
Auto Information (AI)	Mechanical Comprehension (MC)
Shop Information (SI)	Arithmetic Reasoning (AR)
Math Knowledge (MK)	

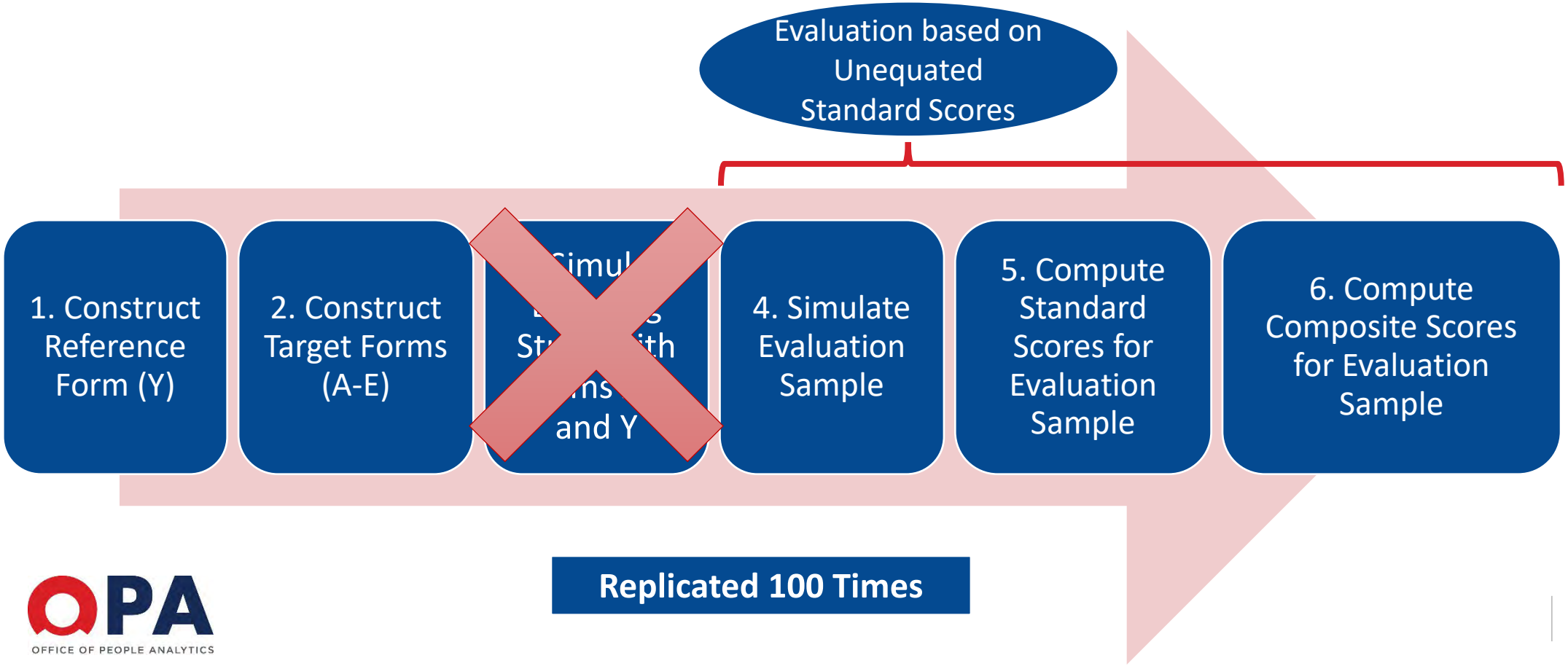
*Except Assembling Objects (AO) due to ongoing research evaluating dimensionality of AO

OPA
OFFICE OF PEOPLE ANALYTICS

Schematic Outline of Simulation Process

In June 2024, we briefed on the results of this entire process, including equating (Step 3)

Today, we will discuss the results of a reduced process when Step 3 is omitted and the reference form's TCs are used to compute all standard scores



OPA
OFFICE OF PEOPLE ANALYTICS

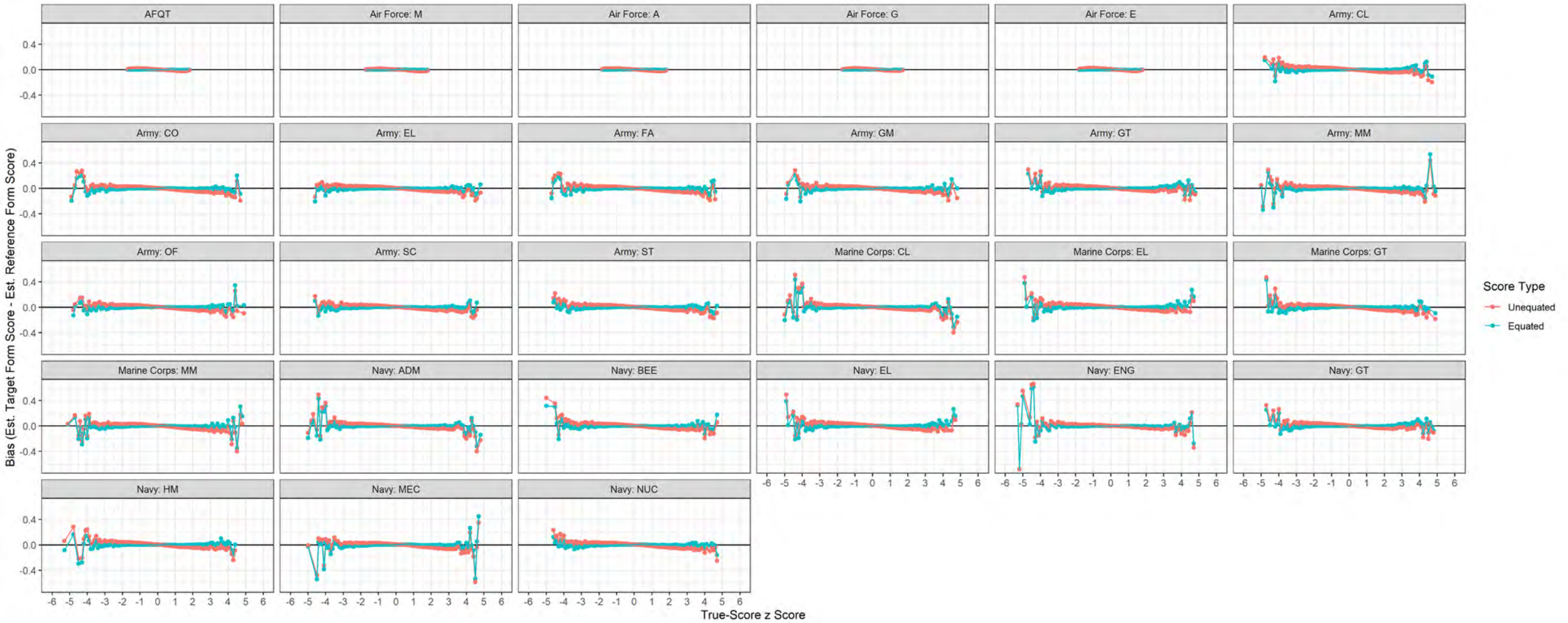
Evaluation of Conditional Score Bias

- Performed conditional bias analyses in two ways:
  - By true-score *z* scores (rounded to 1 decimal place)
    - Detailed, but estimates at the tails of the ability distribution are impacted by large amounts of sampling error
  - By true-score deciles
    - Less detailed, but allows for much more stable estimates of average bias across segments of the ability continuum due to equalized sample sizes across deciles
- Evaluated each combination of composite × form × replication × true score

$$Bias = \sum_{i=1}^{N} \frac{\left(x_{NewForm_i} - x_{ReferenceForm_i}\right)}{N}$$

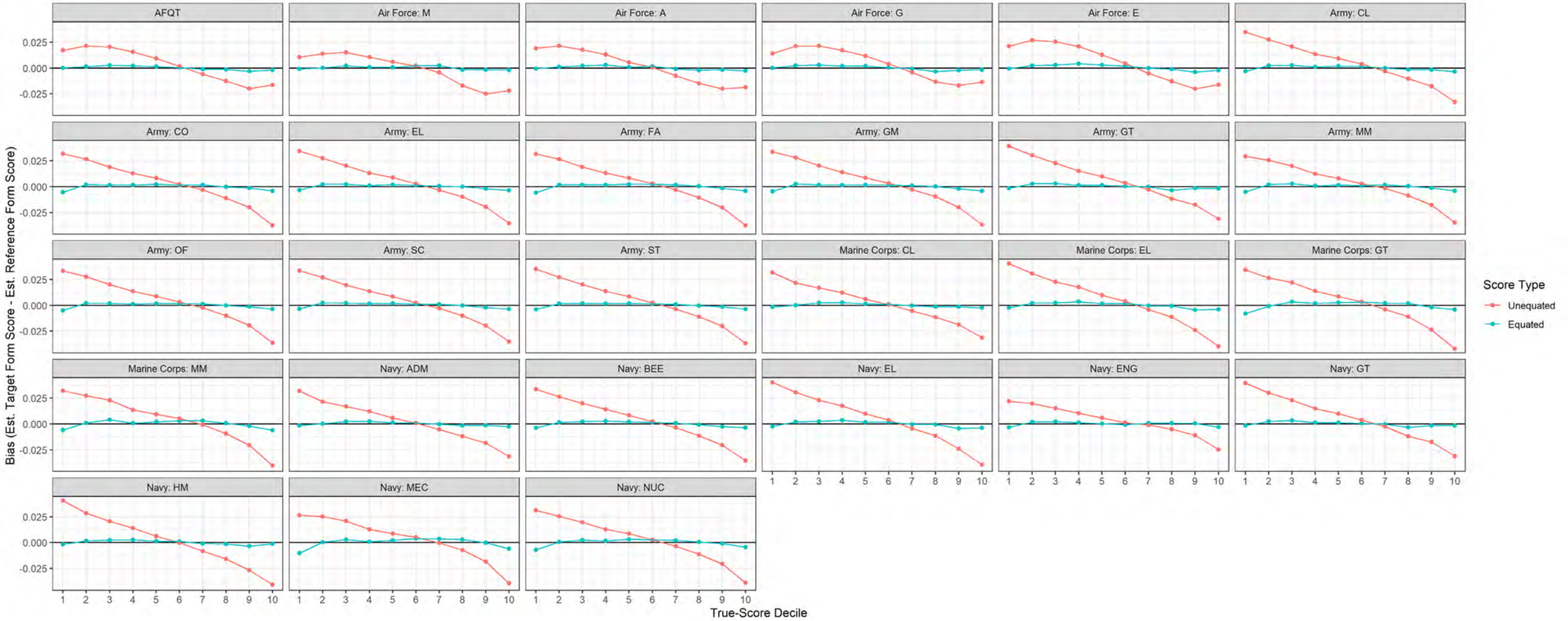
  - Scores evaluated in bias analyses were centered and scaled using the mean and SD of true scores (generating thetas converted to composite scores using generating TCs)
- The following plots depict mean bias effects across forms and replications

Evaluation of Composite Score Bias by True-Score *z* Score



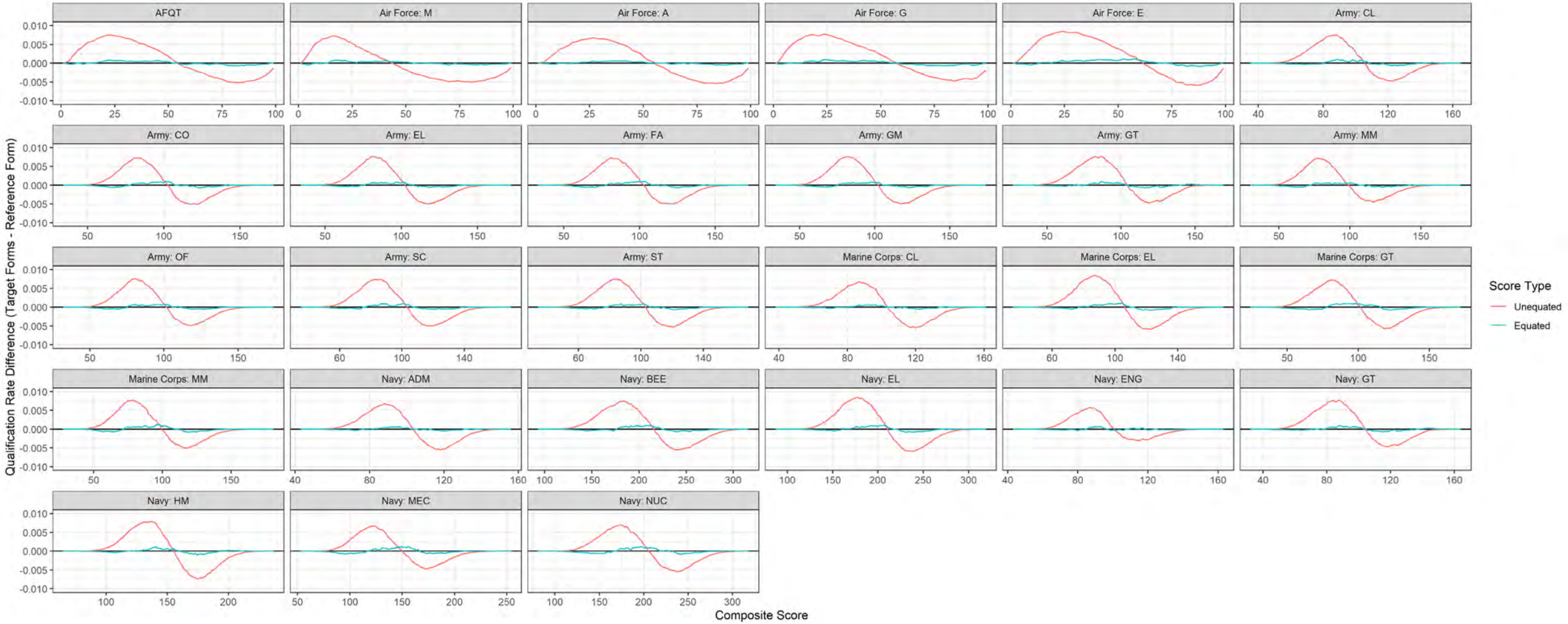
OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation of Composite Score Bias by True-Score Decile



OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation of Qualification Rate Deviations



OPA
OFFICE OF PEOPLE ANALYTICS

Conclusions from Evaluation of Simulated Scores

- Bypassing equating and computing standard scores using the reference form's TCs introduces bias into composite scores
  - In the simulation, lower scores tended to be overestimated, and higher scores tended to be underestimated
  - This bias results in qualification rate differences
- Performing equating nullifies the biases we observed in unequated scores
  - Equated scores are not biased at any point along the ability continuum
  - Equated scores produce qualification rates that are aligned with the reference form's qualification rates
- Key conclusions:
  - Consistent with the findings shared in the June 2024 briefing on equating (Dahlke 2024), equating serves its intended purpose without biasing scores
  - Equating is a remedy for biases that could occur in unequated score distributions

Equating Design Evaluation: Sample Size per Form

OPA
OFFICE OF PEOPLE ANALYTICS

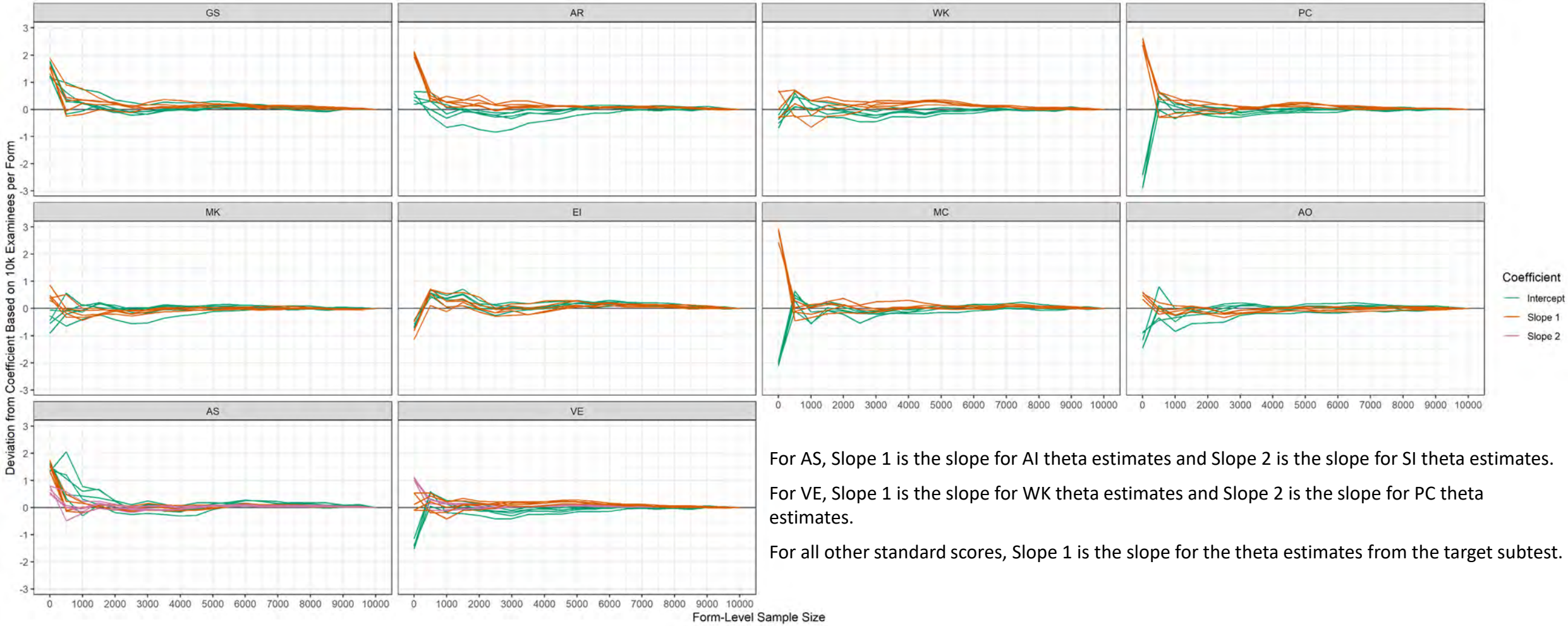
Reducing the Sample Size for CAT-ASVAB Equating Studies

- We reanalyzed data from the equating study for CAT-ASVAB Forms 11–15
- To evaluate different equating study design options, we re-ran equating analyses using varied specifications:
  - Form-level sample sizes varied from 500 to 10k in increments of 500
  - In our main set of analyses, samples were formed by selecting the first *N* records for each form in the order they were collected
  - In a corresponding set of 100 bootstrapped analyses per sample size, equating analyses were based on the first *N* records for each form in the order they appeared in each bootstrapped sample
- For each equating analysis, we estimated TCs based on form-specific equating solutions and pooled equating solutions with all five forms equated together
  - Form-specific equating solutions are the focus of our sample size evaluations
  - Pooled equating solutions were developed to support evaluations involving the number and allocation of equating phases

Convergence of Transformation Constants

OPA
OFFICE OF PEOPLE ANALYTICS

TC Convergence with N_{Form}= 10k Solution for All Coefficients



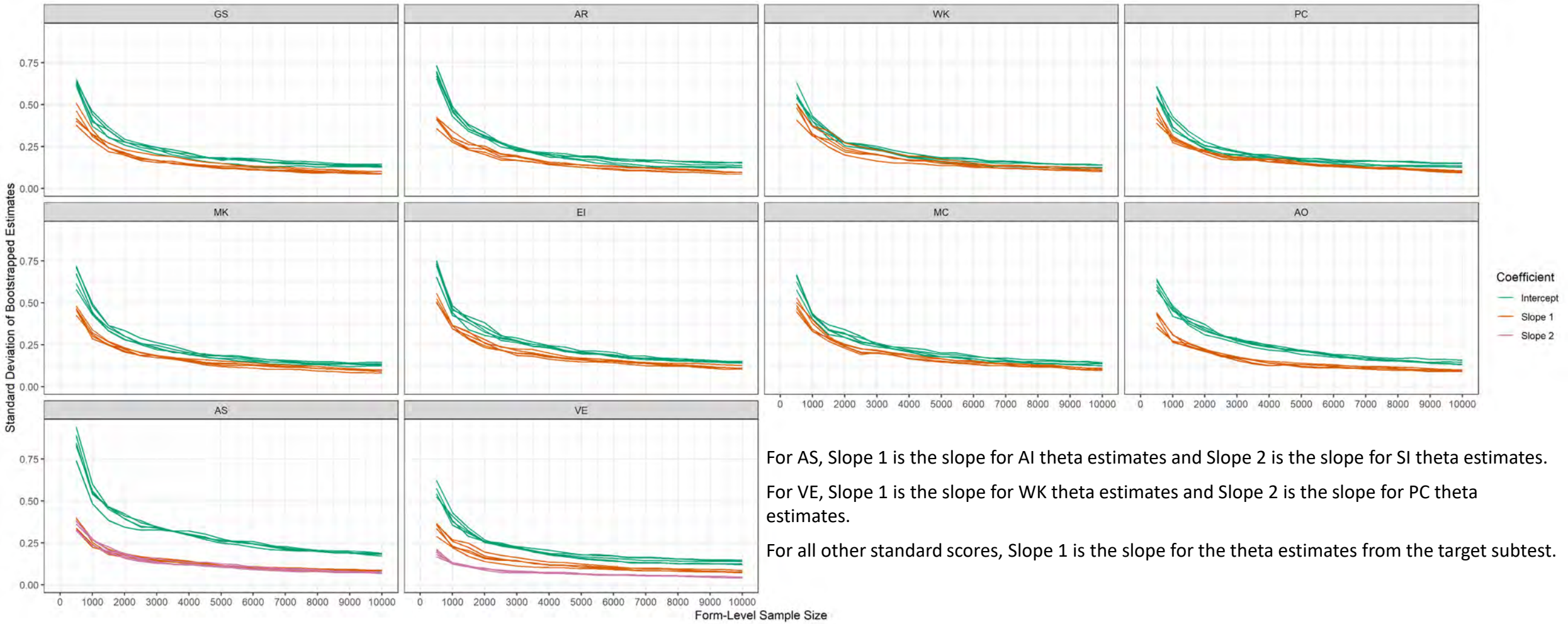
For AS, Slope 1 is the slope for AI theta estimates and Slope 2 is the slope for SI theta estimates.

For VE, Slope 1 is the slope for WK theta estimates and Slope 2 is the slope for PC theta estimates.

For all other standard scores, Slope 1 is the slope for the theta estimates from the target subtest.

OPA
OFFICE OF PEOPLE ANALYTICS

Bootstrapped Standard Errors for All Coefficients



For AS, Slope 1 is the slope for AI theta estimates and Slope 2 is the slope for SI theta estimates.

For VE, Slope 1 is the slope for WK theta estimates and Slope 2 is the slope for PC theta estimates.

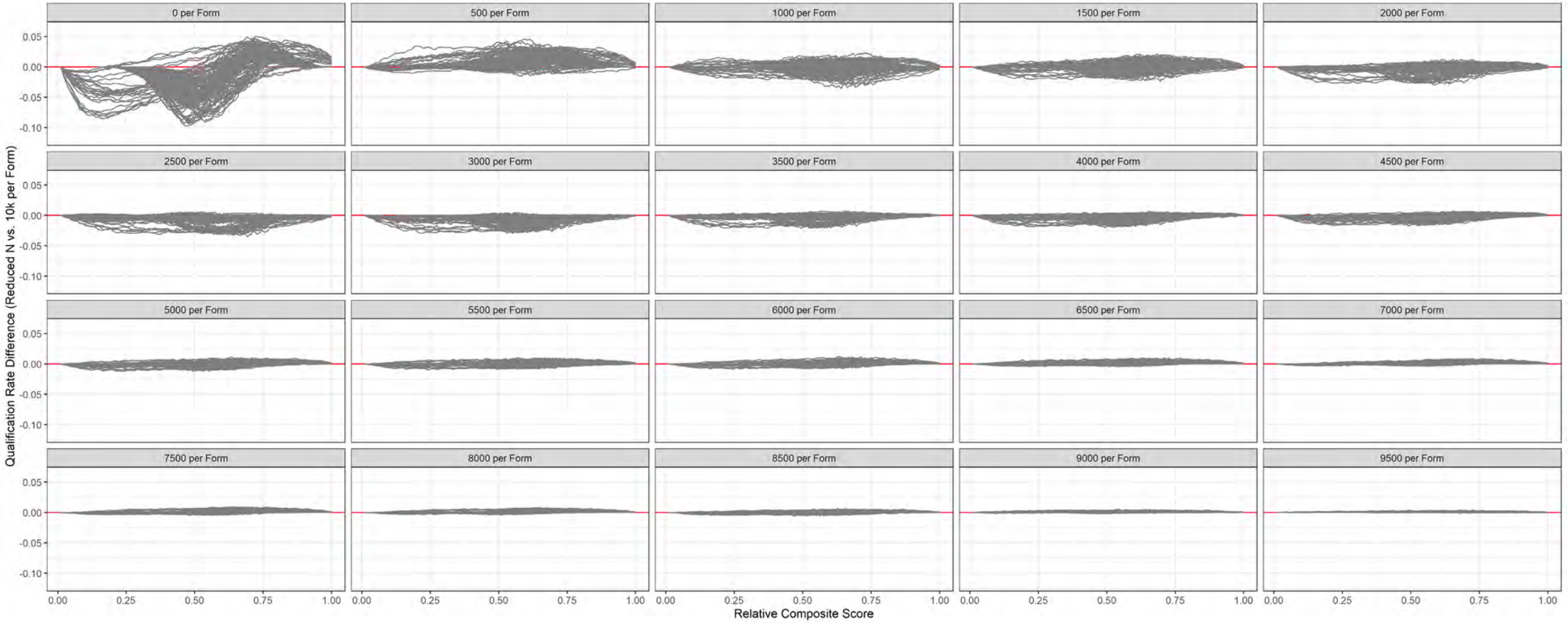
For all other standard scores, Slope 1 is the slope for the theta estimates from the target subtest.

OPA
OFFICE OF PEOPLE ANALYTICS

Qualification Rate Differences: Within-Form Convergence

OPA
OFFICE OF PEOPLE ANALYTICS

Qualification Rate Differences Relative to the *N* = 10k per Form Equating Condition Across All Composites and Forms (Equating Sample)

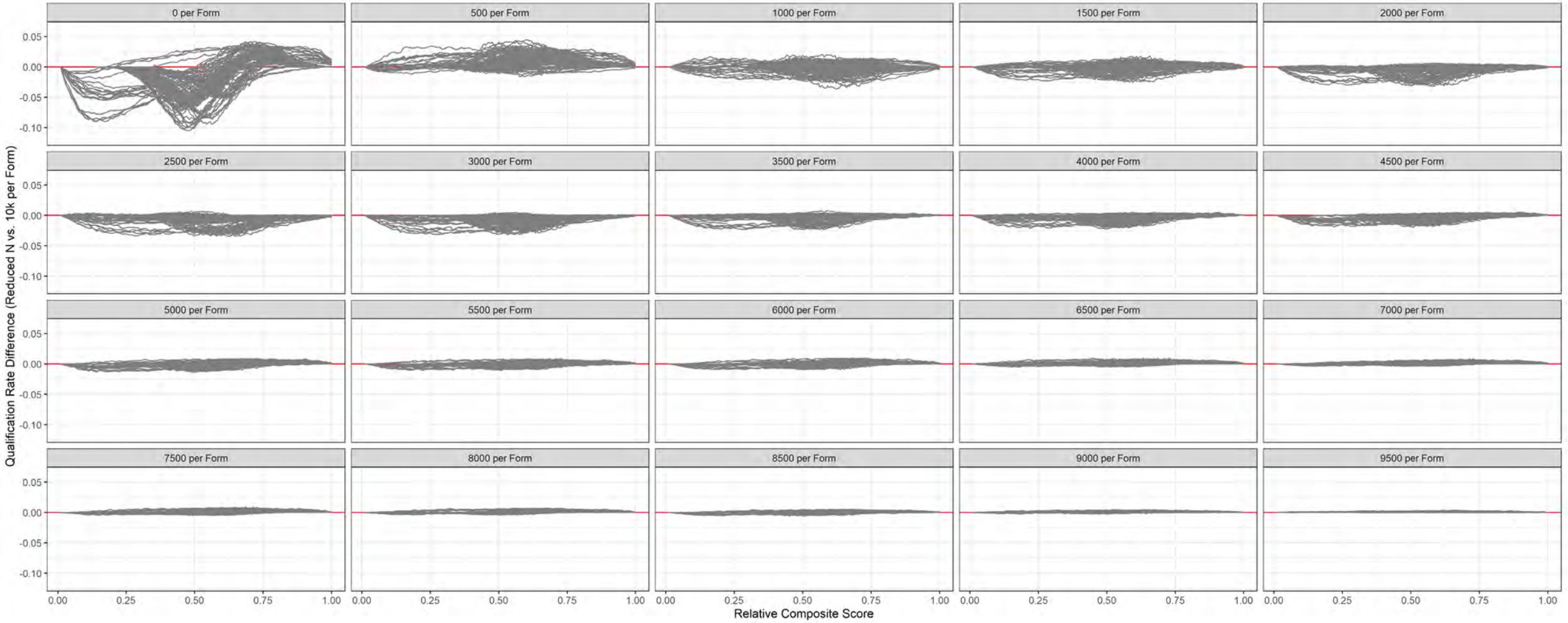


OPA
OFFICE OF PEOPLE ANALYTICS

A Holdout Sample for Evaluating Qualification Rate Convergence

- In addition to examining the convergence of qualification rates using data from the equating study, we also prepared a holdout sample
- The holdout sample consists of 10k records per form for each of the four new forms that have been administered operationally since being equated

Qualification Rate Differences Relative to the *N* = 10k Per Form Equating Condition Across All Composites and Forms (Holdout Sample)

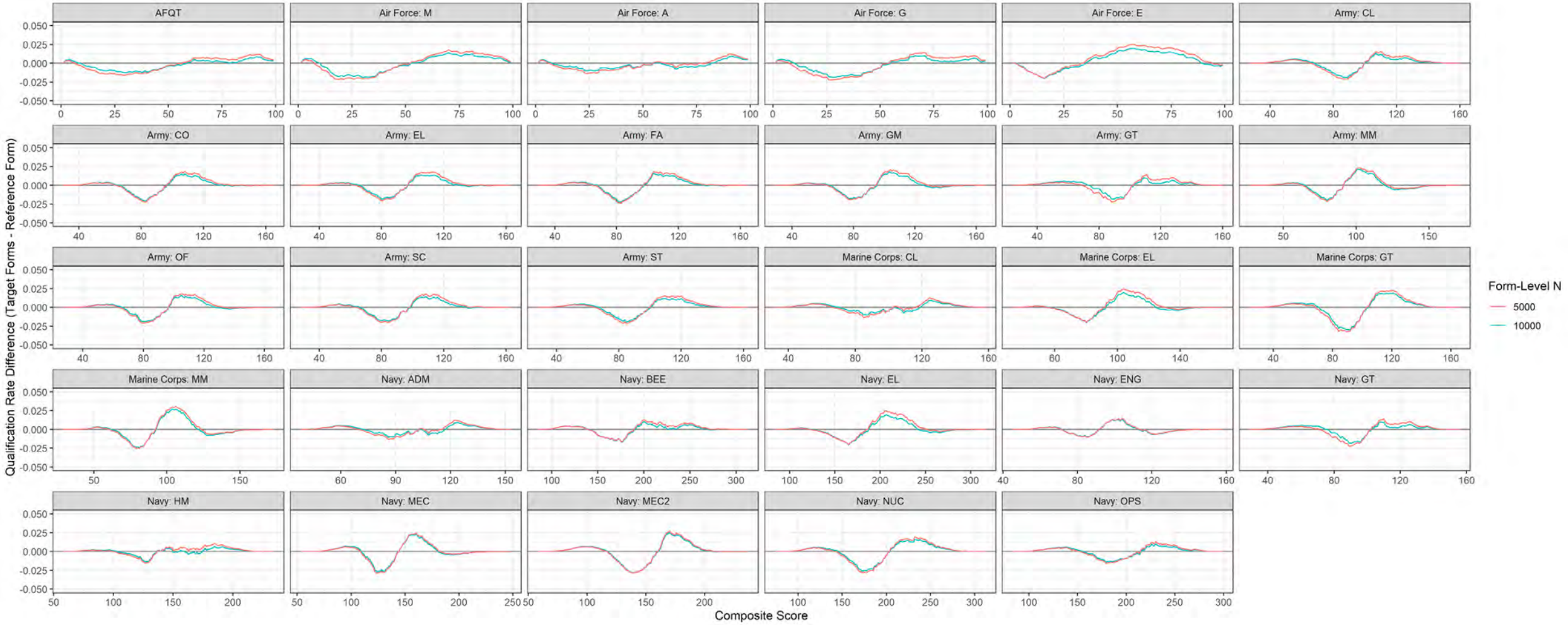


OPA
OFFICE OF PEOPLE ANALYTICS

Qualification Rate Differences: Comparison with Reference Form

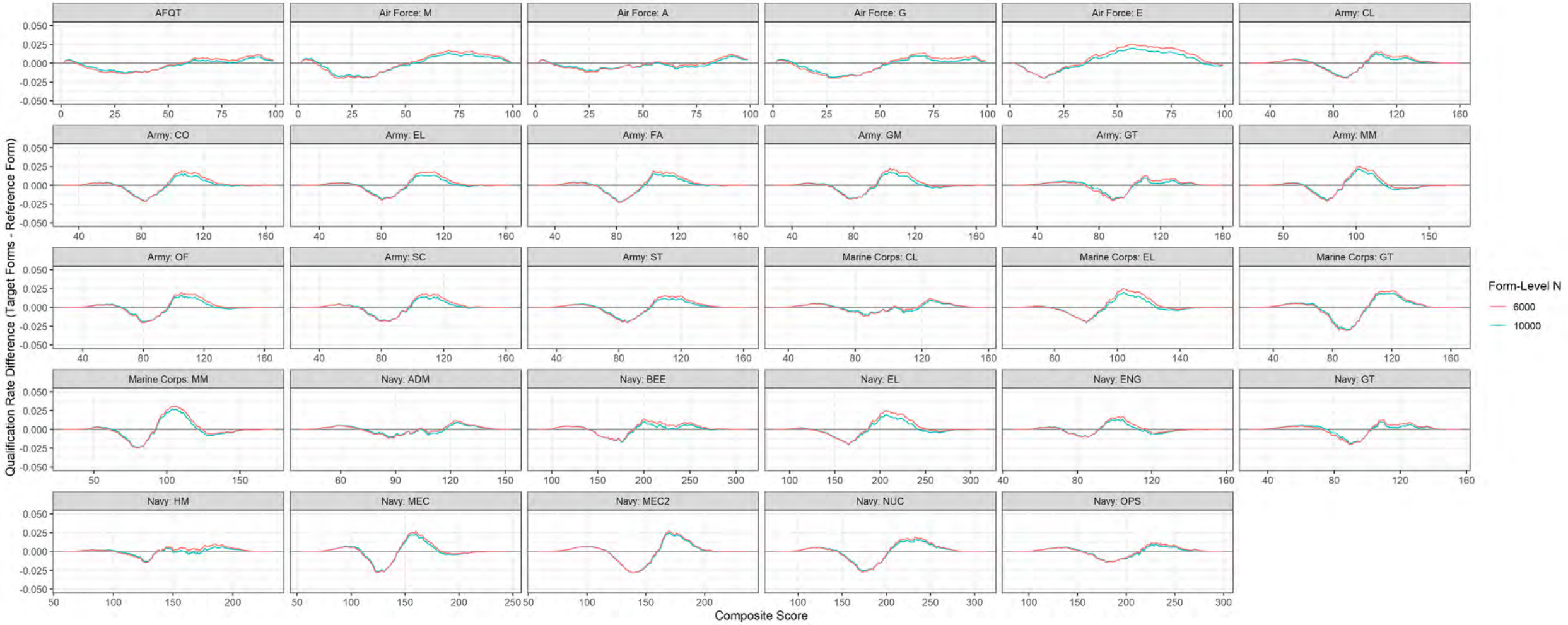
OPA
OFFICE OF PEOPLE ANALYTICS

Qualification Rate Differences Relative to the Reference Form for Equating Conditions with *N* = 5k vs. *N* = 10k per Form (Equating Sample)



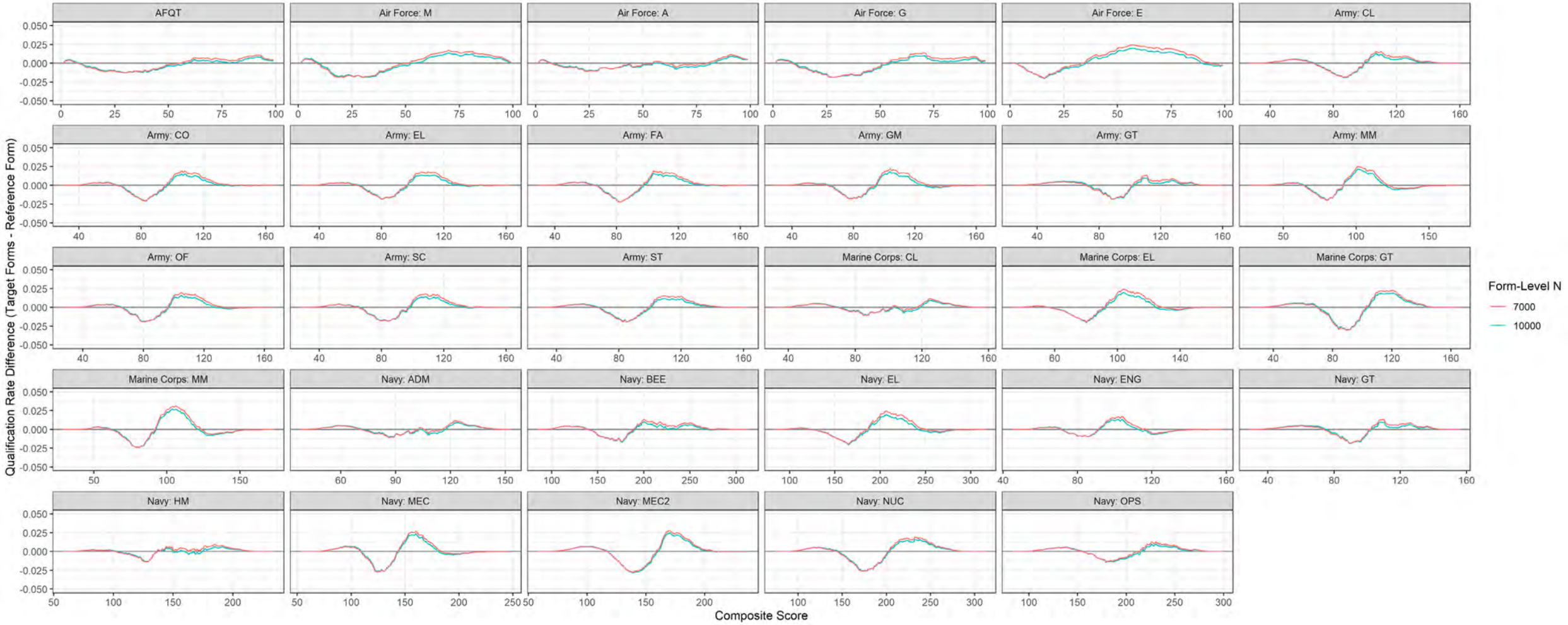
OPA
OFFICE OF PEOPLE ANALYTICS

Qualification Rate Differences Relative to the Reference Form for Equating Conditions with *N* = 6k vs. *N* = 10k per Form (Equating Sample)



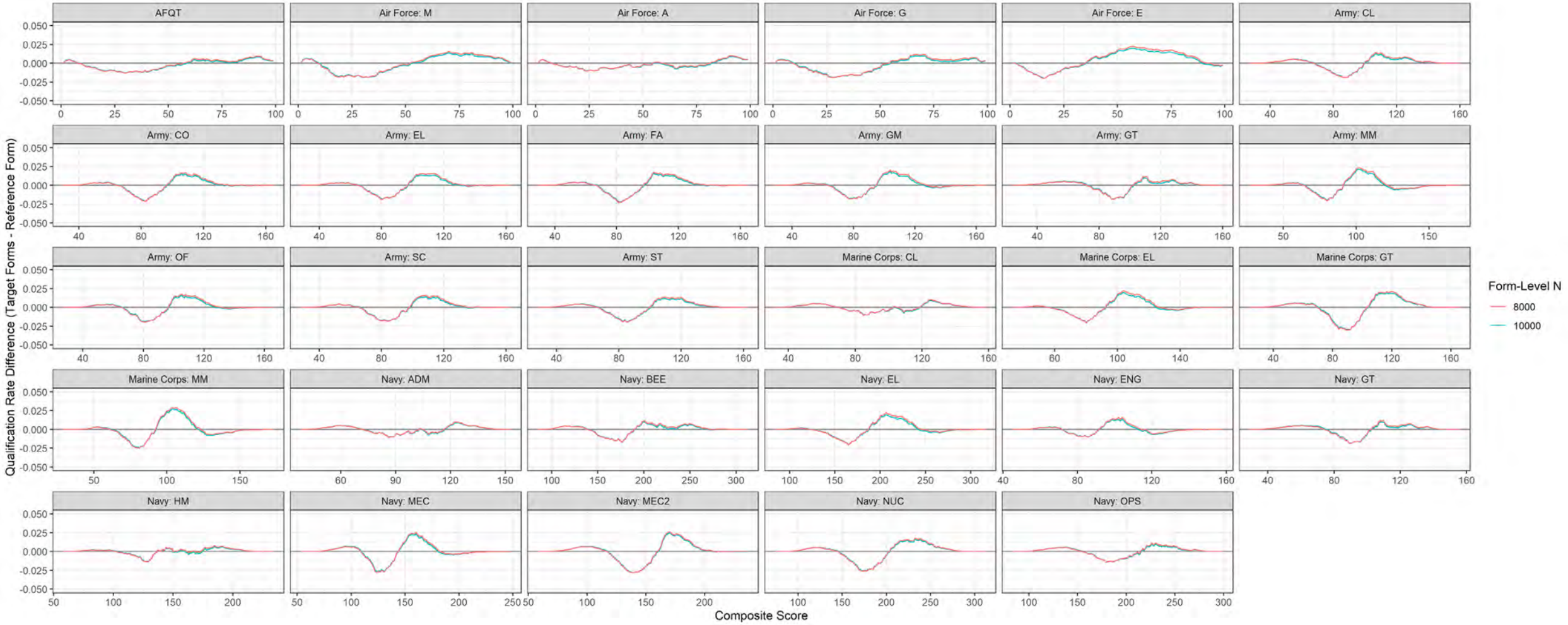
OPA
OFFICE OF PEOPLE ANALYTICS

Qualification Rate Differences Relative to the Reference Form for Equating Conditions with *N* = 7k vs. *N* = 10k per Form (Equating Sample)



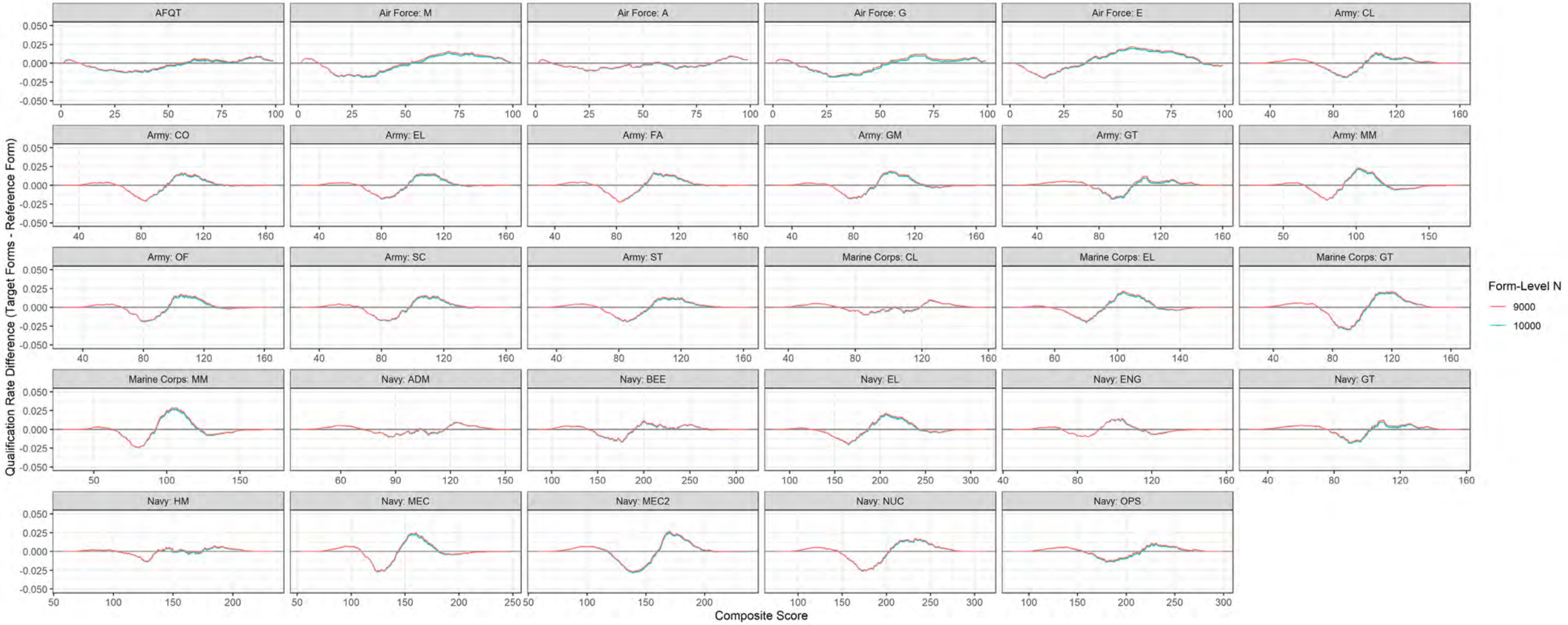
OPA
OFFICE OF PEOPLE ANALYTICS

Qualification Rate Differences Relative to the Reference Form for Equating Conditions with *N* = 8k vs. *N* = 10k per Form (Equating Sample)



OPA
OFFICE OF PEOPLE ANALYTICS

Qualification Rate Differences Relative to the Reference Form for Equating Conditions with *N* = 9k vs. *N* = 10k per Form (Equating Sample)



OPA
OFFICE OF PEOPLE ANALYTICS

Sample Size Recommendation for Future Equating Studies

- Based on our evaluations of TC convergence and qualification rate differences, a target sample size of **6k examinees per form** appears sufficient to achieve functional convergence with analyses based on 10k examinees per form
- Solutions based on as few as 5k examinees per form were quite stable, but using 6k per form allowed the solutions to stabilize even more
  - Compared to 5k, a sample of 6k per form helped TCs to reach closer alignment with the 10k solution (including resolving residuals for forms that were outliers with smaller sample sizes)
  - Compared to 5k, using a sample of 6k per form noticeably improved qualification rate convergence with the reference form for the AFQT

Equating Design Evaluation: Impact of Changing the Number or Allocation of Equating Phases

OPA
OFFICE OF PEOPLE ANALYTICS

Goals for Changing the Number or Allocation of Equating Phases

- Having identified a recommended form-level target sample size for forms' final equating analyses, we next evaluated how other aspects of the equating study design might be altered to:
  - Streamline the administration of the study
  - Reduce differences between scores recorded for examinees who test during an equating study and the scores they would have received if the final equated TCs could be used to recompute their standard scores
- The design factors considered in these evaluations have no additional impact on the final TCs estimated for each form beyond our reduction of the total form-level sample size

OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation Strategy

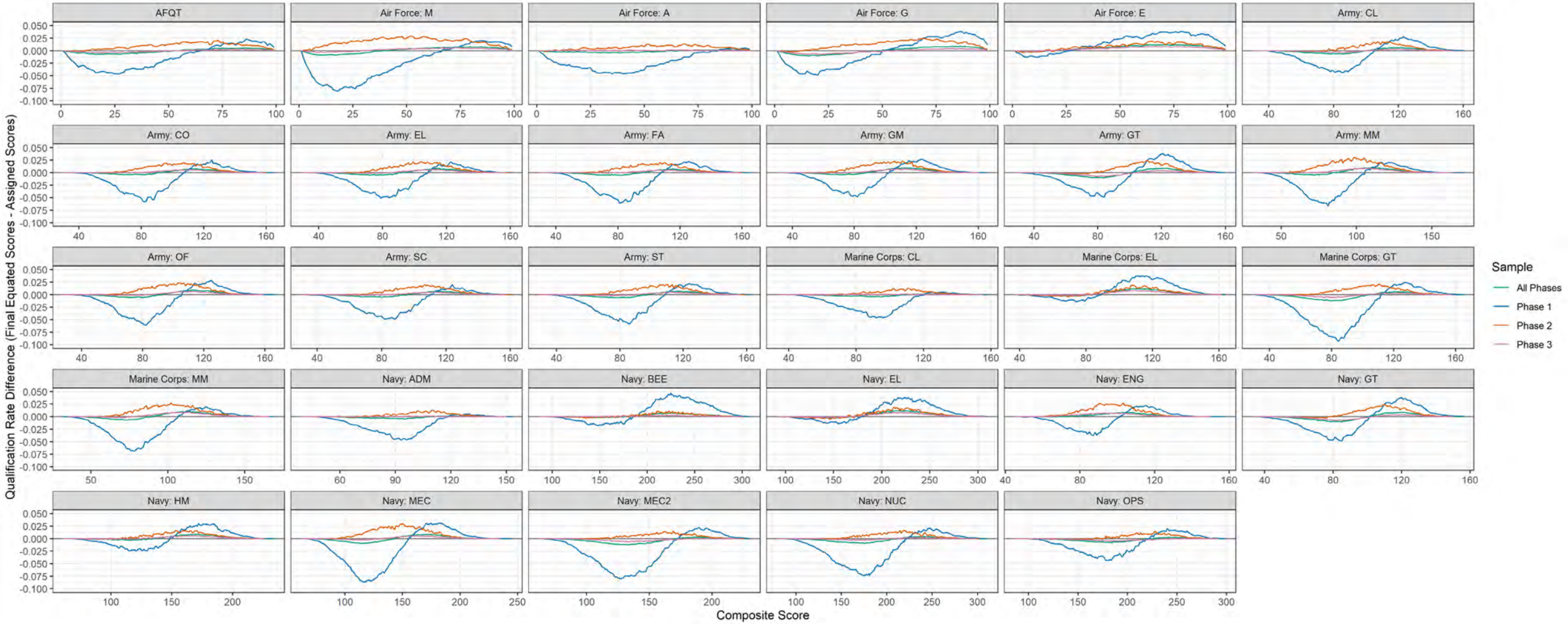
- Each sample was constructed by selecting examinees from the equating data set from CAT-ASVAB Forms 11–15 in the order their results were recorded
- We used a series of four sequential evaluations to identify a recommended configuration for future equating studies:
  1. Using a final form-level sample size of 10k vs. 6k (rehash of sample size evaluation)
  2. Using pooled equating vs. form-specific equating in early phases
  3. Using existing early-phase sample sizes vs. increasing them
  4. Using a three-phase design vs. a two-phase design
- The recommended design feature from each evaluation was carried forward in subsequent evaluations
- The primary basis for making these evaluations is their impact on the qualification rate differences (and the SDs of differences across forms) between:
  a) the equated scores examinees would have earned if the final TCs could be applied retroactively and
  b) the operational scores examinees would have earned at the time they tested, as determined using the TCs specified by the design features in our evaluation

Accounting for the Processing Lag Between Equating Phases

- To enhance the realism of these evaluations, we included a form-level sample size lag of 500 examinees between equating phases
- This accounts for the additional testing that occurs while temporary equating solutions are being computed, replicated, implemented, and released
  - E.g., although the current Phase 1 *N* is 500 per form, the processing lag in our analyses means 500 additional people take each form before the provisional TCs can be replaced with temporary, equated TCs
  - The additional testing volume that accumulates while the TCs are being updated represents an additional group of people who are not benefitting from the gradual updates we make to the TCs during the study period

OPA
OFFICE OF PEOPLE ANALYTICS

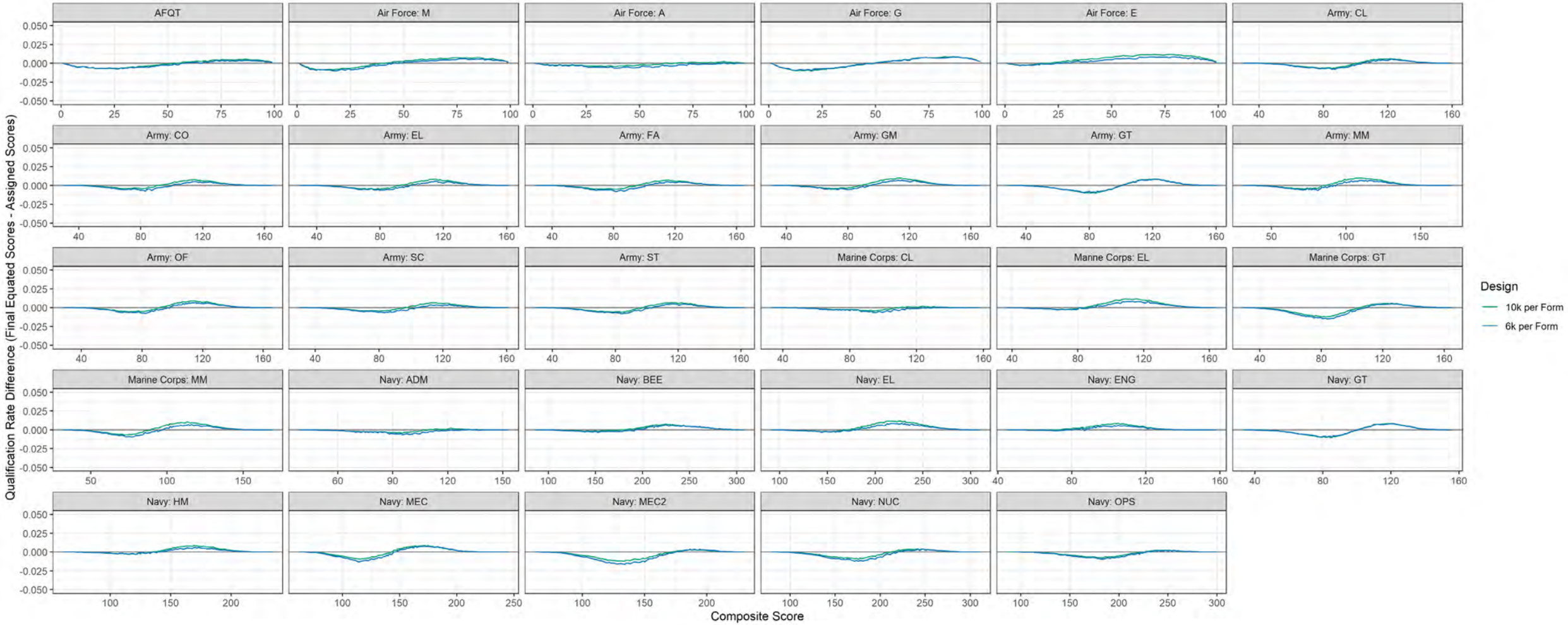
Current Design: Qualification Rate Differences for Reported Scores Compared to Scores Based on Final Equating Constants



OPA
OFFICE OF PEOPLE ANALYTICS

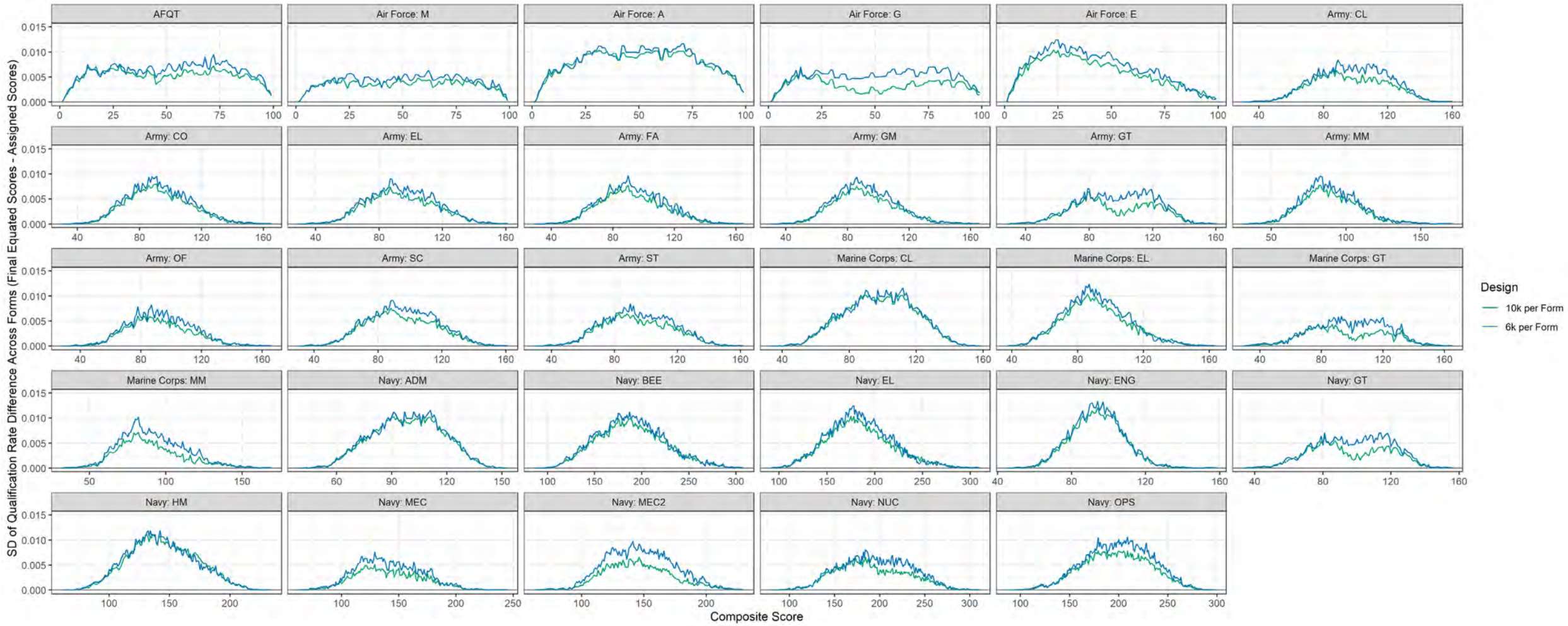
Note: Phase-specific samples are non-cumulative in this figure.

Evaluation 1: Using Final Form-Level *N* of 10k vs. 6k (Overall Qualification Rate Differences Across Forms)



OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation 1: Using Final Form-Level *N* of 10k vs. 6k (Standard Deviations of Qualification Rate Differences Across Forms)



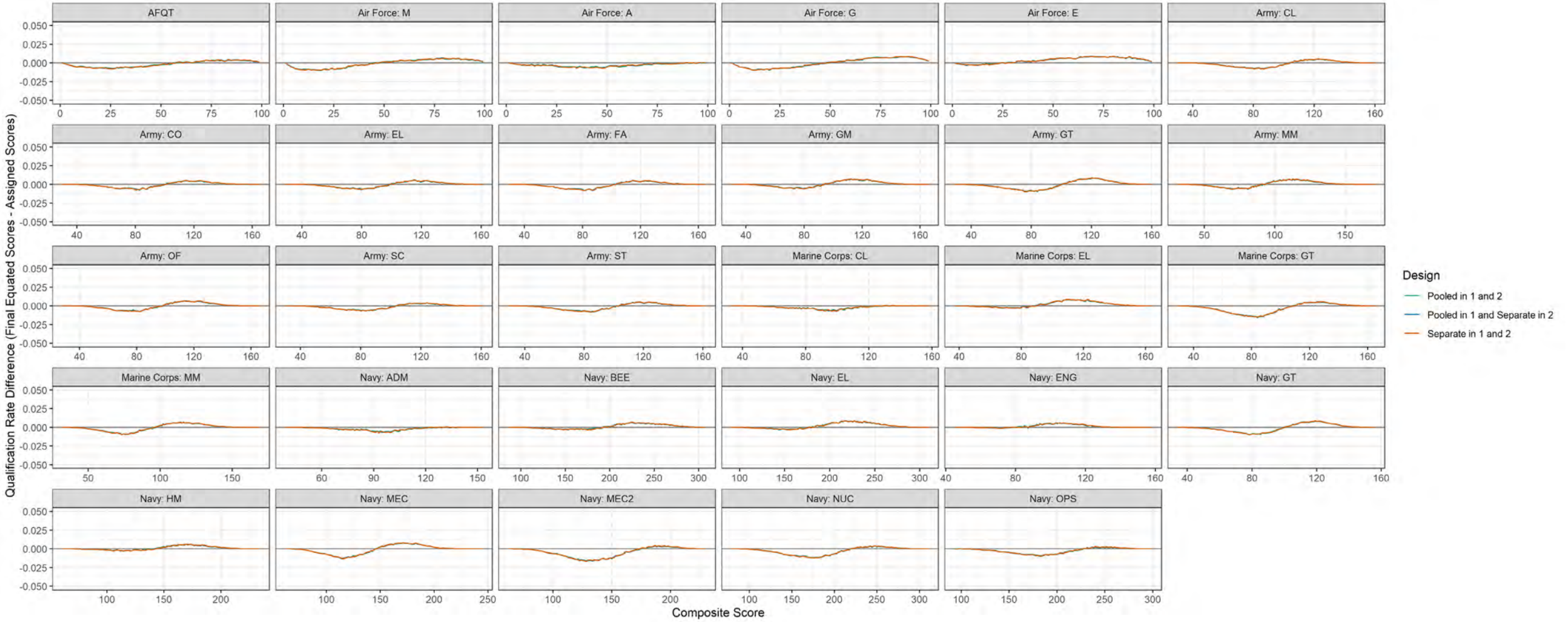
OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation 1 Winner: 6k Examinees per Form

- Allows a substantial reduction in the duration of an equating study
- Has minimal impact on the overall quality of examinees' scores

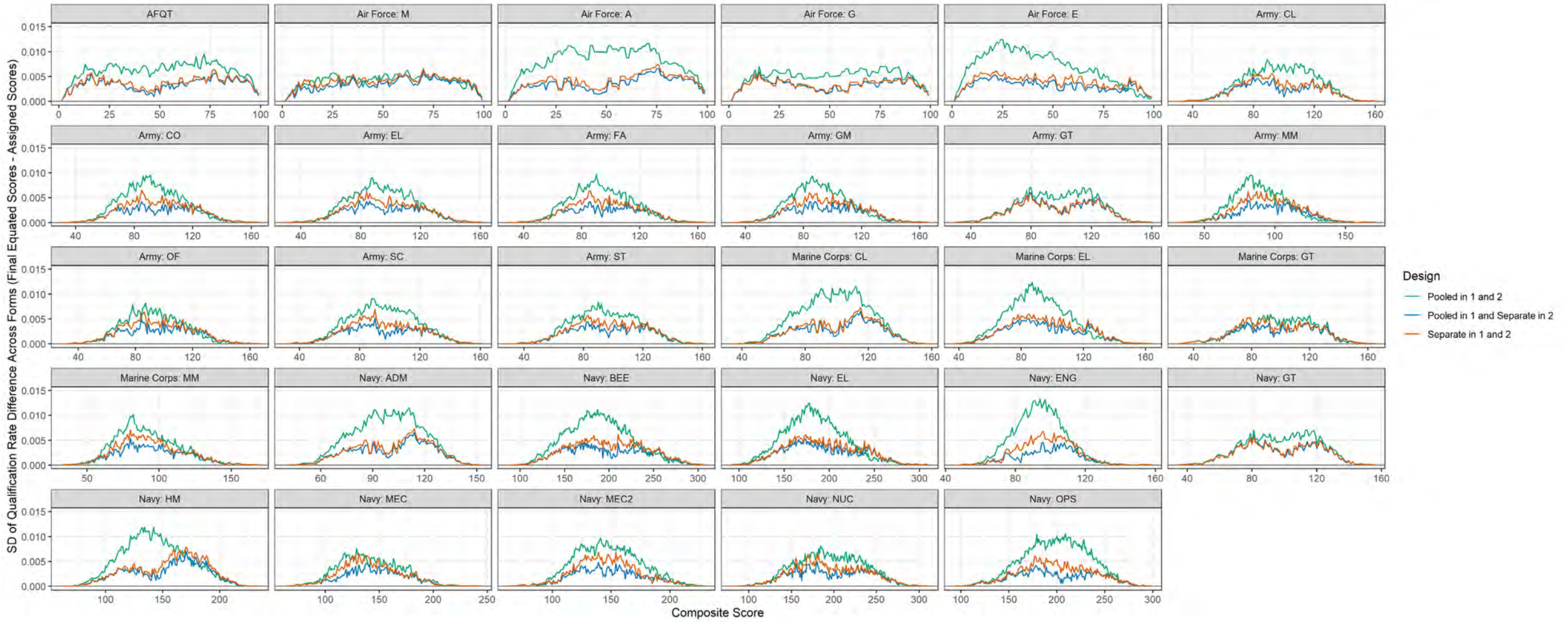
OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation 2: Using Pooled vs. Separate Equating in Early Phases (Overall Qualification Rate Differences Across Forms)



OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation 2: Using Pooled vs. Separate Equating in Early Phases (Standard Deviations of Qualification Rate Differences Across Forms)



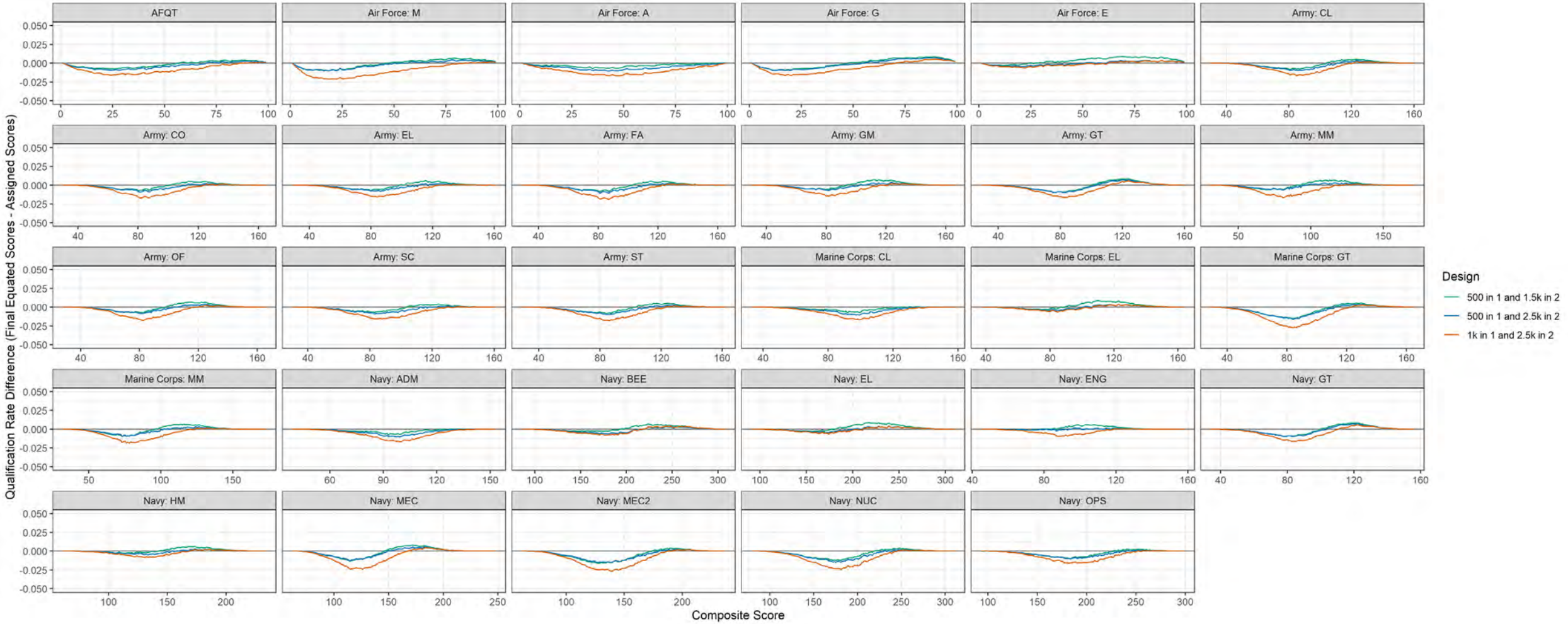
OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation 2 Winner: Pooled Equating in Phase 1 with Separate Equating in Phase 2

- Using form-specific equating analyses in Phase 2 improves the overall quality of reported scores by reducing the variability in quality across forms during Phase 3

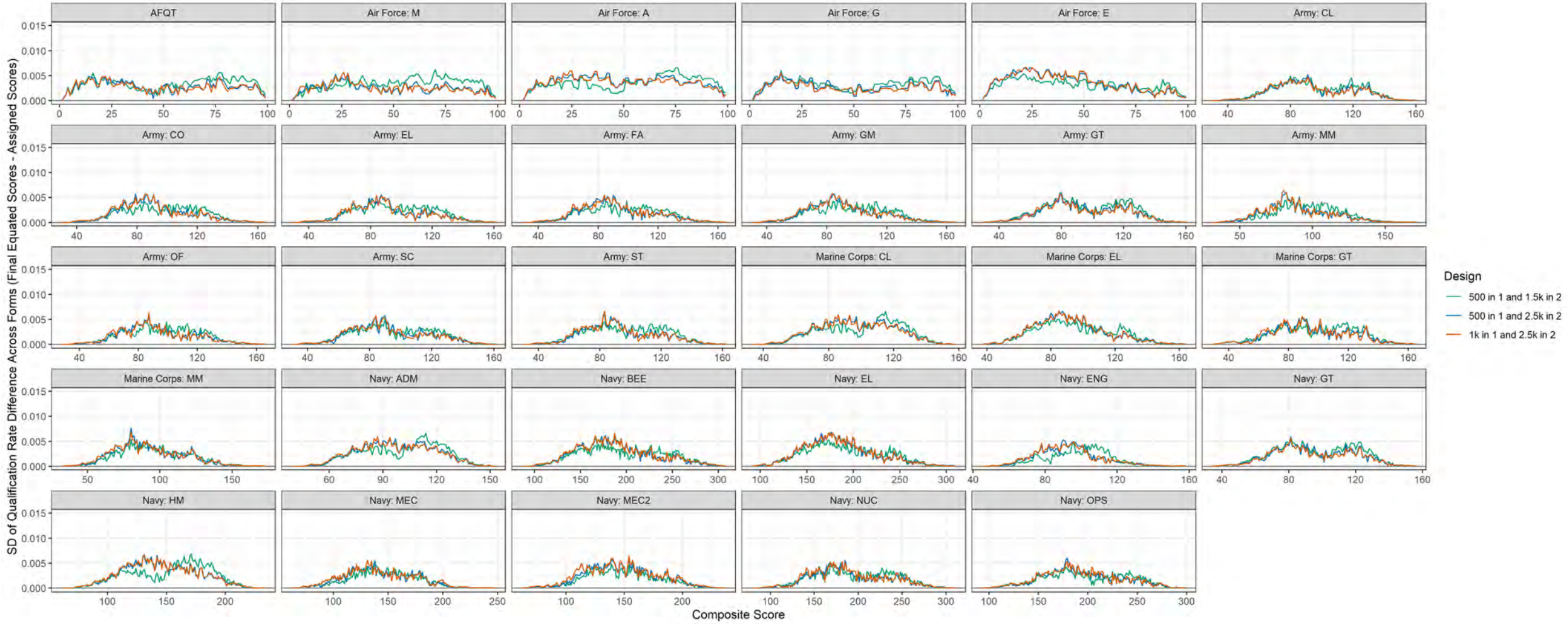
OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation 3: Using Existing Early-Phase *Ns* vs. Increased *Ns* (Overall Qualification Rate Differences Across Forms)



OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation 3: Using Existing Early-Phase *Ns* vs. Increased *Ns* (Standard Deviations of Qualification Rate Differences Across Forms)



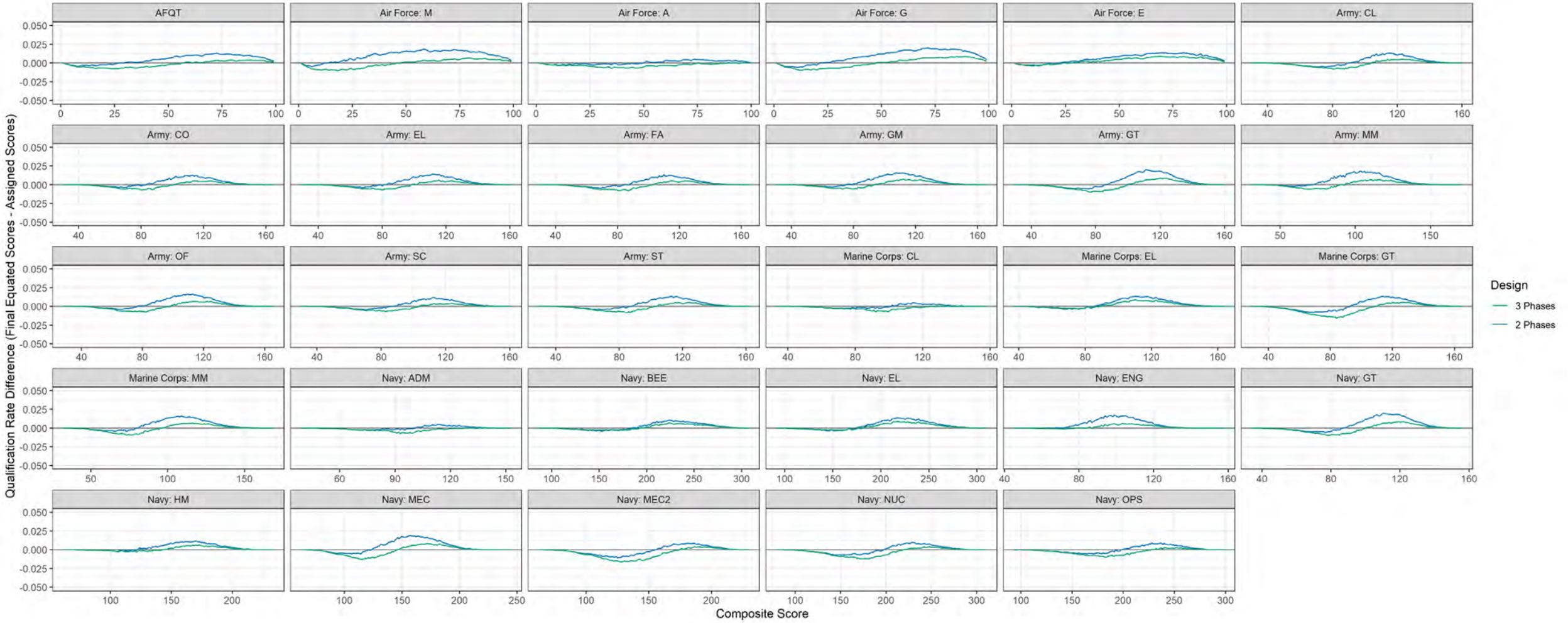
OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation 3 Winner: *N* = 500 per Form in Phase 1 and *N* = 1,500 per Form in Phase 2

- No change
- These sample size targets are effective at mitigating the impact of provisional TCs on the quality of reported scores

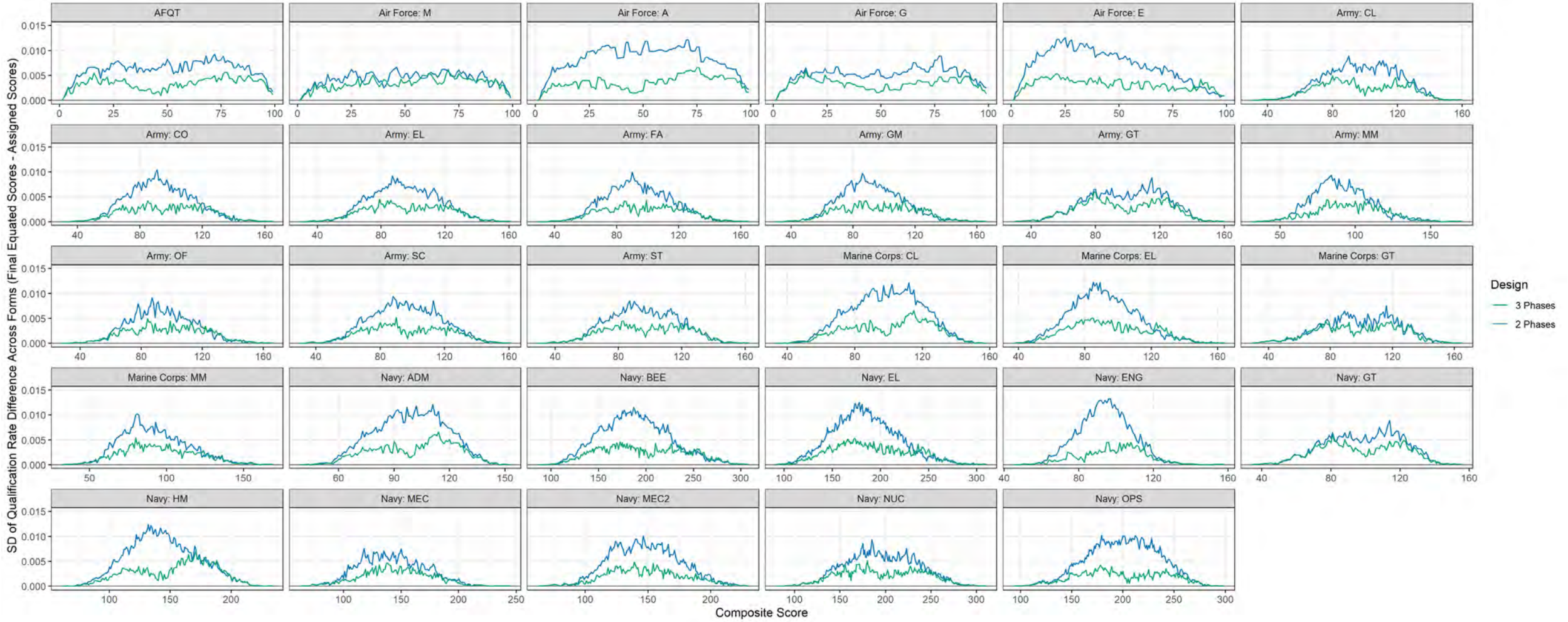
OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation 4: Using a Three-Phase Design vs. a Two-Phase Design (Overall Qualification Rate Differences Across Forms)



OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation 4: Using a Three-Phase Design vs. a Two-Phase Design (Standard Deviations of Qualification Rate Differences Across Forms)



OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation 4 Winner: Three-Phase Equating Design

- No change
- A three-phase design is superior to a two-phase design because it allows an additional opportunity to refine the temporary TCs, which improves the quality of reported scores

OPA
OFFICE OF PEOPLE ANALYTICS

Summary of Recommended Alterations to the CAT-ASVAB Equating Design

OPA
OFFICE OF PEOPLE ANALYTICS

Summary of Recommended Alterations to the CAT-ASVAB Equating Design

- We recommend that future CAT-ASVAB equating studies continue using a three-phase design with the following specifications (changes **bolded**):
  - Phase 1: Target *N* = 500 per form; estimate temporary TCs using a pooled equating analysis across forms
  - Phase 2: Target *N* = 1,500 per form; estimate temporary TCs using a **separate equating analysis per form**
  - Phase 3: Target *N* = **6,000 per form**; estimate final TCs using a separate equating analysis per form
- This design will reduce the duration and number of examinees involved in equating studies, while converging well with the results of a 10k-per-form equating solution and improving the quality of scores reported during Phase 3

Questions for the DAC

OPA
OFFICE OF PEOPLE ANALYTICS

Question for the DAC

- Does the DAC concur with the recommended design changes for future CAT-ASVAB equating studies? (changes **bolded**)
  - Phase 1: 500 per form (pooled equating)
  - Phase 2: 1,500 per form (cumulative *N*; **separate equating for each form**)
  - Phase 3: **6,000 per form** (cumulative *N*; separate equating for each form)

OPA
OFFICE OF PEOPLE ANALYTICS

Thank You!

For more information,
please contact:

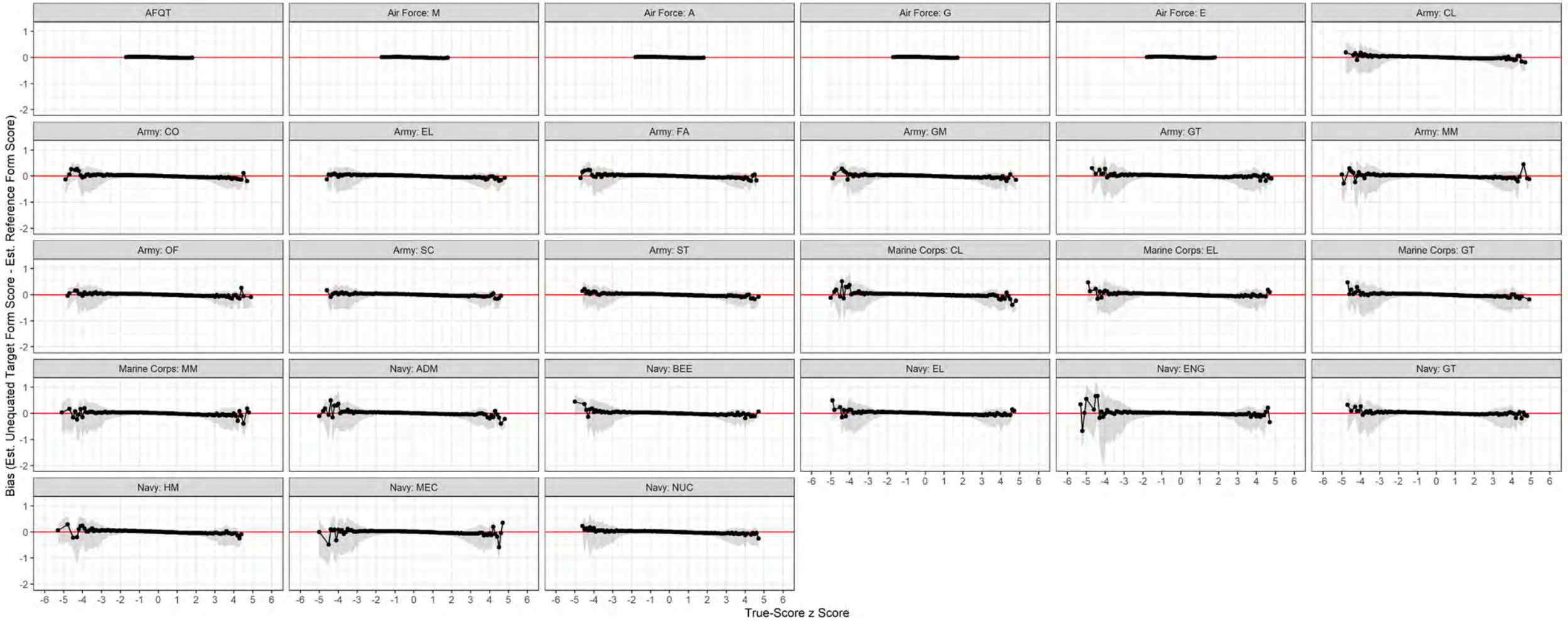
Jeff Dahlke
jdahlke@humrro.org
jeffrey.a.dahlke.ctr@mail.mil

OPA
OFFICE OF PEOPLE ANALYTICS

Supplemental Slides: Simulation-Based Evaluation of Unequated Scores

OPA
OFFICE OF PEOPLE ANALYTICS

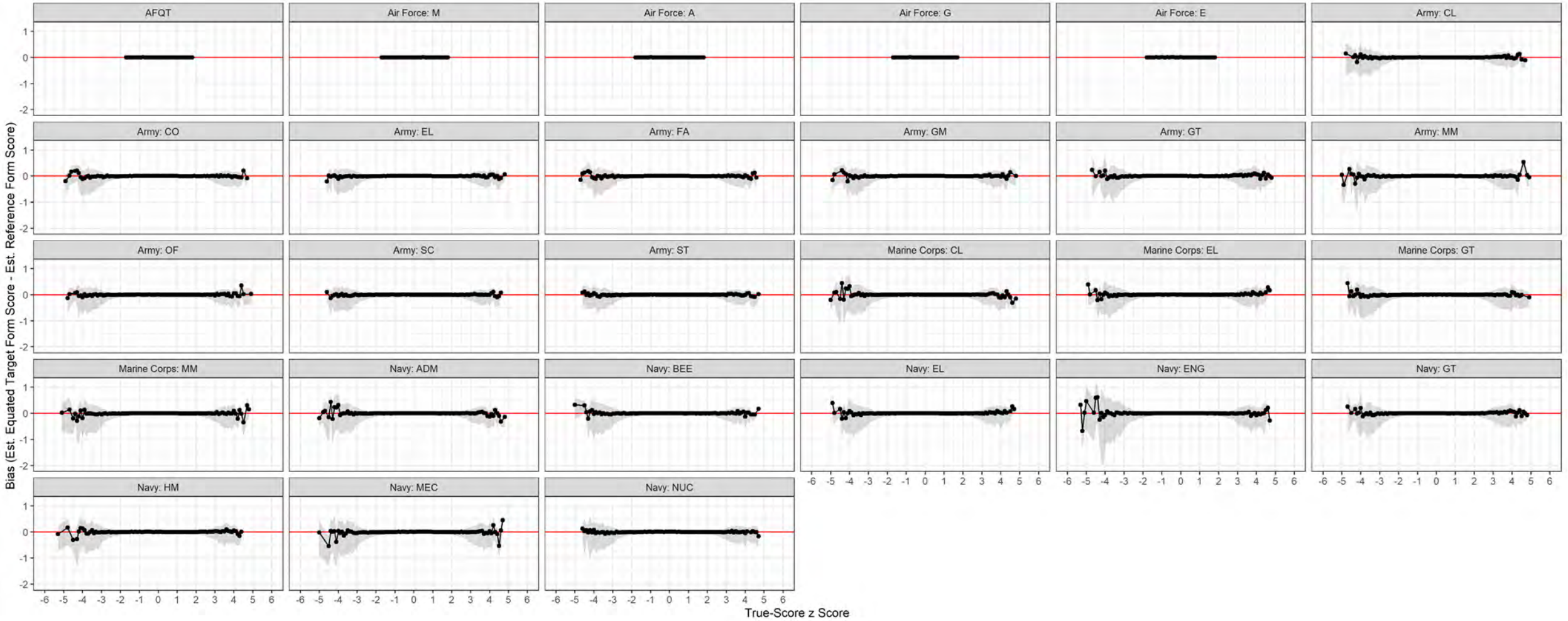
Evaluation of Unequated Composite Score Bias by True-Score *z* Score



OPA
OFFICE OF PEOPLE ANALYTICS

Note. Error ribbons represent 95% confidence intervals.

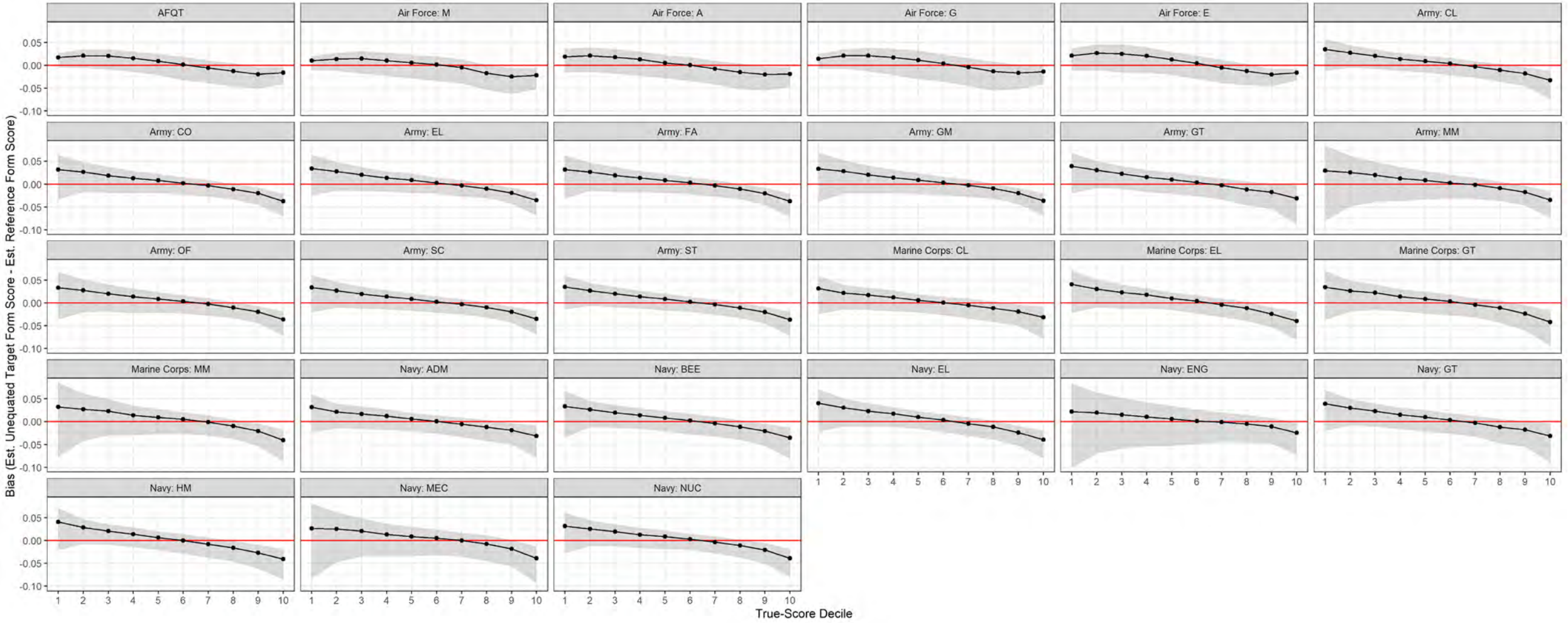
Evaluation of Equated Composite Score Bias by True-Score *z* Score



OPA
OFFICE OF PEOPLE ANALYTICS

Note. Error ribbons represent 95% confidence intervals.

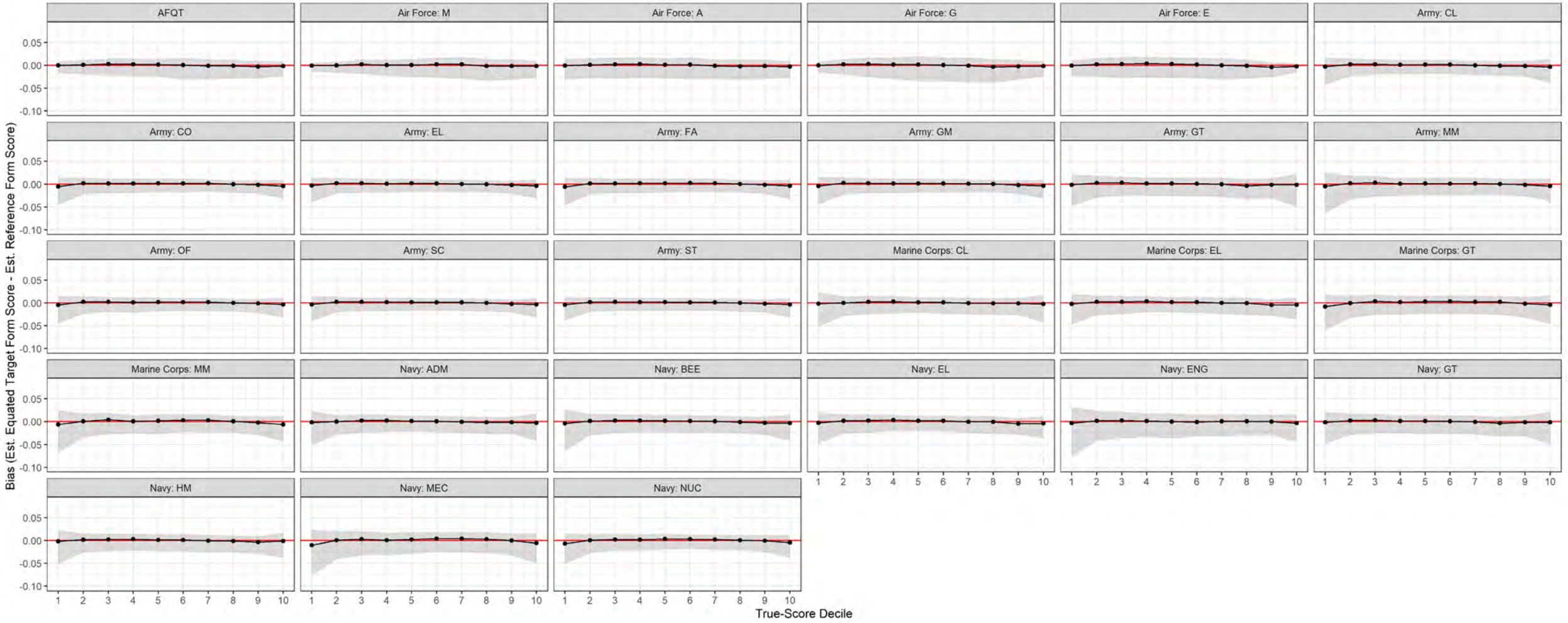
Evaluation of Unequated Composite Score Bias by True-Score Decile



OPA
OFFICE OF PEOPLE ANALYTICS

Note. Error ribbons represent 95% confidence intervals.

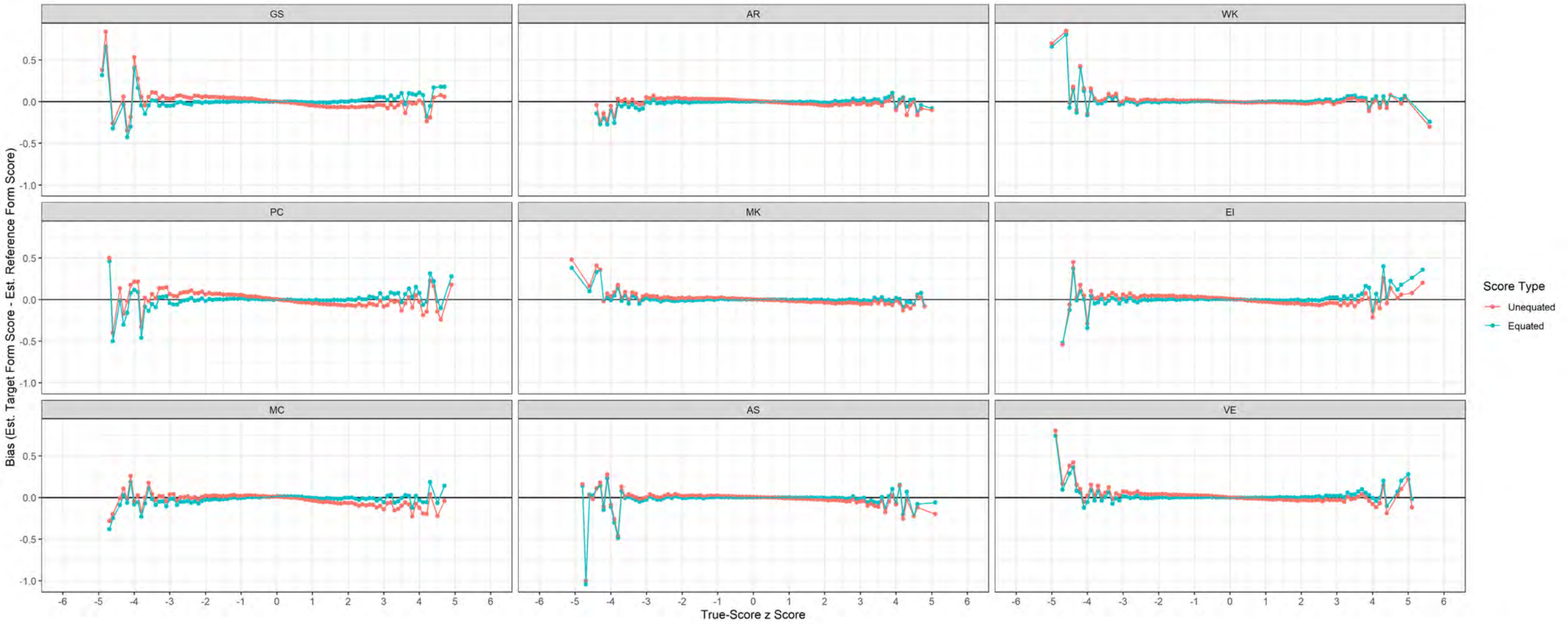
Evaluation of Equated Composite Score Bias by True-Score Decile



OPA
OFFICE OF PEOPLE ANALYTICS

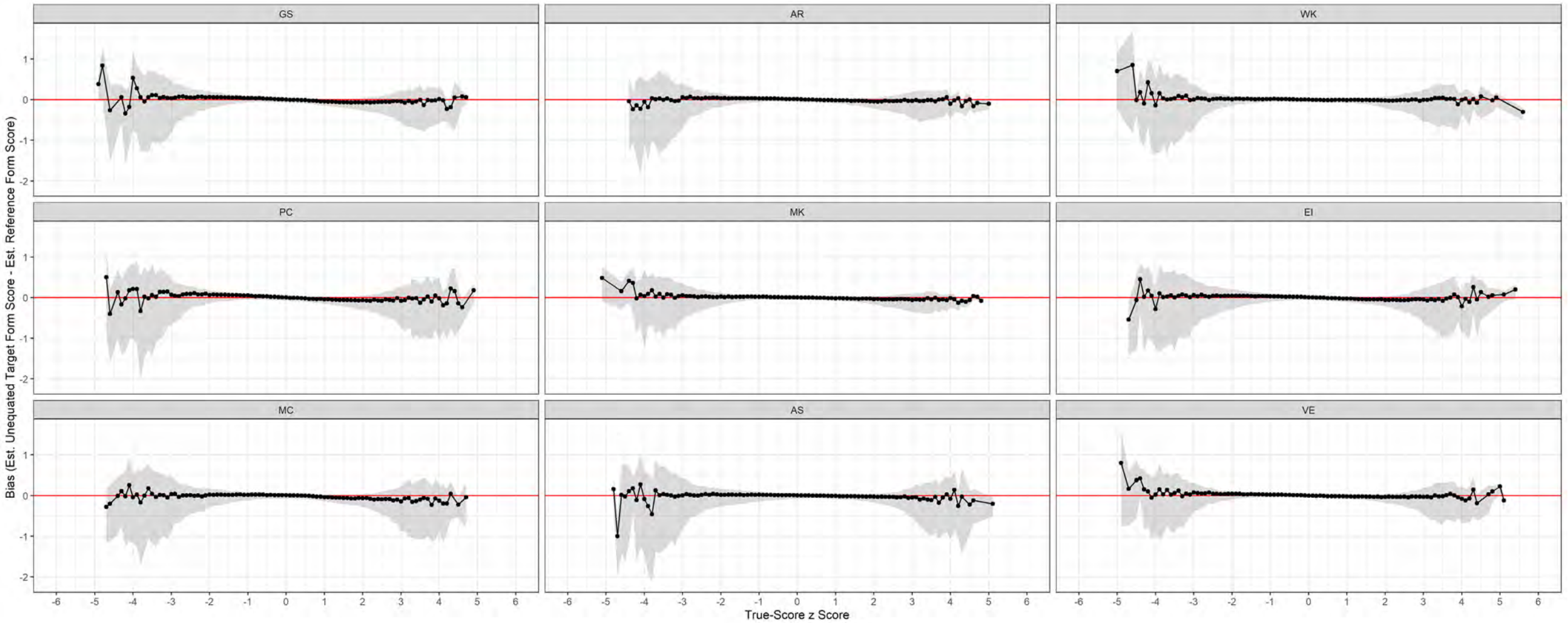
Note. Error ribbons represent 95% confidence intervals.

Evaluation of Standard Score Bias by True-Score *z* Score



OPA
OFFICE OF PEOPLE ANALYTICS

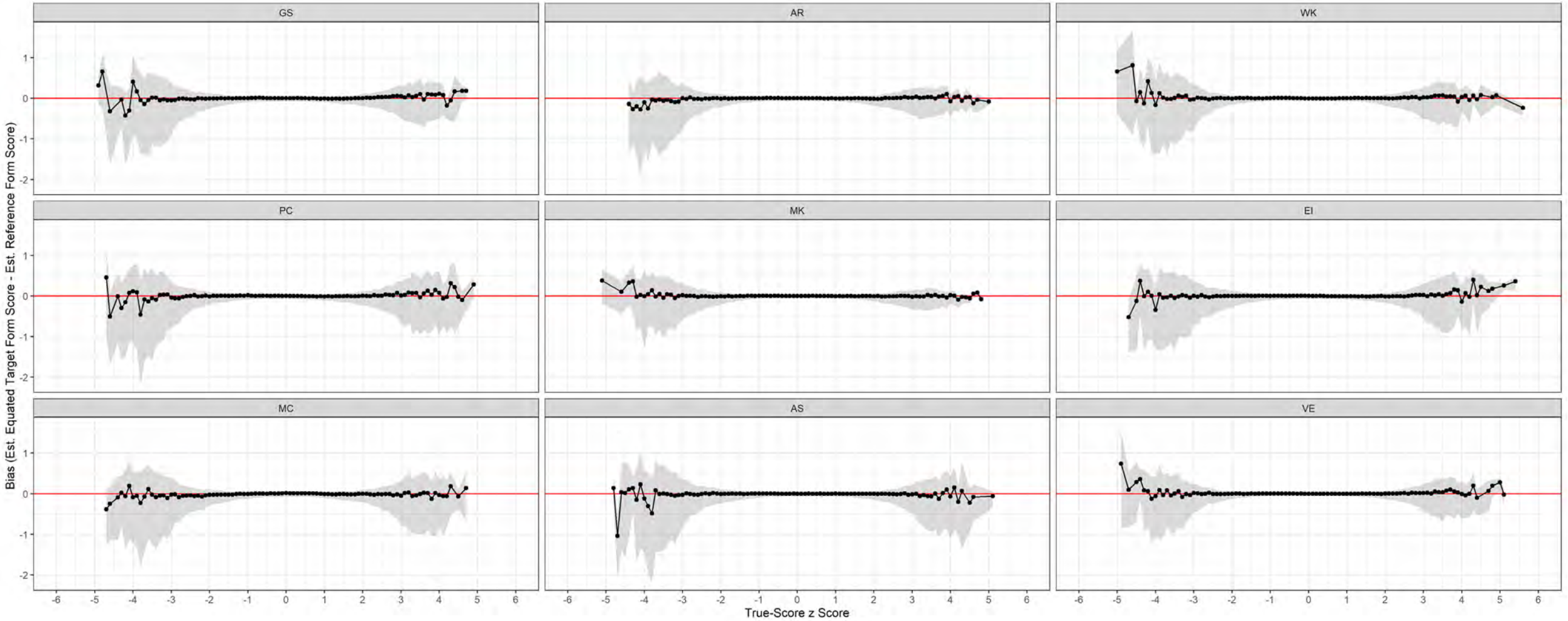
Evaluation of Unequated Standard Score Bias by True-Score *z* Score



OPA
OFFICE OF PEOPLE ANALYTICS

Note. Error ribbons represent 95% confidence intervals.

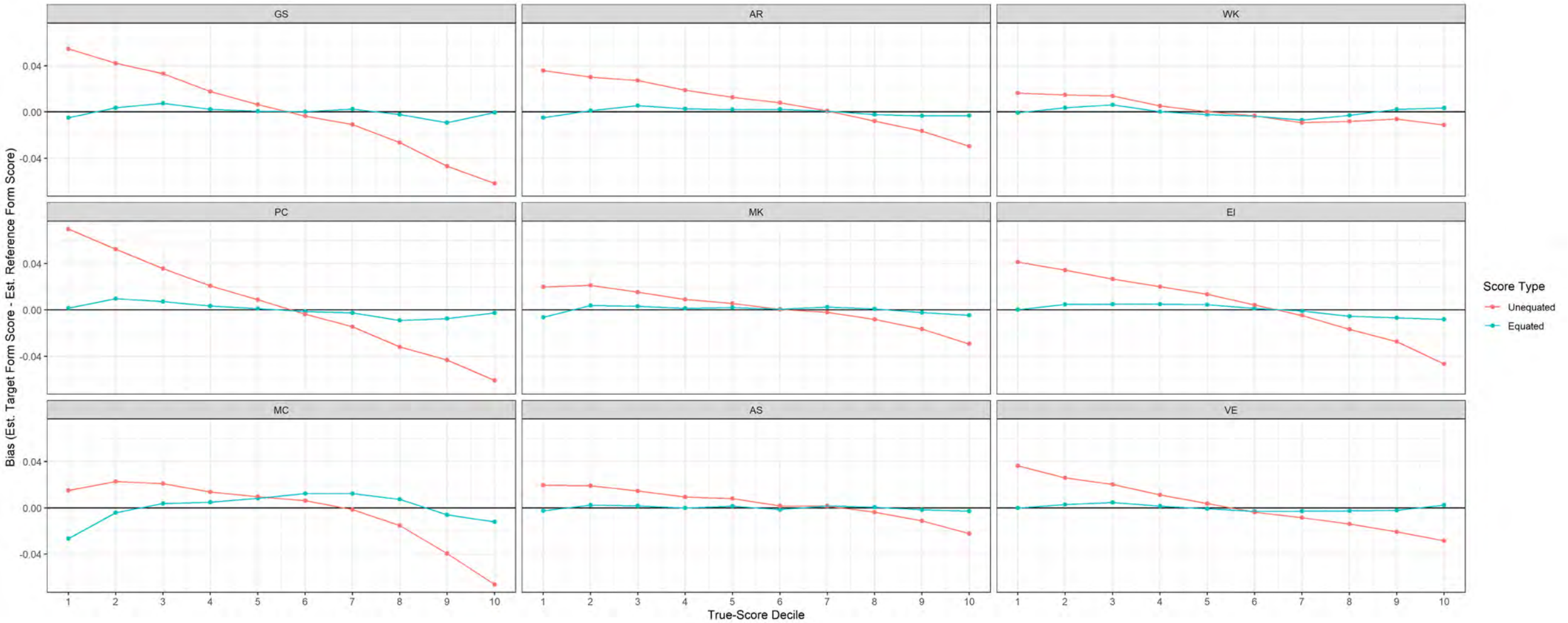
Evaluation of Equated Standard Score Bias by True-Score *z* Score



OPA
OFFICE OF PEOPLE ANALYTICS

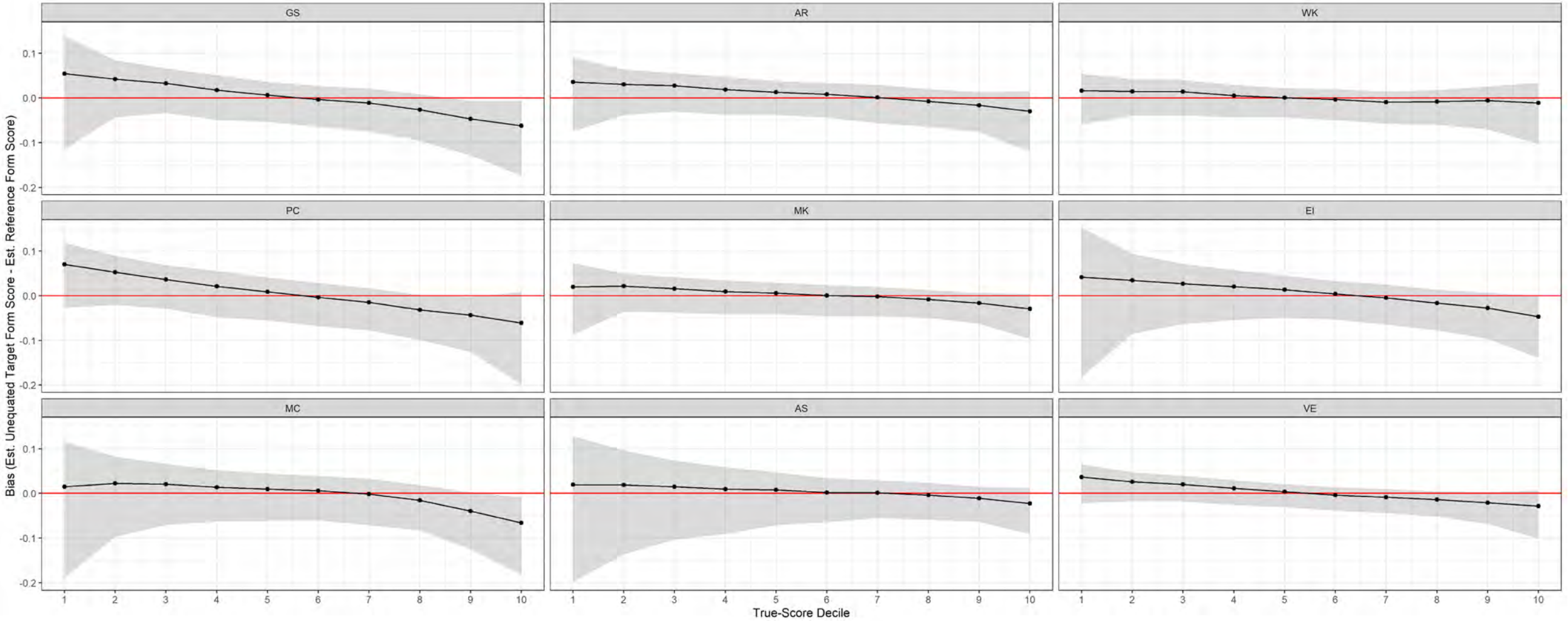
Note. Error ribbons represent 95% confidence intervals.

Evaluation of Standard Score Bias by True-Score Decile



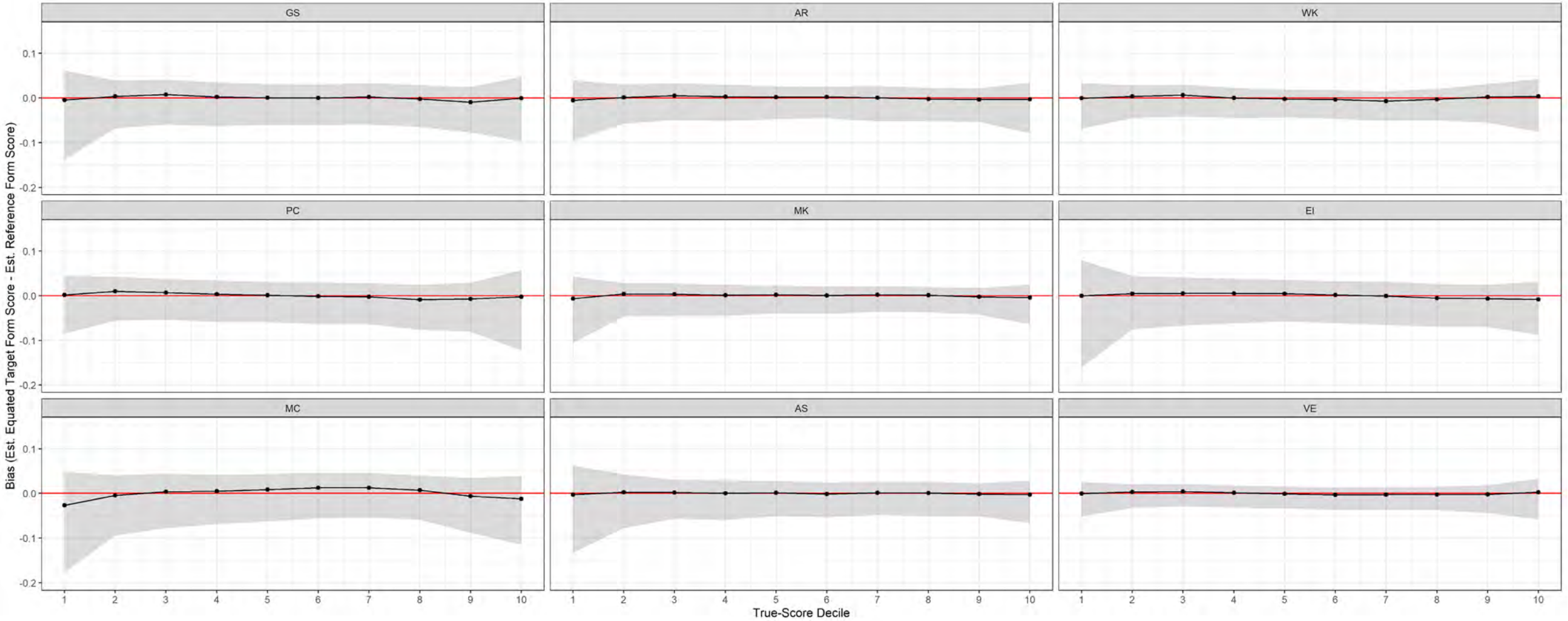
OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation of Unequated Standard Score Bias by True-Score Decile



Note. Error ribbons represent 95% confidence intervals.

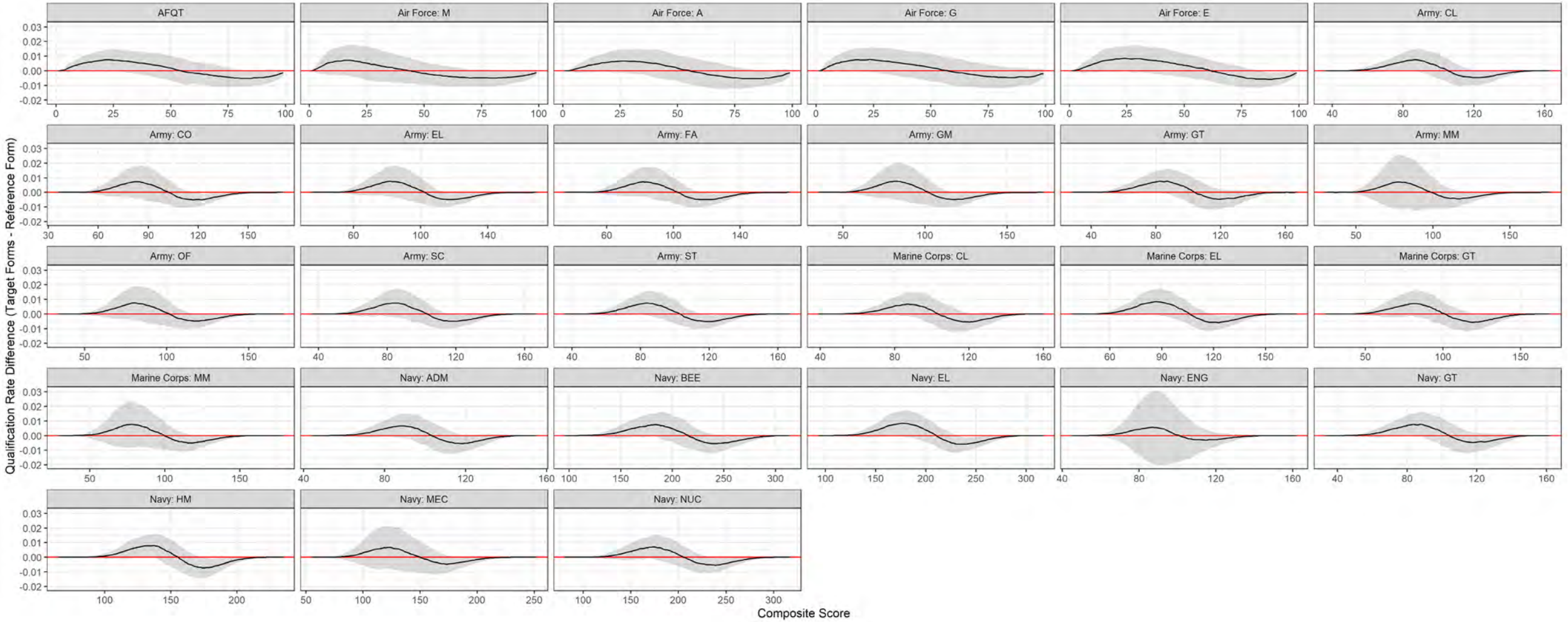
Evaluation of Equated Standard Score Bias by True-Score Decile



OPA
OFFICE OF PEOPLE ANALYTICS

Note. Error ribbons represent 95% confidence intervals.

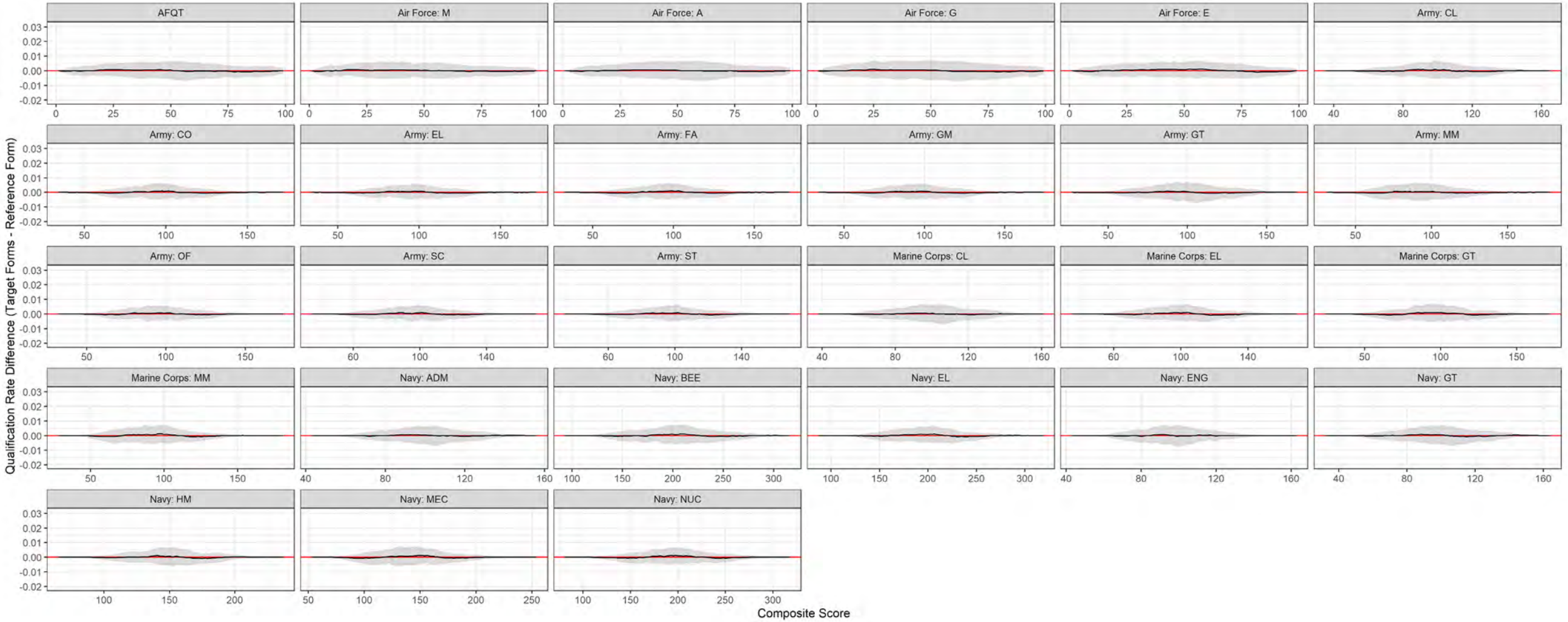
Qualification Rate Differences for Unequated Composite Scores



OPA
OFFICE OF PEOPLE ANALYTICS

Note. Error ribbons represent 95% confidence intervals.

Qualification Rate Differences for Equated Composite Scores



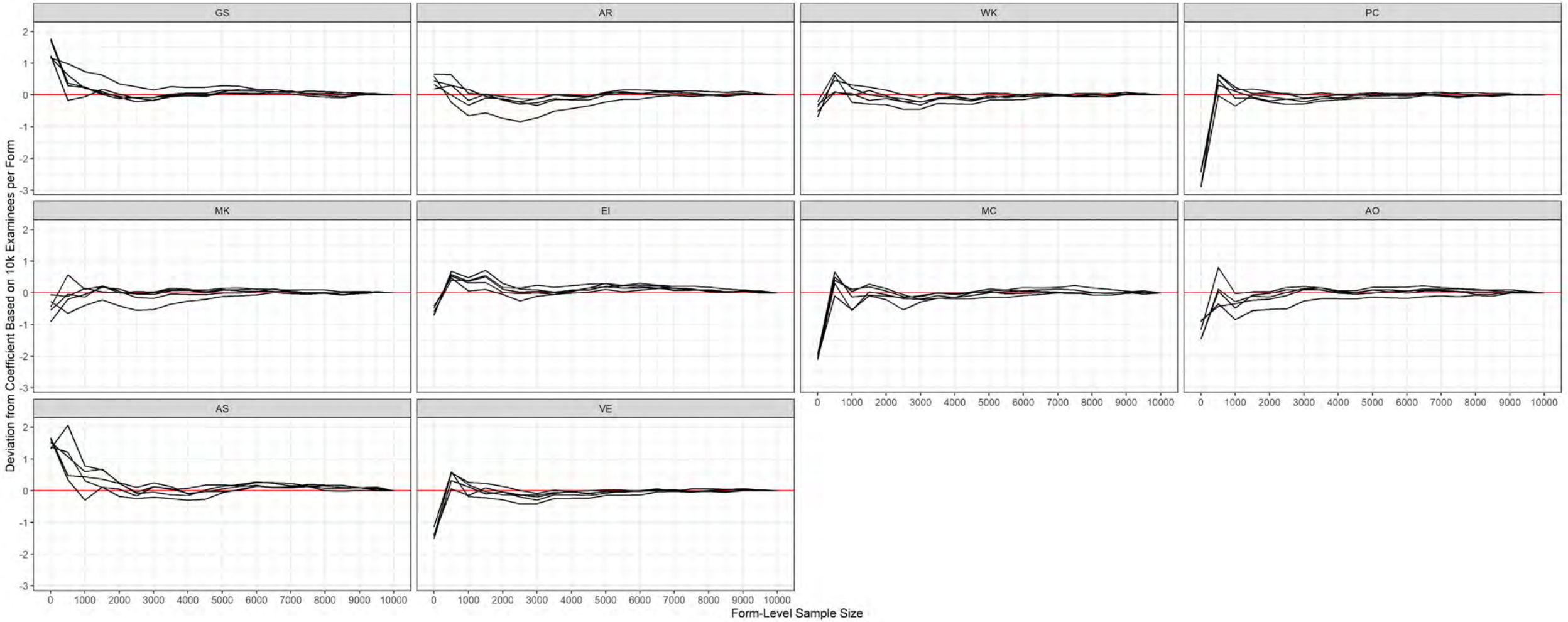
OPA
OFFICE OF PEOPLE ANALYTICS

Note. Error ribbons represent 95% confidence intervals.

Supplemental Slides: Plots of TC Convergence and Sampling Error per TC Coefficient

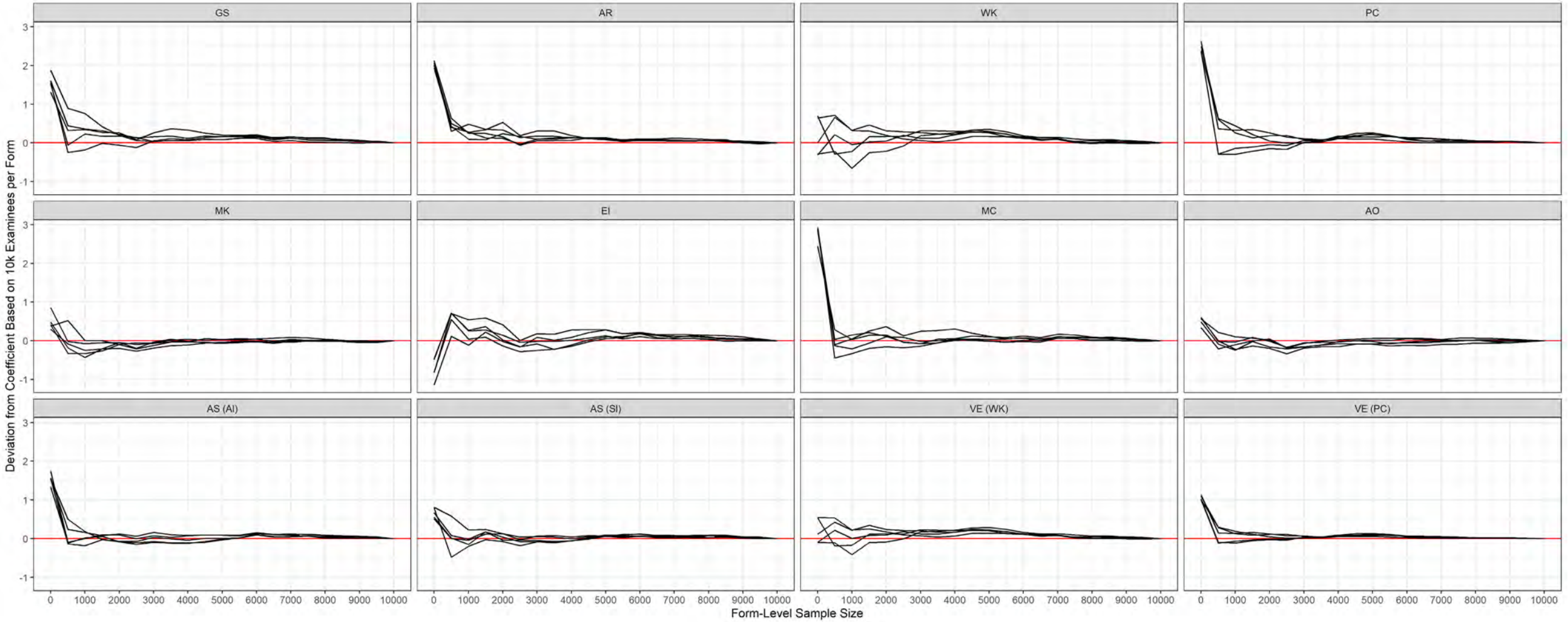
OPA
OFFICE OF PEOPLE ANALYTICS

TC Convergence with N_{Form}= 10k Solution for Intercept Coefficients



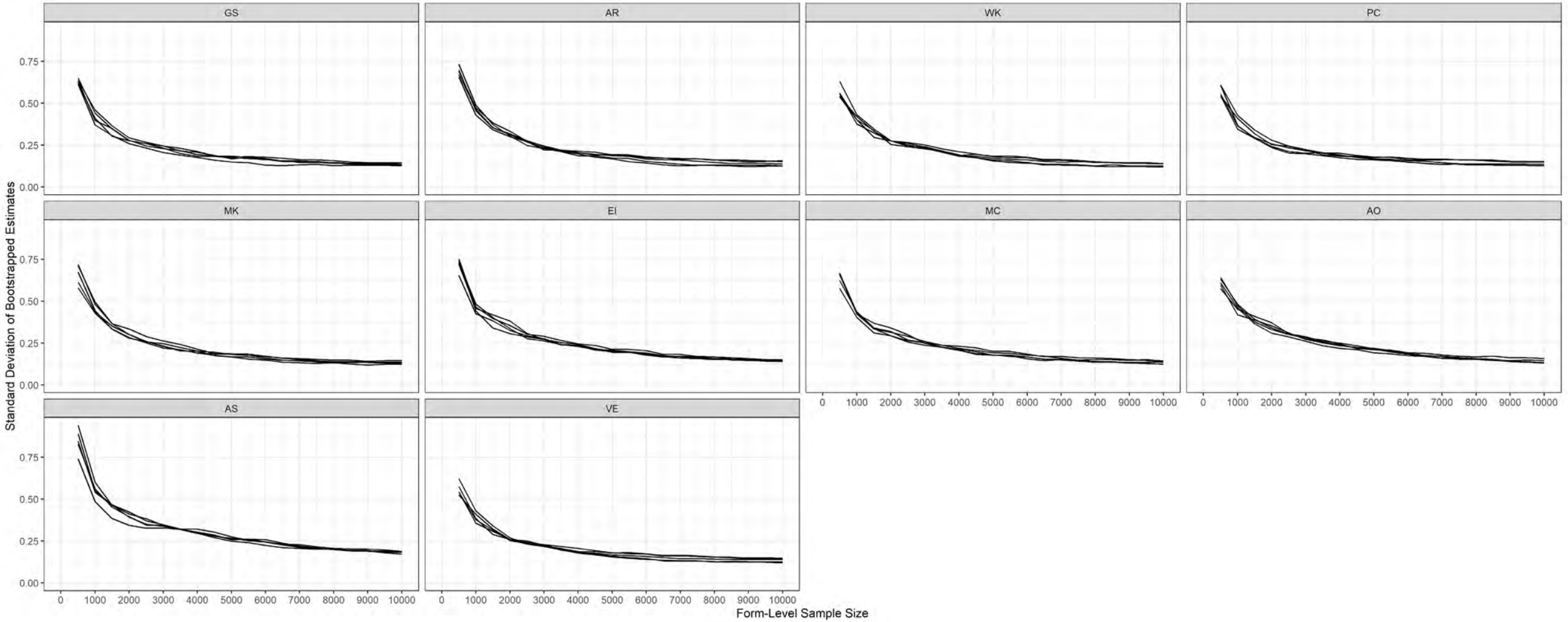
OPA
OFFICE OF PEOPLE ANALYTICS

TC Convergence with N_{Form}= 10k Solution for Slope Coefficients



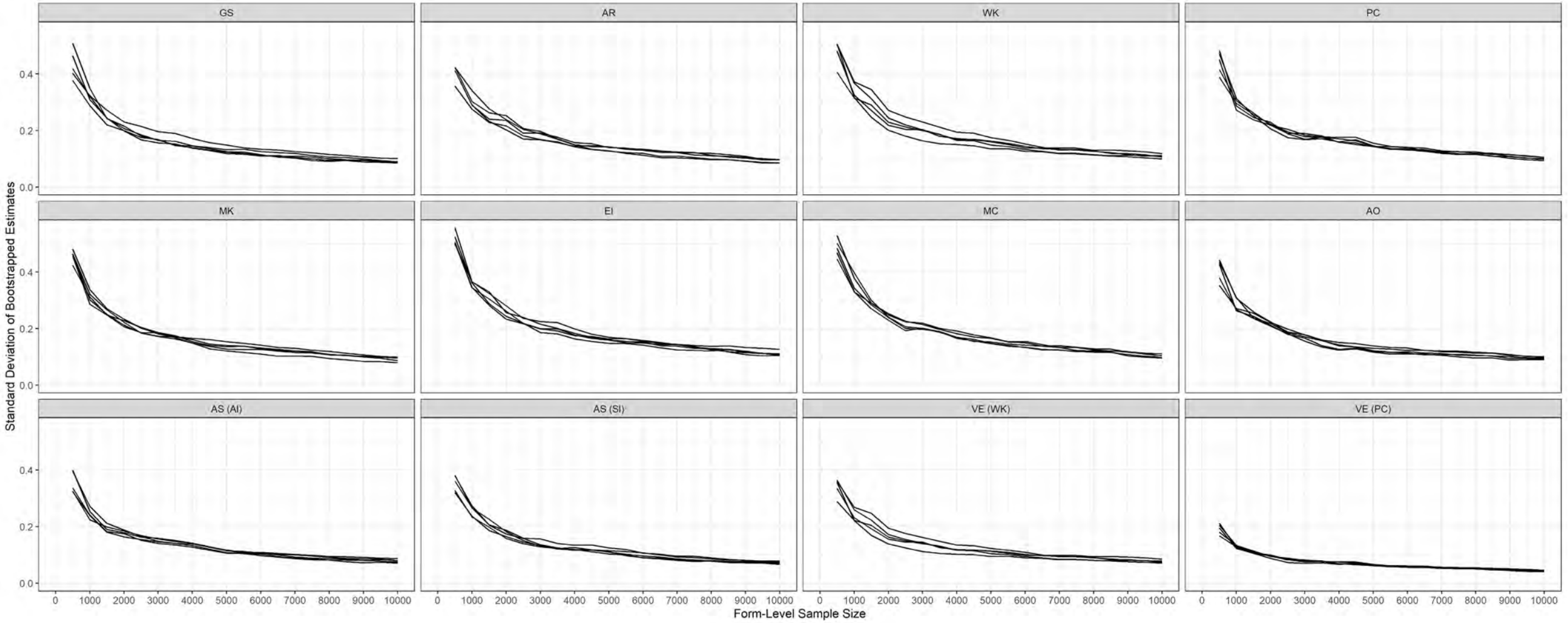
OPA
OFFICE OF PEOPLE ANALYTICS

Bootstrapped Standard Errors for Intercept Coefficients



OPA
OFFICE OF PEOPLE ANALYTICS

Bootstrapped Standard Errors for Slope Coefficients



OPA
OFFICE OF PEOPLE ANALYTICS

Tab J

OPA

OFFICE OF PEOPLE ANALYTICS

Development of a Complex Reasoning (CR) Test

Katherine Klein

Human Resources Research Organization (HumRRO)

Briefing presented to the DACMPT

January 22, 2024

Agenda

- Background
- Development Update
- Overview of New Task Order
- Pilot Study Three

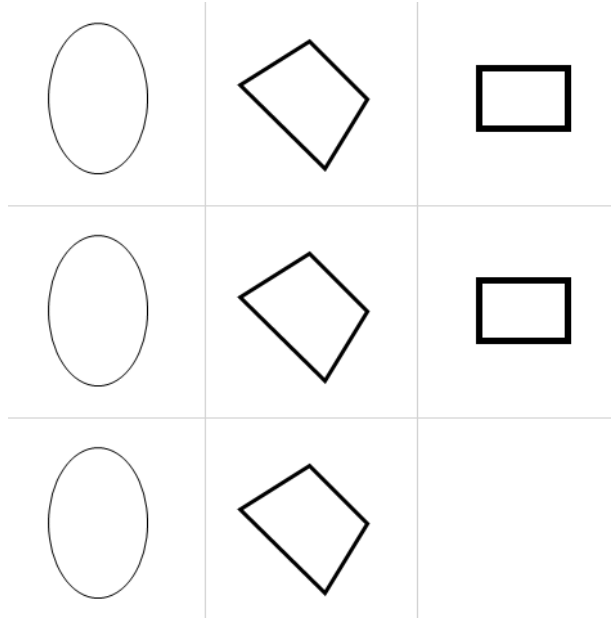
OPA
OFFICE OF PEOPLE ANALYTICS

Background

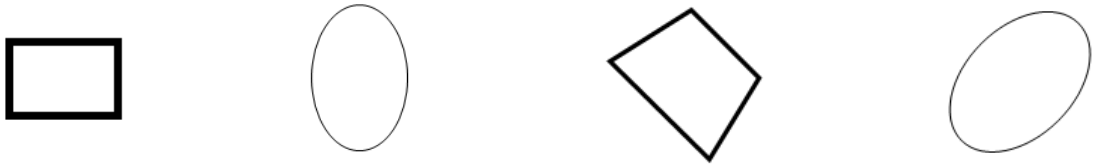
- ***What is complex reasoning?***
  - Non-verbal reasoning; ability to analyze visual information and to solve problems using visual reasoning
- ***Why a complex reasoning test?***
  - Fluid intelligence has been found to be a strong predictor of training and job success
    - Complex (non-verbal) reasoning is one element of fluid intelligence
    - ASVAB Review Panel (2006) recommended that DoD consider adding tests of fluid intelligence to balance the ASVAB's composition (between fluid and crystalized intelligence)
  - Potential benefits to the ASVAB testing program
    - Improved prediction of training and job success in military jobs
    - Lower susceptibility to test compromise
    - Less adverse impact; increased qualification rates for non-native and non-heritage English speakers

Sample Transformation Item

Look at the 3X3 grid below. Identify the pattern(s).



Which of the following images best completes the pattern(s) in the grid?



- **Transformation item features**
  - Types of shapes
  - Orientation of shape(s)
  - Size of shape(s)
  - Number of shape(s)
  - Line weighting on shape(s)
- **Direction(s) of transformations**
  - Vertical
  - Horizontal
  - Diagonal

OPA
OFFICE OF PEOPLE ANALYTICS

Development Update

Launched on the ASVAB Platform

- August 13, 2024
  - Four forms are static, and the 24 items constituting each form are administered in a specified presentation order

Available to Applicants

- September 16, 2024
  - A total of 9,837 applicants have taken the assessment between September 24 – November 4.

OPA
OFFICE OF PEOPLE ANALYTICS

CR Operational Descriptives		
	Raw	Standard Score
Mean	17.03	52.30
Standard Deviation	5.04	10.34
Min	0	17
5th Pct	7	32
25th Pct	14	46
50th Pct	18	54
75th Pct	21	60
95th Pct	23	65
Max	24	67
Correlation with ASVAB and Special Tests		
Armed Forces Qualification Test (AFQT)	.56	
Assembling Objects (AO)	.56	
Arithmetic Reasoning (AR)	.52	
Mechanical Comprehension (MC)	.51	
Math Knowledge (MK)	.49	
Cyber Test (CT)	.46	
General Science (GS)	.45	
Paragraph Comprehension (PC)	.45	
Verbal Expression (VE)	.45	
Electronics Information (EI)	.40	
Word Knowledge (WK)	.40	
Auto-Shop Information (AS)	.26	

Note. Correlations are observed and uncorrected; VE is a composite of WK and PC

Complex Reasoning (CR) Task Order

Line of Effort (LOE)
LOE 1: Design CR Items & Piloting Procedures • Dimensionality analyses and calibrations • Design CR item piloting data collection • Develop test blueprint for CAT version • Develop new CR items
LOE 2: Pilot New Items and Assemble CAT Pools • Pilot new CR items • Conduct item analysis • Develop CAT pools and conventional forms • Scale and equate scores
LOE 3: Recommend Refinements to Procedures • Identify refinements for test blueprints, item generation, and form assembly
LOE 4: Evaluate CR and CompT Scores • Create research plans to evaluate construct validity, criterion-related validity, ongoing psychometrics analysis, and coachability and practice effects
LOE 5: Document CR and CompT • Document task order efforts

OPA
OFFICE OF PEOPLE ANALYTICS

Pilot Study Three

OPA
OFFICE OF PEOPLE ANALYTICS

Wave 1

MEPS CR Form Version 1 (24 items)

- MEPS CR Form 1a w/CAT-like Order (24 items)
- MEPS CR Form 1b w/CAT-like Order (24 items)
- MEPS CR Form 1c w/CAT-like Order (24 items)
- MEPS CR Form 1d w/CAT-like Order (24 items)

Wave 2

MEPS CR Form Version 1 (24 items)

- New CR Tryout Set A (24 new items)
- New CR Tryout Set B (24 new items)
- New CR Tryout Set C (24 new items)
- New CR Tryout Set D (24 new items)

Wave 3

MEPS CR Form Version 1 (24 items)

- New CR Tryout Set E (24 new items)
- New CR Tryout Set F (24 new items)
- New CR Tryout Set G (24 new items)
- New CR Tryout Set H (24 new items)

Wave 4

MEPS CR Form Version 1 (24 items)

- New CR Tryout Set I (24 new items)
- New CR Tryout Set J (24 new items)
- New CR Tryout Set K (24 new items)
- New CR Tryout Set L (24 new items)

OPA
OFFICE OF PEOPLE ANALYTICS

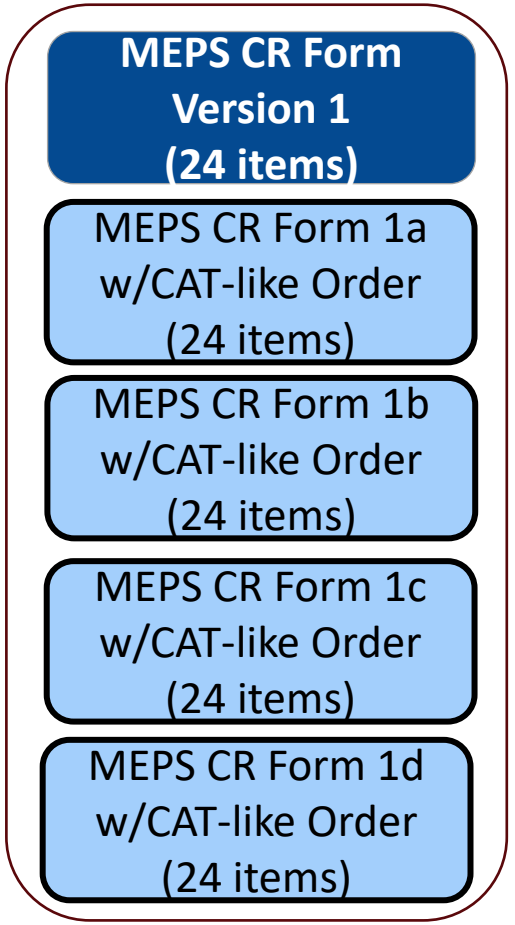
Wave 1 Overview

Objective

- Determine whether non-progressive item order impacts item functioning and test performance
  - Findings influence the feasibility of a CAT CR

Design and Measures

- 24 CR items, 5 static forms
- Pre- and post-test questionnaire
- Two CR attention-check items + insufficient effort



Sample

- Non-military sample representative of military applicants, ages 18–35, U.S. citizen, HS degree/GED/<1 year of college
- Targeted *N* = 5,250 participants (~1,050 participants per form)

Method

- Administered on Qualtrics platform
- Participants randomly assigned to one CR form
- 35-minute fixed time limit
- Record time to completion
- Desktop or laptop only

OPA
OFFICE OF PEOPLE ANALYTICS

Wave 1 Data Collection (as of 1 November)

Group	Pilot 3 (as of 1 November 2024)					
	MEPS Version	Form 1a	Form 1b	Form 1c	Form 1d	All Forms (Combined)
Total	109	107	109	94	101	502
Female	67	63	64	58	59	311
Asian	2	2	6	2	5	17
Black	31	30	18	25	29	133
Hispanic	27	28	21	17	19	112

OPA
OFFICE OF PEOPLE ANALYTICS

Waves 2 – 4 Overview

Objective

- Pilot test 288 new CR items for potential inclusion on the ASVAB platform
- Evaluate, calibrate, and link new CR items to new base IRT scale (estimated with operational CR data)

Sample

- Non-military sample representative of military applicants, ages 18–35, U.S. citizen, HS degree/ GED/<1 year of college
- Targeted *N* = 5,250 participants (~525 participants per form; 1,050 responses per item)

Design and Measures

- 24 CR items per examinee, multiple static forms with overlapping items
- Pre- and post-test questionnaire
- Two CR attention check items + insufficient effort

Method

- Administered on Qualtrics platform
- Within each wave, participants randomly assigned to one CR form
- 35-minute fixed time limit
- Record time to completion
- Desktop or laptop only

OPA
OFFICE OF PEOPLE ANALYTICS

Challenge & Methodology

- Determine how to calibrate and link the new CR items to base scale estimated from operational data on applicants
  - Conducted a simulation study (100 replications) to evaluate the three data collection designs and the four calibration designs to determine which resulted in the best psychometric solution

Data Collection Design Options Included:

1. Gold Standard—Operational + randomly seeded new items*
2. Fully Crossed—Every combination of evens and odds of new item sets with operational (e.g., even A, odd B)
3. Daisy Chain—Chained combinations of even and odd new item sets with operational
4. Random groups—Randomly assign one of five intact item sets (operational or one of four new item sets)

Scaling Method Options Included:

1. BILOG-Scaled Params*
2. True-Scaled Params*
3. Fixed OP Params
4. Fixed OP Params (Rescaled)
5. Latent Mu-Sigma Scaled
6. Stocking-Lord Equated

OPA
OFFICE OF PEOPLE ANALYTICS

*Comparison group only; option not being considered

Solution

Daisy Chain Design: 10 combinations of even-odd item sets across the operational form and four experimental item sets

Reasons for Recommendation:

1. All designs performed very similarly on psychometric metrics
2. Allows for common items, guards against deviations from randomly equivalent groups
3. Less intensive effort compared to fully crossed design

	OP Even	A Even	B Even	C Even	D Even
OP Odd	X	X			
A Odd		X	X		
B Odd			X	X	
C Odd				X	X
D Odd	X				X

OP = Operational Form

Note. Results can be reviewed in the back-up slides.

OPA
OFFICE OF PEOPLE ANALYTICS

Steps

1. Collect sufficient data at MEPS from military applicants on operational CR form (4 versions, same 24 items). MEPS military applicant sample and CR form to establish the new IRT base scale ***completed**
2. Calibrate operational CR form (24 items), derive new base scale using operational data on MEPS military applicant sample (Step 1) ***completed**
3. Pilot 288 new CR items (96 items per wave) using the daisy-chain design with non-military sample
4. Calibrate 288 new CR items (96 items per wave) using data collected (Step 3) and link to the new base scale (Step 2), scaling approach TBD (e.g., fixing parameters to operational MEPS sample, scaling to latent mu-sigma of operational MEPS sample, Stockard-Lord equating)

OPA
OFFICE OF PEOPLE ANALYTICS

Questions for the DAC

- Does the DAC have any feedback on the Daisy-Chain design and plan for scaling and linking new CR items to the new base scale in Waves 2–4?
- Are there are any analyses we should consider for evaluating the feasibility of an adaptive CR version from the Wave 1 data?
- Are there any thoughts on creating an adaptive version of CR?

OPA
OFFICE OF PEOPLE ANALYTICS

Acknowledgments

Matthew Brown, *HumRRO*

Mike Ingerick, *HumRRO*

Scott Oppler, *HumRRO*

Sergio Marquez, *HumRRO*

Nathanial Voss, *HumRRO*

Alex Burgoyne, *HumRRO*

Leilani Seged, *HumRRO*

Robert Wellman, *HumRRO*

Sachi Phillips, *HumRRO*

Furong Gao, *HumRRO*

Jeff Dahlke, *HumRRO*

Mary Pommerich, *DTAC*

Matt Trippe, *DTAC*

Tia Fechter, *DTAC*

Jeff Harber, *DTAC*

Ping Yin, *DTAC*

OPA
OFFICE OF PEOPLE ANALYTICS

Thank you!

For more information
please contact:

Katherine Klein
KKlein@HumRRO.org
651.370.210

OPA
OFFICE OF PEOPLE ANALYTICS

Back-up Slides

OPA
OFFICE OF PEOPLE ANALYTICS

Simulation Results — Bias

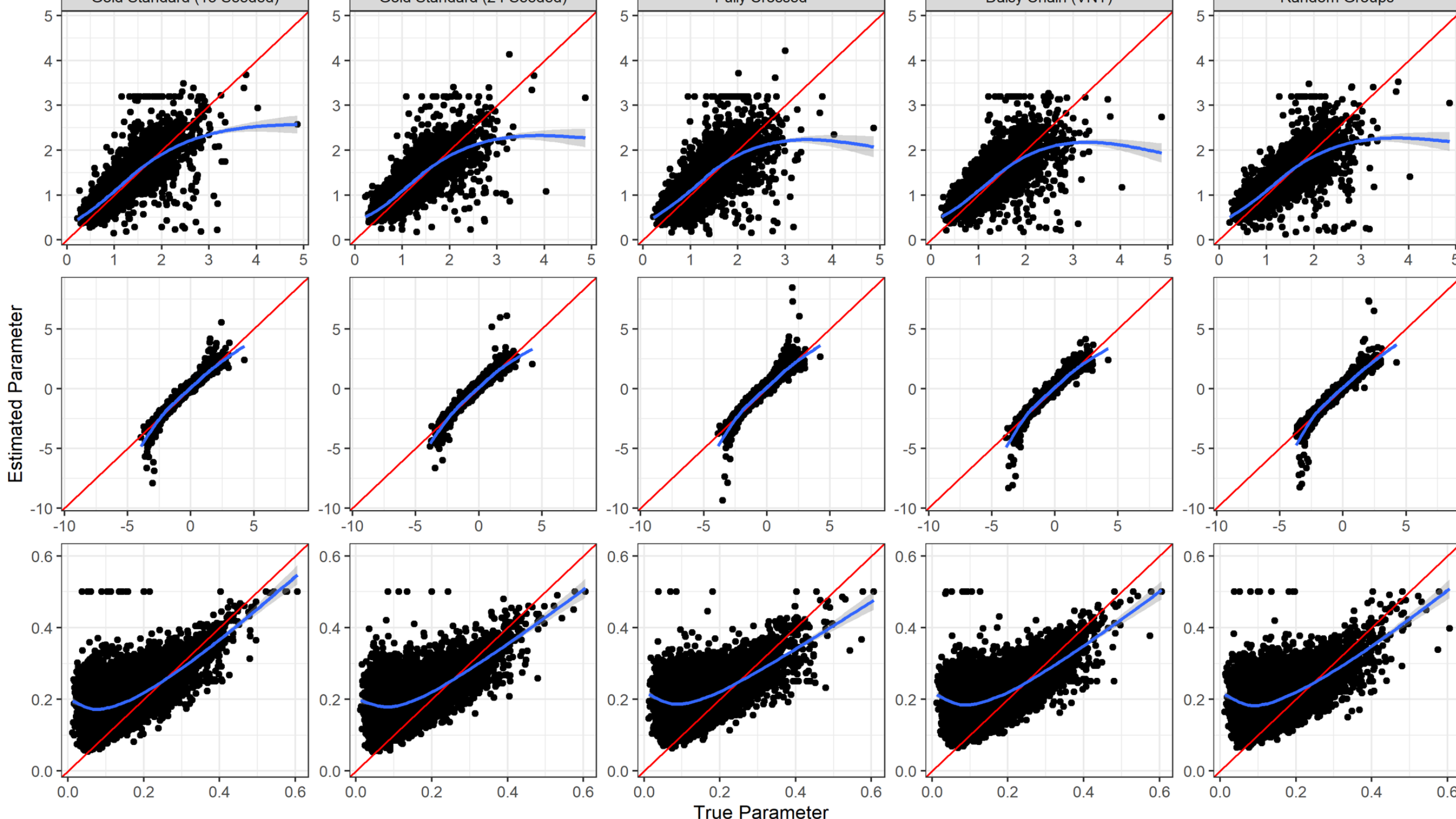
Evaluation	Scaling Method	Gold Standard (15 Seeded)	Gold Standard (24 Seeded)	Fully Crossed	Daisy Chain (VNT)	Random Groups
ICC	BILOG-Scaled Params	-0.047	-0.047	-0.047	-0.047	-0.047
ICC	True-Scaled Params	0.000	0.000	0.000	0.000	0.000
ICC	Fixed OP Params	-0.008	-0.009	-0.034	-0.044	-0.052
ICC	Fixed OP Params (Rescaled)	-0.003	-0.004	-0.003	-0.003	-0.002
ICC	Latent Mu-Sigma Scaled	0.000	0.000	-0.003	0.002	0.000
ICC	Stocking-Lord Equated	0.001	0.001	0.001	0.002	0.001
a	BILOG-Scaled Params	0.013	0.018	0.025	0.019	0.017
a	True-Scaled Params	0.067	0.073	0.081	0.074	0.072
a	Fixed OP Params	0.028	0.033	0.011	0.008	0.010
a	Fixed OP Params (Rescaled)	0.049	0.053	0.065	0.054	0.050
a	Latent Mu-Sigma Scaled	0.062	0.066	0.091	0.073	0.065
a	Stocking-Lord Equated	0.060	0.066	0.082	0.068	0.064
b	BILOG-Scaled Params	0.336	0.343	0.359	0.349	0.343
b	True-Scaled Params	0.081	0.088	0.103	0.093	0.087
b	Fixed OP Params	0.101	0.108	0.208	0.245	0.284
b	Fixed OP Params (Rescaled)	0.076	0.081	0.086	0.076	0.078
b	Latent Mu-Sigma Scaled	0.081	0.087	0.121	0.083	0.087
b	Stocking-Lord Equated	0.075	0.080	0.094	0.081	0.081
c	BILOG-Scaled Params	0.055	0.059	0.063	0.062	0.061
c	True-Scaled Params	0.055	0.059	0.063	0.062	0.061
c	Fixed OP Params	0.050	0.052	0.049	0.046	0.048
c	Fixed OP Params (Rescaled)	0.046	0.048	0.051	0.046	0.051
c	Latent Mu-Sigma Scaled	0.055	0.059	0.063	0.062	0.061
c	Stocking-Lord Equated	0.055	0.059	0.063	0.062	0.061

Simulation Results — RMSE

Evaluation	Scaling Method	Gold Standard (15 Seeded)	Gold Standard (24 Seeded)	Fully Crossed	Daisy Chain (VNT)	Random Groups
ICC	BILOG-Scaled Params	0.061	0.062	0.063	0.062	0.062
ICC	True-Scaled Params	0.022	0.024	0.023	0.023	0.023
ICC	Fixed OP Params	0.024	0.029	0.048	0.061	0.069
ICC	Fixed OP Params (Rescaled)	0.029	0.033	0.029	0.035	0.028
ICC	Latent Mu-Sigma Scaled	0.022	0.024	0.029	0.025	0.024
ICC	Stocking-Lord Equated	0.022	0.024	0.024	0.024	0.025
a	BILOG-Scaled Params	0.213	0.223	0.245	0.234	0.230
a	True-Scaled Params	0.226	0.236	0.261	0.247	0.243
a	Fixed OP Params	0.193	0.212	0.211	0.216	0.214
a	Fixed OP Params (Rescaled)	0.210	0.227	0.231	0.227	0.221
a	Latent Mu-Sigma Scaled	0.225	0.233	0.284	0.251	0.242
a	Stocking-Lord Equated	0.229	0.240	0.273	0.250	0.246
b	BILOG-Scaled Params	0.390	0.399	0.435	0.411	0.417
b	True-Scaled Params	0.202	0.211	0.251	0.225	0.240
b	Fixed OP Params	0.214	0.245	0.327	0.385	0.399
b	Fixed OP Params (Rescaled)	0.256	0.274	0.237	0.288	0.234
b	Latent Mu-Sigma Scaled	0.204	0.213	0.271	0.231	0.244
b	Stocking-Lord Equated	0.200	0.204	0.237	0.221	0.238
c	BILOG-Scaled Params	0.083	0.087	0.092	0.090	0.089
c	True-Scaled Params	0.083	0.087	0.092	0.090	0.089
c	Fixed OP Params	0.081	0.086	0.086	0.088	0.087
c	Fixed OP Params (Rescaled)	0.085	0.090	0.086	0.088	0.085
c	Latent Mu-Sigma Scaled	0.083	0.087	0.092	0.090	0.089
c	Stocking-Lord Equated	0.083	0.087	0.092	0.090	0.089

Simulation Results — r

Evaluation	Scaling Method	Gold Standard (15 Seeded)	Gold Standard (24 Seeded)	Fully Crossed	Daisy Chain (VNT)	Random Groups
a	BILOG-Scaled Params	0.862	0.847	0.817	0.831	0.836
a	True-Scaled Params	0.862	0.847	0.817	0.831	0.836
a	Fixed OP Params	0.889	0.865	0.863	0.856	0.858
a	Fixed OP Params (Rescaled)	0.873	0.850	0.851	0.850	0.858
a	Latent Mu-Sigma Scaled	0.861	0.848	0.793	0.825	0.833
a	Stocking-Lord Equated	0.856	0.839	0.802	0.824	0.828
b	BILOG-Scaled Params	0.983	0.982	0.975	0.980	0.976
b	True-Scaled Params	0.983	0.982	0.975	0.980	0.976
b	Fixed OP Params	0.983	0.977	0.970	0.958	0.963
b	Fixed OP Params (Rescaled)	0.971	0.966	0.976	0.962	0.976
b	Latent Mu-Sigma Scaled	0.983	0.982	0.972	0.978	0.975
b	Stocking-Lord Equated	0.984	0.983	0.978	0.980	0.976
c	BILOG-Scaled Params	0.586	0.554	0.500	0.518	0.519
c	True-Scaled Params	0.586	0.554	0.500	0.518	0.519
c	Fixed OP Params	0.569	0.508	0.499	0.457	0.476
c	Fixed OP Params (Rescaled)	0.507	0.447	0.506	0.457	0.515
c	Latent Mu-Sigma Scaled	0.586	0.554	0.500	0.518	0.519
c	Stocking-Lord Equated	0.586	0.554	0.500	0.518	0.519


Gold Standard (10 Seeded)
Gold Standard (24 Seeded)
Fully Crossed
Baby Shaln (VNT)
Random Groups
Estimated Parameter
True Parameter

Tab K

OPA

OFFICE OF PEOPLE ANALYTICS

Computational Thinking

Kimberly Adams and Scott Oppler
Human Resources Research Organization

Briefing presented to the DACMPT
January 22, 2025

Briefing Agenda

- Background and Project Overview
  - Phase 1: Computational Thinking Score
  - Phase 2: Computational Thinking Score Validation
- Predictors and Criterion
- Phase 2 Results
- Closing

OPA
OFFICE OF PEOPLE ANALYTICS

Congressional Mandate

- William M. (Mac) Thornberry National Defense Authorization Act (NDAA) for Fiscal Year 2021 (HR 6395), Section 594
  - Must assess six (6) computational thinking construct domains
    - Problem Decomposition
    - Abstraction
    - Pattern Recognition
    - Analytical Ability
    - Identifying Variables for Data Representation
    - Creating Algorithms and Solution Expressions
  - Must be available for operational use by October 1, 2024

OPA
OFFICE OF PEOPLE ANALYTICS

Computational Thinking Construct Domains

Construct Domains	Descriptions
1. Problem decomposition	▪ Break down a problem/task into smaller/easier components (e.g., describe a system as a sequence of processes)
2. Abstraction	▪ Focus on the most relevant information and ignore extraneous information to interpret meaning and reduce complexity of a problem/task
3. Pattern recognition	▪ Identify and use repeated information or patterns to predict outcomes or determine actions for a problem/task
4. Analytical ability	▪ Inspect, cleanse, transform, and model data with the goal of discovering useful information for a problem/task
5. Identifying variables for data representation	▪ Recognize how parts of a solution may be reapplied to, or eliminated from, similar or unique problems/tasks
6. Creating algorithms and solution expressions	▪ Recognize and evaluate options against outcomes to simplify or automate processes for efficiency and resource utilization improvements

Where We Started

- Existing measures of computational thinking were not viable
  - Those used for selection require specific programming language skills
  - Those used for skill acquisition are developed for the K–12 classroom environment, which are free on the internet (lack test security)
- NDAA-specified deadline of 01 October 2024 did not support creating a new, valid measure of computational thinking
- Belief that the Complex Reasoning Test (CR) already under development, and possibly some of the ASVAB subtests [e.g., Arithmetic Reasoning (AR), Assembling Objects (AO)] and other special tests [e.g., Cyber (CT), Coding Speed (CS), Mental Counters (MCt)] were likely assessing the computational thinking construct domains

Project Overview

Phase 1: Define Computational Thinking Score Equation

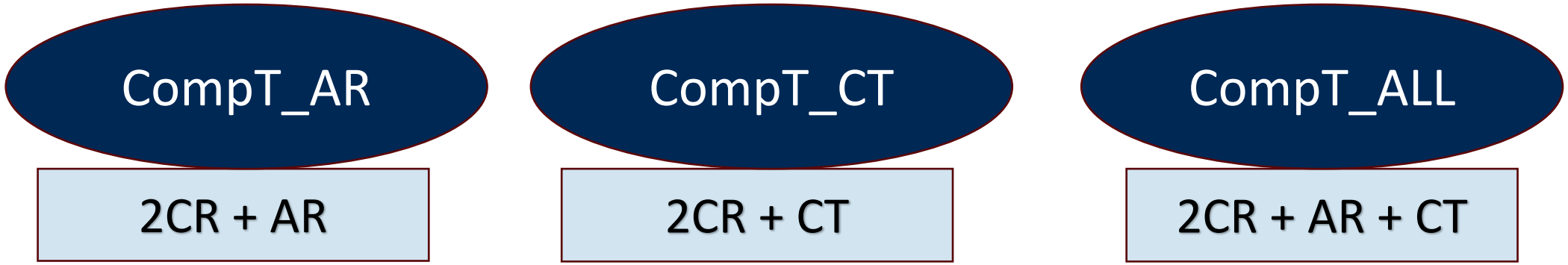
- Gather empirical & SME-estimated correlations
- Specify & analyze prediction models
- Generate, evaluate, finalize synthetic CompT score equations
- Submit software requirements & specifications

Phase 2: Verify Validity of Computational Thinking Scores

- Select computational thinking marker test
- Develop & implement data collection plan at MEPS
- Match shippers' ASVAB & CT scores to study data & clean
- Conduct analyses & summarize results

OPA
OFFICE OF PEOPLE ANALYTICS

Computational Thinking Score Equations



Note: Scores are a weighted sum of CR, AR, and CT standard (T) scores with X = 50, std = 10. The AR, CR, and CT standard (T) scores are normed to the PAY97 sample.

OPA
OFFICE OF PEOPLE ANALYTICS

Validation Data Collection

Collected Data	Matched Data	Cleaned Data
▪ MEPS administered the Qualtrics data collection tool between 4/15 – 5/20 ✓Complex Reasoning (CR) ✓Computational Thinking Assessment for Middle Schoolers (CTA-M) ✓Background questions ▪ Shippers = 1,044	▪ HumRRO sent DTAC participant IDs from Qualtrics on weekly basis ▪ DTAC used participant IDs and MEPS rosters to pull ASVAB and CT scores into a de-identified dataset ▪ HumRRO appended with responses on CR, CTA-M, and background questions ▪ Shippers = 922	▪ Removed any that showed a lack of motivation using: ✓Two CR attention-check items ✓Self-report question at end ✓Time spent on CR and CTA-M (no more than 2 standard deviations below the mean for time spent) ✓Checks for careless response patterns ✓Checks for CR and CTA-M scores that were at or below chance ▪ Removed any that left study early for transportation ▪ Shippers = 722

OPA
OFFICE OF PEOPLE ANALYTICS

Sample by Demographic Group

Sex		Race-Ethnicity		Service*	
Female	106	Hispanic White (HW)	166	Air Force	232
Male	608	Non-Hispanic Asian (NHA)	35	Army	22
NA	8	Non-Hispanic Black (NHB)	172	Coast Guard	0
		Non-Hispanic White (NHW)	291	Marine Corps	214
		Other or NA	58	Navy	238
				Space Force	16
Total	722		722		722

*Participation was limited to Shippers with a pre-enlistment CT score. Therefore, an equal distribution across Services was not expected given Services have different policies for administering CT to applicants.

OPA
OFFICE OF PEOPLE ANALYTICS

Sample by Type of Service and Component

Component	Service						
	Army	Air Force	Coast Guard	Marine Corps	Navy	Space Force	Total
Active Duty	21	232	0	205	235	16	709
Guard	1	0	0	9	3	0	13
Reserve	0	0	0	0	0	0	0
Total	**22**	**232**	**0**	**214**	**238**	**16**	**722**

OPA
OFFICE OF PEOPLE ANALYTICS

Predictors and Criterion

OPA
OFFICE OF PEOPLE ANALYTICS

Overview of Predictors

- Components of operational equation-based Computational Thinking scores
  - AR
  - CT
  - CR
- Operational equation-based Computational Thinking scores derived from Phase 1 study
  - CompT_AR = 2CR + AR
  - CompT_CT = 2CR + CT
  - CompT_ALL = 2 CR + AR + CT

OPA
OFFICE OF PEOPLE ANALYTICS

Overview of Criterion

- Computational Thinking Abilities – Middle Grades Assessment (CTA-M)
  - Developed by Wiebe et al, 2019
  - Designed for classroom use with middle school students
- Consists of 23 items administered with a 45-minute time limit
  - 15 Computational Thinking Test (CTt) items (Gonzalez et al., 2015)
  - 8 Bebras items (2016 UK Bebras Challenge)
- Items map to two or three of the six construct domains based on consensus judgments by HumRRO team members
  - Problem Decomposition
  - Solving for Algorithms
  - Analytical Ability

OPA
OFFICE OF PEOPLE ANALYTICS

Predictor and Criterion Analyses

- Calculated score for each Shipper on:
  - CTA-M (criterion)
  - CR (predictor)
- Calculated the three CompT scores using the operational equations from Phase 1
  - CompT_AR
  - CompT_CT
  - CompT_ALL
- Computed predictor and criterion descriptive statistics
- Computed predictor and criterion reliability estimates (except AR and CT*)
- Computed predictor and criterion subgroup differences (except AR and CT*)

*For AR and CT, used existing estimates of reliability documented in psychometric checklists (Sinclair et al., 2003) and current estimates of subgroup differences for FY23 applicant data (Johnston-Fisher et al., 2024).

Predictor and Criterion Descriptives

Variable	Variable Type	Mean	Median	SD	Min	Max
CTA-M	Criterion	13.8	14.0	4.1	6	23
AR	Predictor	52.8	52.5	7.9	30	72
CT	Predictor	51.6	52.0	8.9	22	76
CR	Predictor	55.0	57.0	8.1	35	67
CompT_AR	Equation Score	162.8	167.0	20.8	103	202
CompT_CT	Equation Score	161.5	164.0	20.5	104	202
CompT_ALL	Equation Score	214.3	217.0	25.6	141	271

OPA
OFFICE OF PEOPLE ANALYTICS

Predictor and Criterion Reliabilities

Variable	Type of Variable	Reliability	
		Cronbach's Alpha	Mosier's Composite Formula
CTA-M	Criterion	0.73	—
AR*	Predictor	0.89	—
CT*	Predictor	0.70	—
CR	Predictor	0.82	—
CompT_AR	Equation Score	—	0.88
CompT_CT	Equation Score	—	0.83
CompT_ALL	Equation Score	—	0.88

*Cronbach's alpha obtained from Psychometrics Checklists as reported in Sinclair et al. (2023).

OPA
OFFICE OF PEOPLE ANALYTICS

Predictor and Criterion Subgroup Differences

Variable	Type of Variable	Cohen's *d* Effect Size			
		M-F	NHW-NHB	NHW-HW	NHW-NHA*
CTA-M	Criterion	0.28	0.54	0.19	0.35
AR	Predictor	0.25	0.43	0.10	-0.01
CT	Predictor	0.44	0.36	0.21	0.18
CR	Predictor	0.07	0.23	0.11	-0.01
CompT_AR	Equation Score	0.15	0.34	0.12	-0.01
CompT_CT	Equation Score	0.25	0.34	0.18	0.08
CompT_ALL	Equation Score	0.28	0.41	0.17	0.06

Effect Size Category	
Less than Small	<0.20
Small	0.20 – 0.49
Moderate	0.50 – 0.79

*Sample size for Non-Hispanic Asian subgroup is too small to support interpretation of effect sizes.

OPA
OFFICE OF PEOPLE ANALYTICS

Phase 2 Results

OPA
OFFICE OF PEOPLE ANALYTICS

Data Analysis Plan

- Calculate zero-order correlations between CTA-M and the three components (AR, CT, CR) in the three Computational Thinking score equations
  - Correct results for range restriction
  - Disattenuate results for criterion unreliability
- Calculate zero-order correlations between CTA-M and the three operational equation-based Computational Thinking scores developed in Phase 1
  - Correct results for range restriction
  - Disattenuate results for criterion unreliability
- Estimate empirical validity of non-negative least square (NNLS) regression equations using data from Phase 2 validation study
  - Correct results for range restriction
  - Disattenuate results for criterion unreliability
  - Adjust results for shrinkage
- Conduct post-hoc analysis to recompute estimates using all 9 ASVAB subtests, CT, and CR

Correlation of Equation Component Tests with CTA-M

Equation Component Test	Correlation with CTA-M	
	Observed	Corrected*
AR	0.48	0.71
CT	0.40	0.61
CR	0.54	0.73

*Results are corrected for multivariate range restriction and disattenuated for criterion unreliability.

OPA
OFFICE OF PEOPLE ANALYTICS

Correlation of Operational Equation-based Scores with CTA-M

Operational Equation-Based Score	Correlation with CTA-M	
	Observed	Corrected*
CompT_AR	0.61	0.81
CompT_CT	0.60	0.80
CompT_ALL	0.63	0.83

*Results are corrected for multivariate range restriction and disattenuated for criterion unreliability.

OPA
OFFICE OF PEOPLE ANALYTICS

NNLS Regression Results by Operational Equation Scores

Regression Coefficient/ Multiple R	CompT_AR		CompT_CT		CompT_ALL	
	Observed	Corrected*	Observed	Corrected*	Observed	Corrected*
AR	0.15	0.18	—	—	0.11	0.13
CT	—	—	0.12	0.15	0.09	0.09
CR	0.2	0.23	0.23	0.27	0.20	0.22
Multiple R	0.61	0.81	0.60	0.80	0.63	0.83
R Shrinkage	0.61	0.81	0.60	0.80	0.63	0.83

*Results are corrected for multivariate range restriction and disattenuated for criterion unreliability.

OPA
OFFICE OF PEOPLE ANALYTICS

Operational vs. NNLS Regression Validity Results

Computational Thinking Score	Validity Estimates			
	Operational Equations Based on Phase 1 Synthetic Validity Study		NNLS Regression Equations Based on Phase 2 Criterion-Related Validity Study	
	Observed	Corrected*	Observed	Corrected*
CompT_AR	0.61	0.81	0.61	0.81
CompT_CT	0.60	0.80	0.60	0.80
CompT_ALL	0.63	0.83	0.63	0.83

*Results are corrected for multivariate range restriction and disattenuated for criterion unreliability.

OPA
OFFICE OF PEOPLE ANALYTICS

Post-Hoc Validity Estimates with All ASVAB Subtests + CT + CR

Computational Thinking Score	Validity Estimates			
	Operational Equations Based on Phase 1 Synthetic Validity Study		NNLS Regression Equations Based on Phase 2 Criterion-Related Validity Study	
	Observed	Corrected*	Observed	Corrected*
CompT_AR	0.61	0.81	0.61	0.81
CompT_CT	0.60	0.80	0.60	0.80
CompT_ALL	0.63	0.83	0.63	0.83
Post-hoc = All ASVAB subtests + CT + CR	--	--	0.67	0.87

*Results are corrected for multivariate range restriction and disattenuated for criterion unreliability. Yellow highlights identify post-hoc results to use for comparison to empirical results for computational thinking scores (same as slide 23).

OPA
OFFICE OF PEOPLE ANALYTICS

Results Conclusion

- All three equation-based scores (CompT_AR, CompT_CT, CompT_ALL) were strong predictors of the computation thinking construct, at least as it was operationalized in the Phase 2 validity study (i.e., CTA-M)
- Empirical weights for the score components (AR, CT, CR) derived from the Phase 2 validity study did not outperform the operational weights derived from the Phase 1 synthetic validity study
- Empirical validity estimates using all ASVAB subtests, CT, and CR resulted in relatively small increases (delta R = 0.04) in prediction in CTA-M scores

Closing

OPA
OFFICE OF PEOPLE ANALYTICS

Software Updates (Completed)

- CR is available for administration on the iCAT platform
- Applicant's completion of CR triggers calculation of CompT scores
  - Requires an AR and/or CT score within the last 2 years
  - Uses most recent AR and/or CT score when multiple records are found
  - Submits a blank score if an eligible AR and/or CT score is not found
- Saves each CompT score within the applicant's CR record
- MEPCOM receives all 4 scores: CR as well as 3 CompT scores

OPA
OFFICE OF PEOPLE ANALYTICS

Questions?

OPA
OFFICE OF PEOPLE ANALYTICS

Response to June 2024 DAC Recommendation

- In process of preparing research designs for CR and CompT that DTAC may consider for future research
  - Applicant data containing one to three of the CompT scores is slowly accumulating, which will support additional analyses
    - Demographic information will likely be available for future subgroup differences research
    - Shippers' occupational training criteria may be useful for future research, should it be made available
    - ASVAB Training Relevance Survey results may be used to identify military occupations with high computational thinking relevance results to further research

OPA
OFFICE OF PEOPLE ANALYTICS

Questions to DAC

- Does the DAC have any suggestions for conducting additional research on fairness issues and/or validity?

OPA
OFFICE OF PEOPLE ANALYTICS

Acknowledgments

- DTAC team
  - Mary Pommerich, Matt Trippe, Liz Waterbury, Greg Manley, Ping Yin
- Phase 1 HumRRO team
  - Scott Oppler, Ted Diaz, Dan Putka, Sam Posnock, Kate Klein, Mike Ingerick, Matt Brown, Sergio Marquez, Sachi Phillips
- Phase 2 HumRRO team
  - Scott Oppler, Robert Wellman, Susan Rowe, Karla Castillo-Guerra, Furong Gao, Jeff Dahlke, Ted Diaz, Rae Powell, Kate Klein, Matt Brown, Evan Good, Cheryl Paullin, Sachi Phillips
- Many, many HumRRO software development and QA team members

OPA
OFFICE OF PEOPLE ANALYTICS

Thank you!

For more information please contact:

Kimberly Adams
kadams@humrro.org
703.236.4303

OPA
OFFICE OF PEOPLE ANALYTICS

Tab L

OPA

OFFICE OF PEOPLE ANALYTICS

Update on Calculator Impact Study

Kevin Bradley

Human Resources Research Organization

Briefing presented to the DACMPT

January 22, 2025

Briefing Agenda

- Overview
- Results
  - Research Question 1: Does calculator availability meaningfully impact the dimensionality of Arithmetic Reasoning (AR) and Mathematics Knowledge (MK) subtests?
  - Research Question 2: Do psychometric properties differ based on calculator availability?
  - Research Question 3: Does calculator availability impact subgroup performance differences?
  - Research Question 4: Does calculator availability impact the amount of time needed to complete each math subtest?
  - Supplemental Analyses
- General Conclusions and Implications
- Questions for the DAC

OPA
OFFICE OF PEOPLE ANALYTICS

Bottom Line Up Front (BLUF)

- Allowing calculators did not meaningfully impact the underlying dimensionality of the AR and MK subtests. Subtests remained predominantly unidimensional in both the No Calculator and Calculator conditions.
- Providing calculators made some AR items easier, resulting in modest but nonnegligible increases in average AR scores (standardized mean difference = 0.37), but had relatively little effect on MK item difficulty or scores (standardized mean difference = 0.07).
  - Statistical equating will be necessary to maintain statutorily-required AFQT qualification rates and would nullify potential mean score increases.
  - Different degrees of calculator sensitivity across items may create complications for adaptive testing.
- Allowing calculators had no notable impacts on measurement properties such as subtest reliability and item discrimination.
- The impact of calculator availability was generally similar across demographic subgroups.
- Providing calculators reduced the average time spent on AR, but had no impact on time spent on MK.
  - The magnitude of this result was generally similar across demographic subgroups.

Overview

OPA
OFFICE OF PEOPLE ANALYTICS

Overview

- Current ASVAB policy is “no calculators”
- Previous research (Buckland et al., 2021) surveyed subject matter experts (SMEs) across the Services about whether servicemembers are required to apply mathematics knowledge and arithmetic reasoning *without having access* to a calculator or other tool
  - 68% of surveyed military SMEs indicated some form of math, without a calculator, is required in training
  - 56% reported that some form of math, without a calculator, is required on the job
  - Thus, Buckland et al. (2021) recommended the “no calculator” policy continue

Overview (cont.)

- Expressed concerns over current policy with respect to calculators
  - Other national testing programs (e.g., ACT, SAT, GED) allow calculators on the quantitative tests
  - Exclusion of calculators may result in the perception that the ASVAB testing program is not keeping up with trends in assessment
  - High school curricula often allow calculators during instruction and exams
  - Test items requiring manual calculations may result in increased test anxiety as students are not accustomed to performing such calculations without a calculator

Overview (cont.)

- Purpose:
  - Empirically evaluate the impact on examinee test performance and the psychometric properties of the AR and MK subtests when calculators are allowed on the MK and AR subtests of the ASVAB
- Study design considerations:
  - Maximize generalizability to ASVAB applicant population
  - Minimize security risks to existing ASVAB item pools
  - Minimize disruptions to operational testing of applicants
  - Minimize strain or burden on study participants

Overview (cont.)

- Participants were similar to those who take the ASVAB under operational testing conditions, with (relatively) recent operational ASVAB scores
- Designed to be as similar as possible to ASVAB operational testing
- Administered in MEPS by Test Administrators/Test Control Officers
- Included post-test survey (contextual information about participants, motivation, calculator usage)
- Shippers completed the study during a waiting period on their ship day
  - 3,042 participants met all screening criteria (sufficient effort and motivation)
  - 2,870 participants met all screening criteria and were unequivocally matched to their official ASVAB administration
  - Demographic makeup of participant sample is provided in Appendix slides

OPA
OFFICE OF PEOPLE ANALYTICS

Overview (cont.)

- All participants completed the same 30-item AR form and 25-item MK form
- Two conditions: calculator provided/calculator not provided
- To avoid intermingling or “cross-condition” exposure, all participants on a given day assigned to the same condition
  - Odd days (11th, 19th, 25th of month) = calculator not provided
  - Even days (12th, 20th, 30th of month) = calculator provided

OPA
OFFICE OF PEOPLE ANALYTICS

Research Question 1: Does calculator availability meaningfully impact the dimensionality of AR and MK subtests?

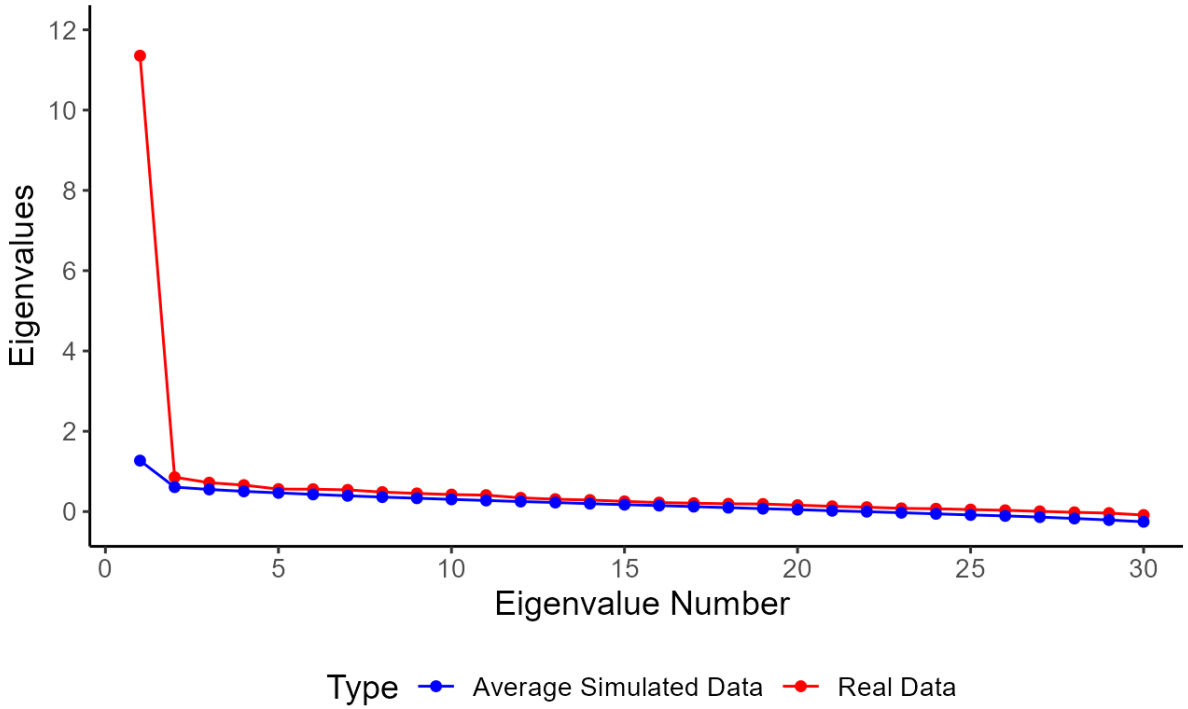
- Parallel analysis
- Bifactor models
- Multiple groups confirmatory factor analysis (CFA)
- Differential functioning of items and tests
- Correlations with other subtest scores

OPA
OFFICE OF PEOPLE ANALYTICS

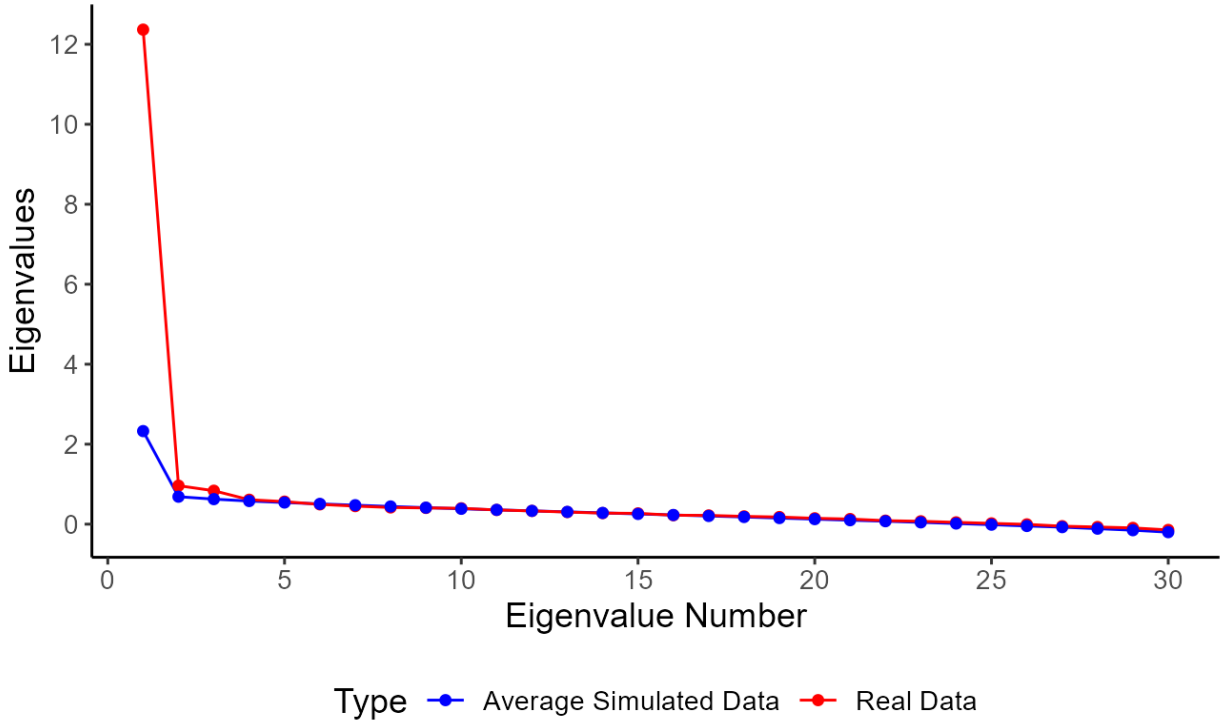
Results

- Parallel analysis results indicate similar AR dimensionality in No Calculator and Calculator conditions

AR No Calculator Condition Scree Plot



AR Calculator Condition Scree Plot

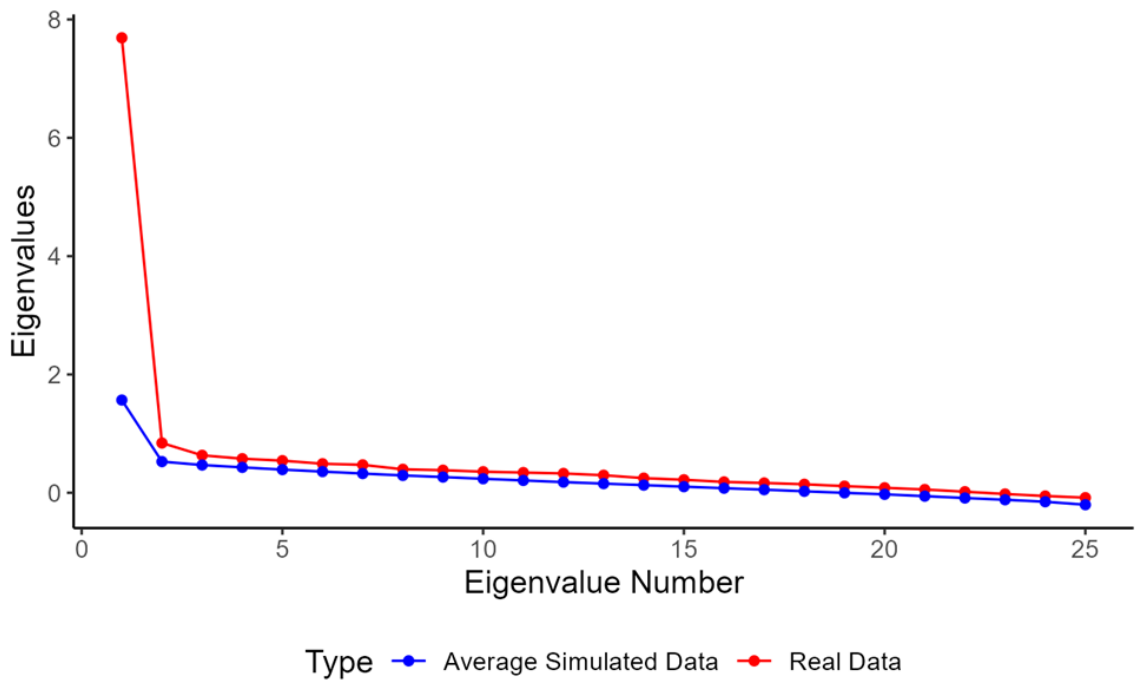


OPA
OFFICE OF PEOPLE ANALYTICS

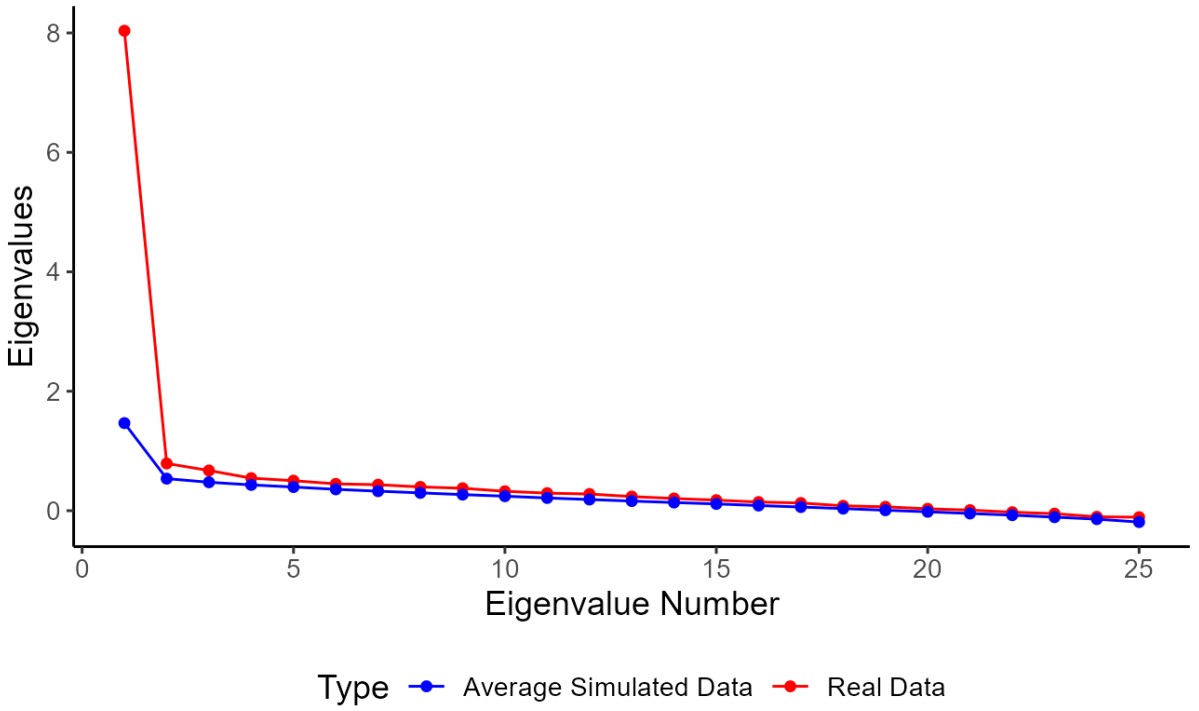
Results (cont.)

- Parallel analysis results indicate similar MK dimensionality in No Calculator and Calculator conditions

MK No Calculator Condition Scree Plot



MK Calculator Condition Scree Plot



OPA
OFFICE OF PEOPLE ANALYTICS

Results (cont.)

- Bifactor model analysis results supported similar dimensionality of AR between conditions, and similar dimensionality of MK between conditions
- Multiple Groups CFA
  - Configural factorial invariance (to test if all items load on a single dimension across groups) was supported for AR and MK
  - Metric (equivalence of factor loadings) invariance partially supported for AR (after removing the equivalence constraints for a subset of items that also demonstrated high non-compensatory differential item functioning [NCDIF] values) and fully supported for MK
  - Scalar (equivalence of intercepts or thresholds) invariance partially supported for AR (after removing the equivalence constraints for a subset of items that also demonstrated high NCDIF values) and MK (after freeing the intercept for one item)

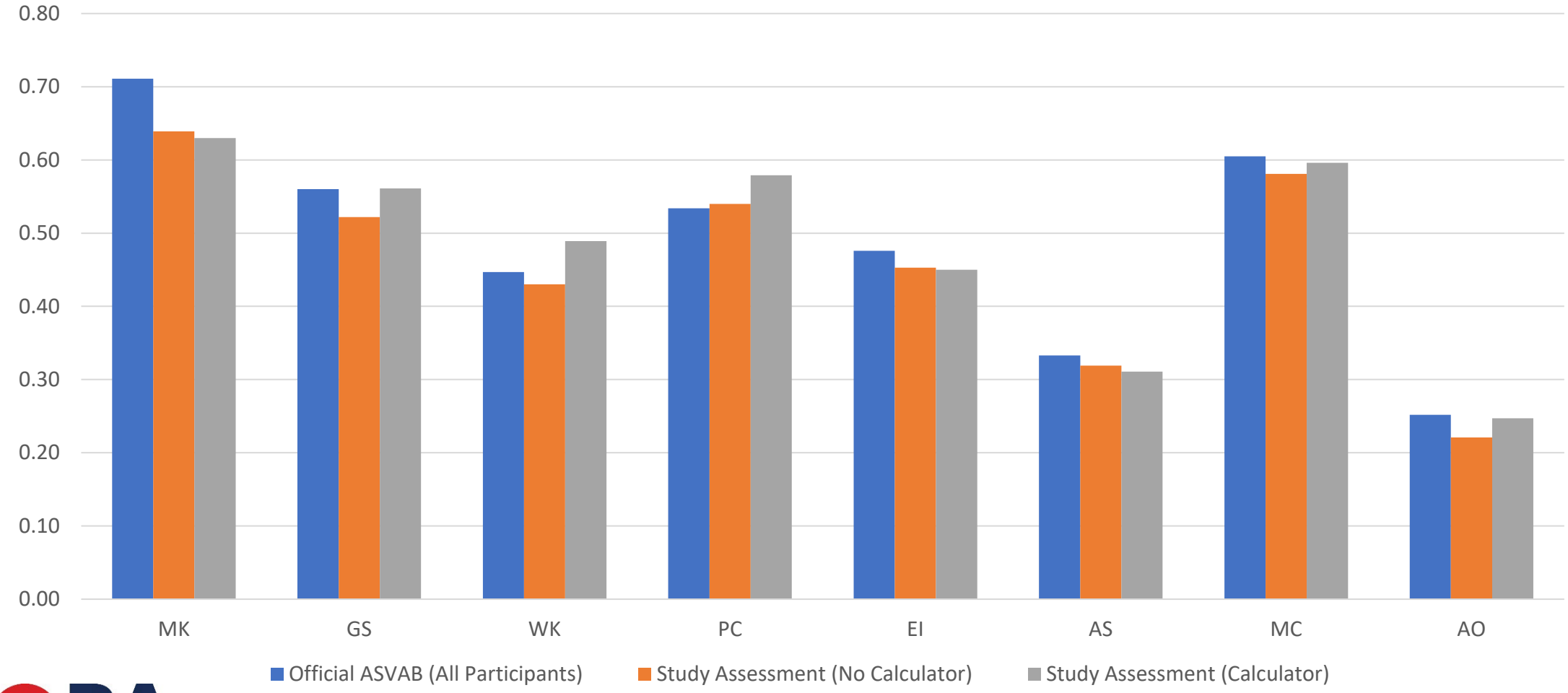
Results (cont.)

- Differential functioning of items and tests (DFIT) test for invariance of item parameters across conditions
  - CDIF & NCDIF
    - 13 AR items and 2 MK items (indicating participants in the Calculator condition more likely to answer correctly)
    - Items exhibiting NCDIF generally exhibited large, positive CDIF values
  - Differential test functioning (DTF)
    - DTF is significant for AR (4.106, compared to significance threshold of .180) but not MK (.068, compared to significance threshold of .150)

OPA
OFFICE OF PEOPLE ANALYTICS

Results (cont.)

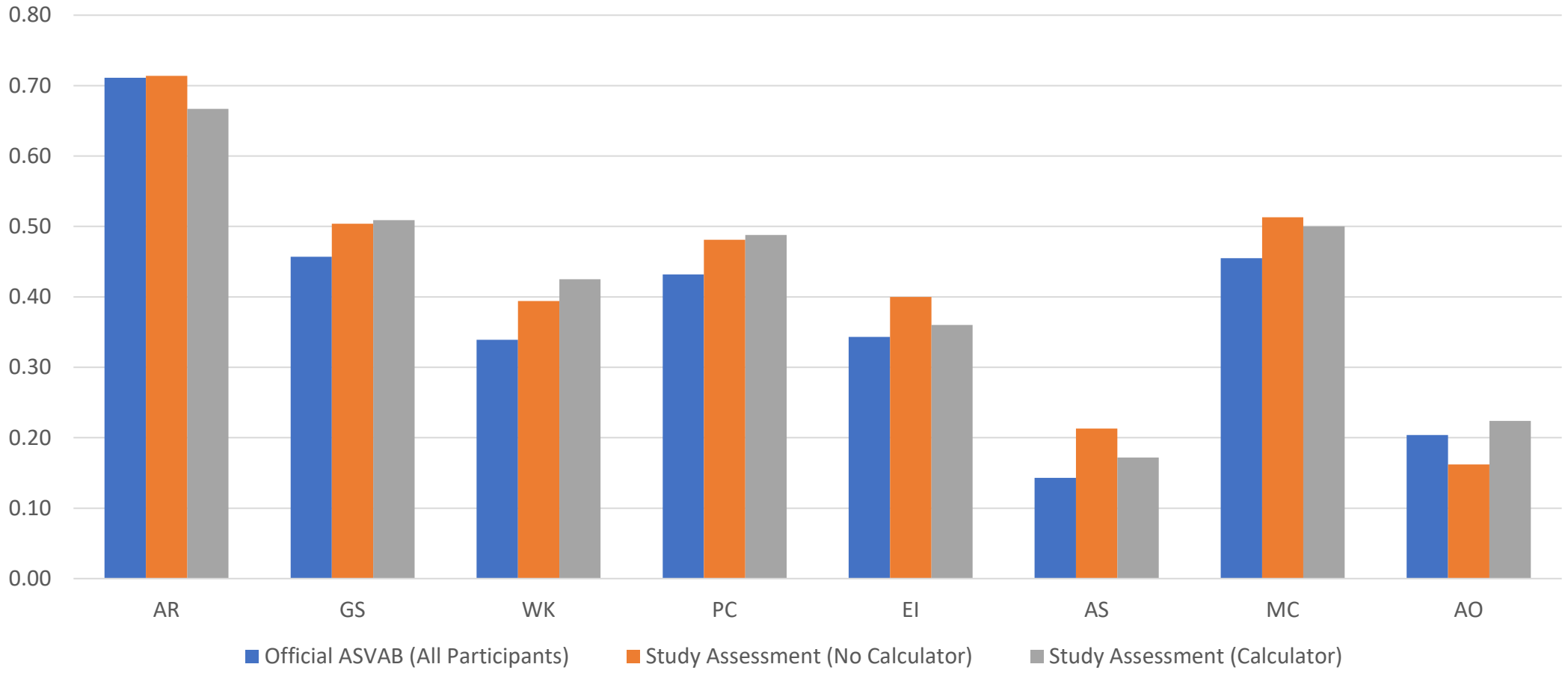
- The pattern and magnitude of AR correlations with other subtest scores are similar for official ASVAB scores, No Calculator condition scores, and Calculator condition scores



OPA
OFFICE OF PEOPLE ANALYTICS

Results (cont.)

- The pattern and magnitude of MK correlations with other subtest scores are similar for official ASVAB scores, No Calculator condition scores, and Calculator condition scores



OPA
OFFICE OF PEOPLE ANALYTICS

Summary and Conclusions for RQ1

- *Does calculator availability meaningfully impact the dimensionality of AR and MK subtests (RQ1)?*
  - Parallel analysis, bifactor models, and CFA results indicate allowing calculators did not meaningfully impact the underlying dimensionality of AR and MK subtests
    - Partial CFA invariance indicates similar factor structure/form across conditions with some items easier in calculator condition
  - DFIT results indicate some items easier in calculator condition
  - Correlations indicate similar patterns across conditions

OPA
OFFICE OF PEOPLE ANALYTICS

Research Question 2: Do psychometric properties differ based on calculator availability?

- Test-level analyses
  - Mean score comparisons
  - Reliability comparisons
  - DTF between conditions
- Item-level analyses
  - Differential item functioning (DIF) between conditions
  - Differences in item statistics between conditions

OPA
OFFICE OF PEOPLE ANALYTICS

Results

- Calculator availability resulted in modest increases in average AR scores but had relatively little effect on MK scores

Subtest	Official Scores					Experimental Scores					Estimated Latent Ability Distributions				
	No Calculator Condition		Calculator Condition		*d*	No Calculator Condition		Calculator Condition		*d*	No Calculator Condition		Calculator Condition		*d*
	M	*SD*	*M*	*SD*		*M*	*SD*	*M*	*SD*		*M*	*SD*	*M*	*SD*	
AR	52.74	8.10	52.26	8.09	−0.06	48.88	9.13	52.26	9.23	0.37	48.55	10.34	52.67	10.56	0.39
MK	53.21	7.06	52.97	6.81	−0.03	49.20	7.04	49.66	7.09	0.07	49.11	7.60	49.64	7.67	0.07

OPA
OFFICE OF PEOPLE ANALYTICS

Results (cont.)

- Allowing calculators had no notable impact on subtest reliability

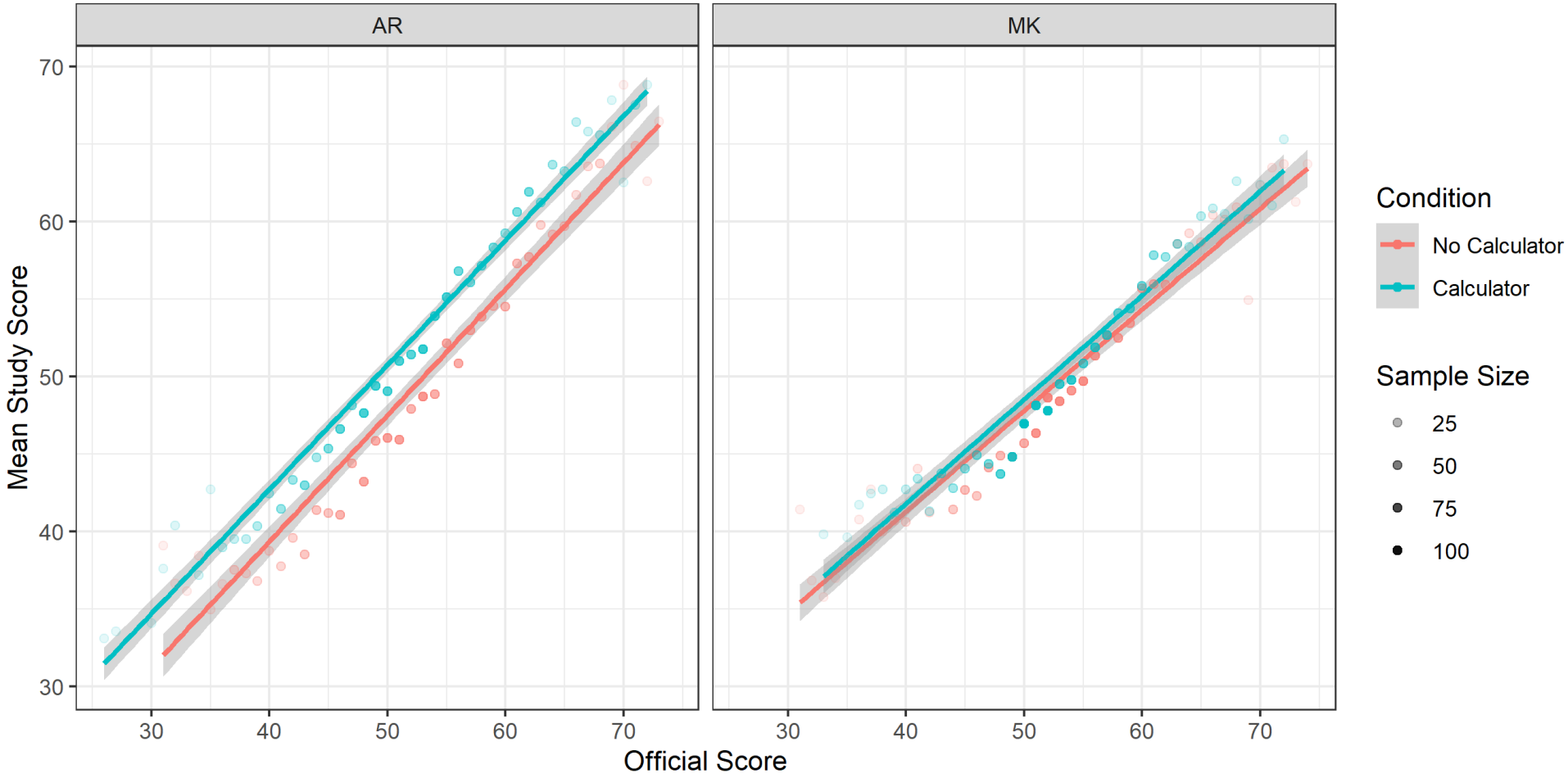
Reliability Type	Method	Subtest	Condition		
			No Calculator Condition	Calculator Condition	Difference
Coefficient alpha	Unstandardized	AR	.895	.902	.007
		MK	.846	.852	.006
	Standardized	AR	.894	.901	.008
		MK	.845	.852	.006
Marginal IRT R_{xx}	Empirical	AR	.858	.864	.006
		MK	.791	.799	.008
	Projected	AR	.873	.868	−.005
		MK	.847	.865	.018

Note: Empirical marginal IRT reliability are based on participants' theta estimates and corresponding standard errors; projected IRT reliability used estimated IRT parameters (no calculator scale) and N(0,1) ability distribution.

OPA
OFFICE OF PEOPLE ANALYTICS

Results (cont.)

Relations between official scores and study scores by condition (DTF between conditions)



Note: For AR, the results indicate that mean study scores (conditional on official scores) are higher for Calculator condition participants compared to No Calculator condition participants. The AR results indicate a simple main effect of calculators on participants' scores. No such pattern was observed for MK, as the conditions' regression lines were not statistically significantly different.

OPA
OFFICE OF PEOPLE ANALYTICS

Results (cont.)

- AR items were generally easier for participants in the Calculator condition (b parameter estimates on the No Calculator scale)
  - Linear equating nullified the mean difficulty differences between conditions
  - Items' a and c parameters were not systematically different between conditions
- MK items were not systematically easier when a calculator was available

3PL Item Parameter	Subtest	No-Calculator Scale				
		No Calculator Condition		Calculator Condition		*d*
		M	*SD*	*M*	*SD*	
a	AR	1.35	0.39	1.43	0.46	0.18
	MK	1.35	0.40	1.46	0.45	0.27
b	AR	0.02	0.47	-0.31	0.53	-0.67
	MK	0.56	0.32	0.53	0.39	-0.10
c	AR	0.22	0.07	0.23	0.06	0.12
	MK	0.22	0.08	0.23	0.10	0.11

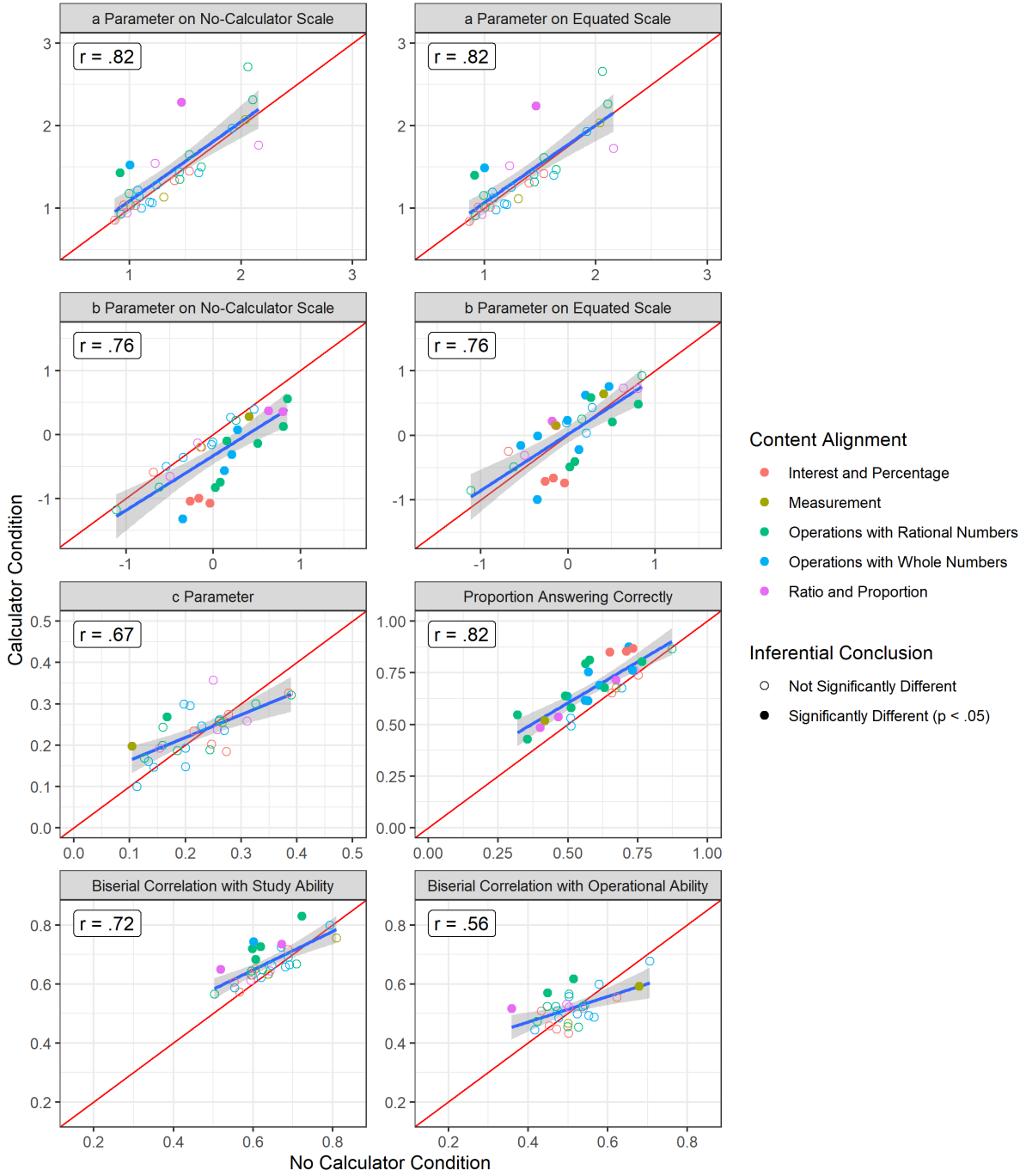
3PL Item Parameter	Subtest	Equated Scale				
		No Calculator Condition		Calculator Condition		*d*
		M	*SD*	*M*	*SD*	
a	AR	1.35	0.39	1.40	0.45	0.11
	MK	1.35	0.40	1.45	0.45	0.24
b	AR	0.02	0.47	0.04	0.54	0.03
	MK	0.56	0.32	0.58	0.39	0.06
c*	AR	0.22	0.07	0.23	0.06	0.12
	MK	0.22	0.08	0.23	0.10	0.11

*c parameter is not transformed

Results (cont.)

- Relations between item-level statistics from study conditions for AR

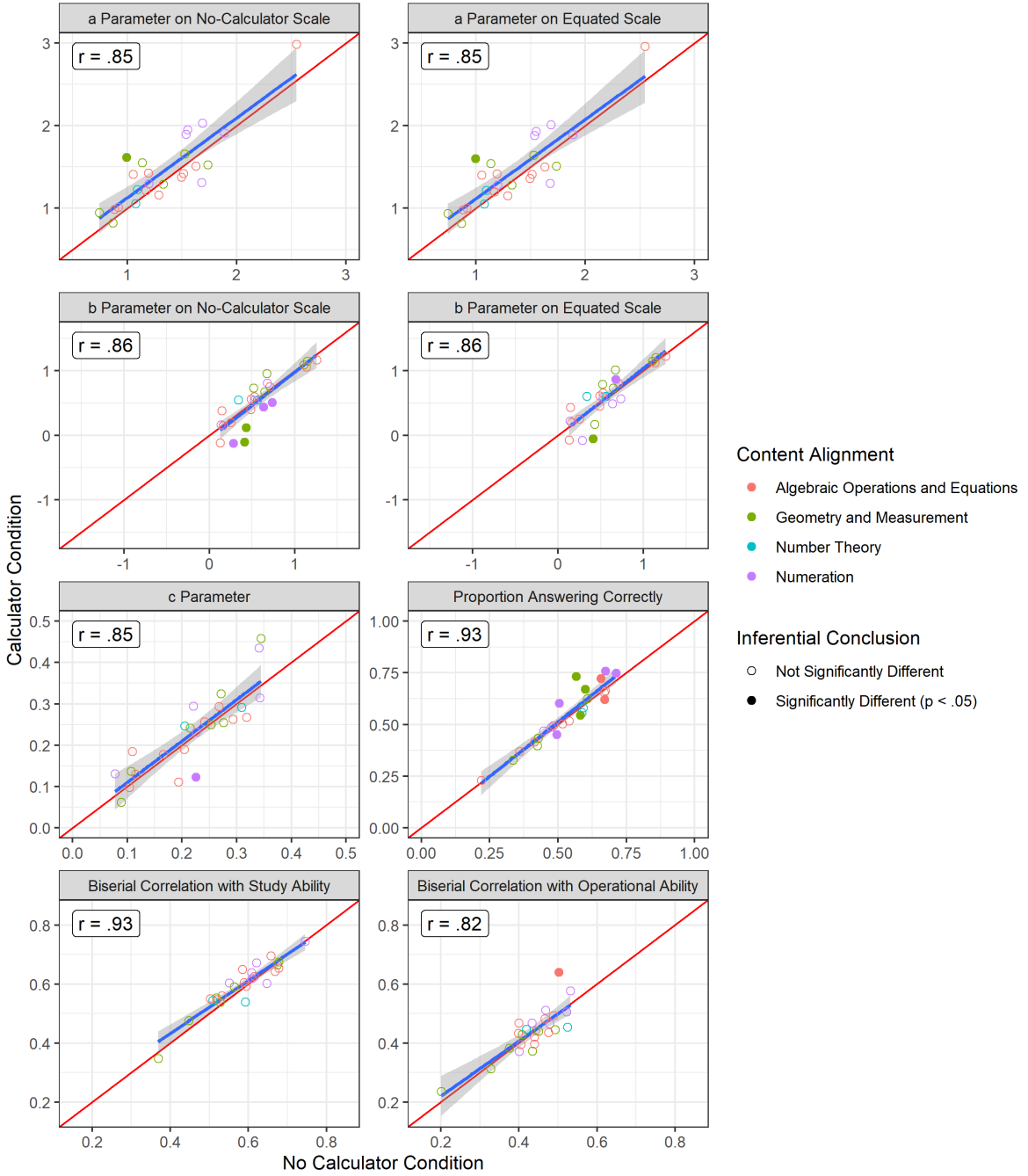
Note: Difficulty parameters and p-values reflect some items were easier in the Calculator condition. All plots reflect a strong positive relationship between item statistics in the No Calculator and Calculator conditions.



Results (cont.)

- Relations between item-level statistics from study conditions for MK

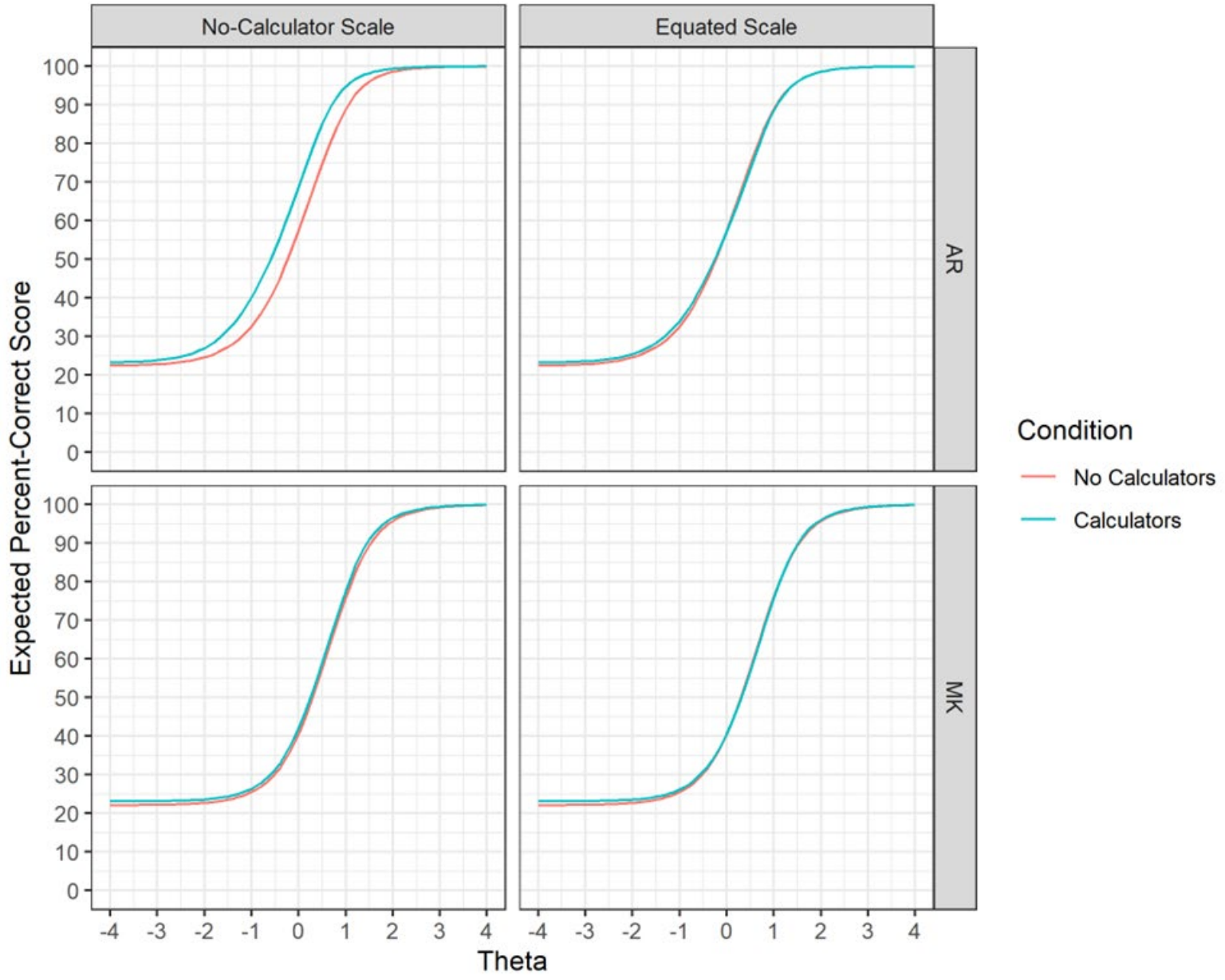
Note: All plots reflect a strong positive relationship between item statistics in the No Calculator and Calculator conditions.



Results (cont.)

- IRT test characteristic curve (TCC) comparisons between conditions

Note: Differences in AR TCCs between conditions were minimal after using linear rescaling to account for the impact of calculators on estimated latent ability distributions.



OPA
OFFICE OF PEOPLE ANALYTICS

Summary & Conclusions for RQ2

- *Do psychometric properties differ based on calculator availability (RQ2)?*
  - Calculators make some AR items easier, but have very little impact on the difficulty of MK items
  - The effects of calculators on scores and item difficulty parameters are primarily linear (after equating, TCCs for No Calculator and Calculator conditions are nearly identical)
    - No Calculator and Calculator conditions could be linked through linear rescaling procedures applied to either scores or item parameters to maintain the interpretability of standard scores and composite scores
      - This finding is likely limited to the individually equated, fixed linear forms used in this study (we do not expect it to generalize to all P&P-ASVAB forms, nor to CAT-ASVAB forms)
      - Even though the mean effects of calculators on item parameters were nullified via IRT equating, there was considerable variance in the differences in AR items' equated b parameters between conditions, and a few items had outlier a parameters in the Calculator condition
      - There was less variance in MK items' parameter differences between conditions, but our DIF analyses showed that a small proportion of MK items are likely to be calculator sensitive
      - This variance in equated item parameters means that a CAT assessment based on equated parameters might encounter inefficiencies due to items' actual parameters differing from the equated parameter estimates
    - Equating would be an essential component of introducing calculators to operational ASVAB testing (to maintain continuity of scores), **resulting in no systematic advantage gained by examinees from using calculators**

Research Question 3: Does calculator availability impact subgroup performance differences?

- Mean score differences across subgroups
- Adverse impact potential by condition
- Within condition DIF analysis

OPA
OFFICE OF PEOPLE ANALYTICS

Results

- The magnitudes of effect sizes between conditions are consistent across subgroups

Subgroup	No Calculator Scores					Calculator Scores					Effect Size	
	n	AR		MK		*n*	AR		MK		AR	MK
		M	*SD*	*M*	*SD*		*M*	*SD*	*M*	*SD*	*d*	*d*
Overall	1,382	48.88	9.13	49.20	7.04	1,488	52.26	9.23	49.66	7.09	.37	.07
Female	158	45.86	7.76	47.89	6.42	169	48.98	9.11	48.80	6.76	.37	.14
Male	1,216	49.23	9.19	49.34	7.11	1,307	52.60	9.21	49.77	7.14	.37	.06
Hispanic White	291	47.65	8.34	47.85	6.87	363	51.18	8.30	48.57	6.78	.43	.11
Non-Hispanic Asian	50	51.30	9.10	52.21	7.32	68	52.89	9.77	51.60	8.77	.17	−.07
Non-Hispanic Black	326	44.81	8.17	47.30	6.26	313	48.34	8.39	47.78	6.55	.43	.07
Non-Hispanic White	629	51.54	9.07	50.78	6.98	652	54.71	9.33	51.15	6.98	.35	.05
English Proficiency: Yes	1,449	49.00	9.15	49.24	7.05	1,517	52.39	9.23	49.74	7.09	.37	.07
English Proficiency: No	28	44.02	7.72	47.43	6.74	30	47.00	8.14	46.27	6.64	.38	−.17

OPA
OFFICE OF PEOPLE ANALYTICS

Results

- Allowing calculators does not appear to alter the potential for adverse impact

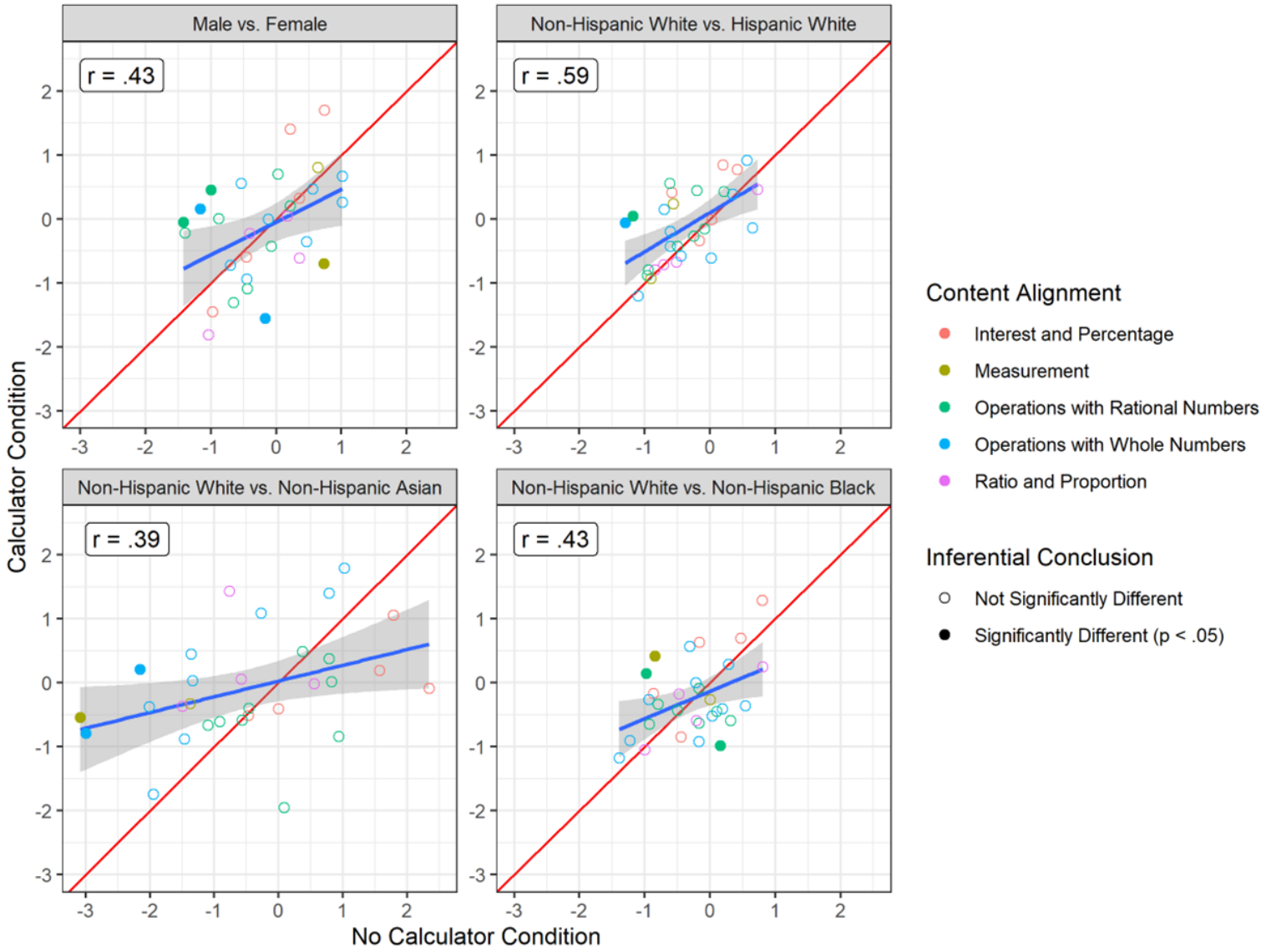
Subtest	Subgroup Contrast	Official Scores			Study Scores		
		d (SE) No Calculators	*d* (SE) Calculators	Difference	*d* (SE) No Calculators	*d* (SE) Calculators	Difference
AR	Male – Female	0.37 (.08)	0.40 (.08)	0.03 (.12)	0.37 (.08)	0.39 (.08)	0.02 (.12)
	Non-Hispanic White – Hispanic White	0.40 (.07)	0.41 (.07)	0.00 (.10)	0.44 (.07)	0.39 (.07)	−0.05 (.10)
	Non-Hispanic White – Non-Hispanic Asian	−0.21 (.15)	0.12 (.15)	0.33 (.19)	0.03 (.15)	0.19 (.15)	0.17 (.19)
	Non-Hispanic White – Non-Hispanic Black	0.83 (.07)	0.75 (.07)	−0.09 (.10)	0.77 (.07)	0.71 (.07)	−0.06 (.10)
	English Proficient – Not English Proficient	0.47 (.19)	0.30 (.19)	−0.17 (.27)	0.55 (.19)	0.58 (.19)	0.04 (.27)
MK	Male – Female	0.07 (.08)	0.01 (.08)	−0.07 (.12)	0.21 (.08)	0.14 (.08)	−0.07 (.12)
	Non-Hispanic White – Hispanic White	0.22 (.07)	0.22 (.07)	−0.00 (.10)	0.42 (.07)	0.37 (.07)	−0.05 (.10)
	Non-Hispanic White – Non-Hispanic Asian	−0.66 (.15)	−0.28 (.15)	0.38 (.20)	−0.20 (.15)	−0.06 (.15)	0.14 (.19)
	Non-Hispanic White – Non-Hispanic Black	0.35 (.07)	0.25 (.07)	−0.10 (.10)	0.52 (.07)	0.49 (.07)	−0.02 (.10)
	English Proficient – Not English Proficient	0.07 (.19)	0.32 (.19)	0.25 (.27)	0.26 (.19)	0.49 (.19)	0.23 (.27)

OPA
OFFICE OF PEOPLE ANALYTICS

Note: *d* indicates standardized mean difference between subgroups.

Results

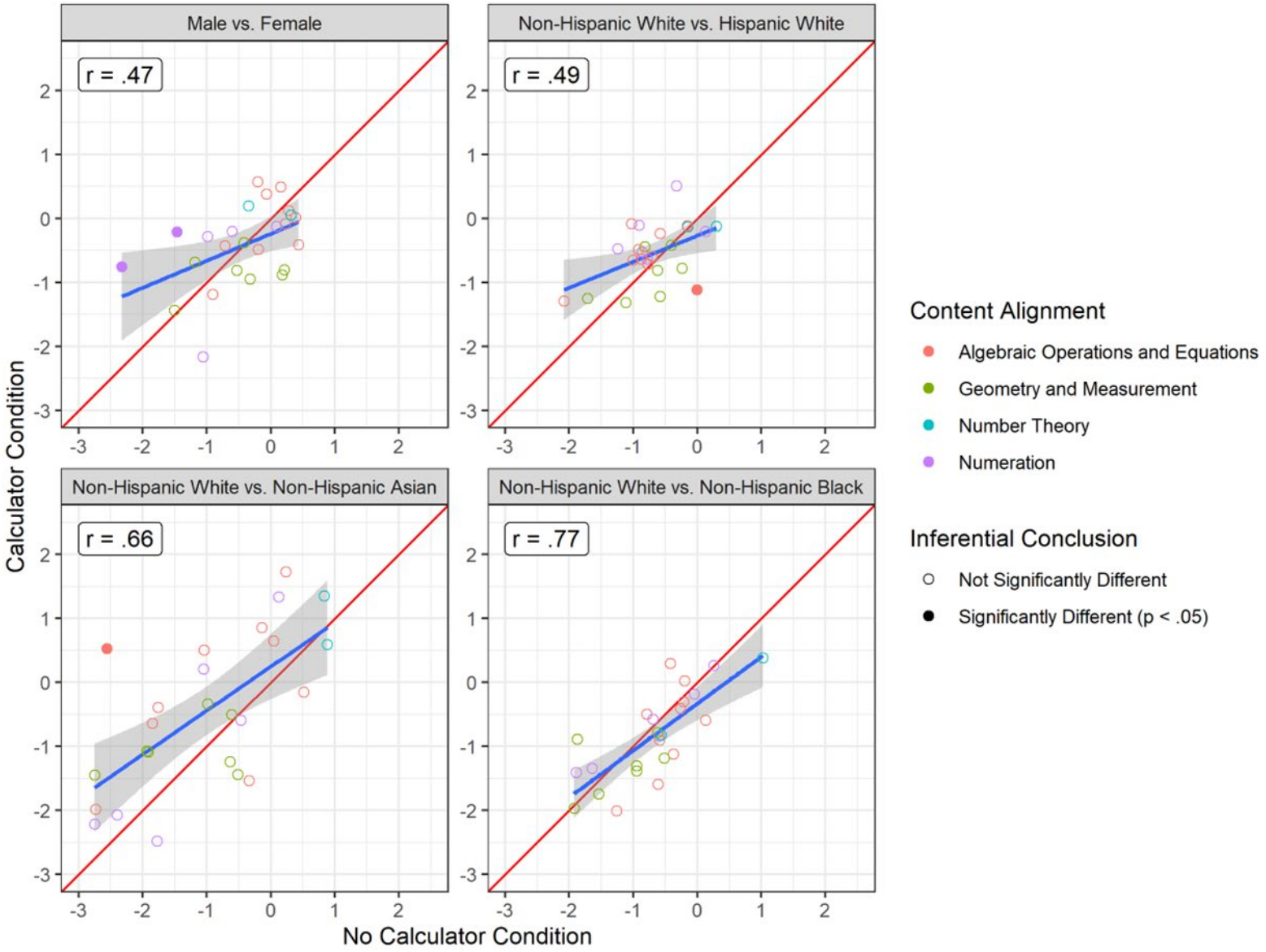
- Within condition DIF analysis (AR), significant differences in DIF between conditions were uncommon across subgroup comparisons



OPA
OFFICE OF PEOPLE ANALYTICS

Results (cont.)

- Within condition DIF analysis (MK), significant differences in DIF between conditions were uncommon across subgroup comparisons



OPA
OFFICE OF PEOPLE ANALYTICS

Summary & Conclusions for RQ3

- Calculators do not appear to differentially impact scores by demographic subgroups
  - Magnitude of between conditions standardized mean difference (*d)* is comparable across subgroups
  - Significant differences in DIF between conditions were uncommon across subgroup contrasts

OPA
OFFICE OF PEOPLE ANALYTICS

Research Question 4: Does calculator availability impact the amount of time needed to complete each math subtest?

- Mean differences in test times between conditions by subgroup

OPA
OFFICE OF PEOPLE ANALYTICS

Results (cont.)

Note: Bold numbers indicate statistically significant differences. Statistically non-significant moderate effect sizes are associated with small sample sizes.

Subgroup	No Calculator				Calculator				Effect Size	
	AR		MK		AR		MK		AR	MK
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*d*	*d*
Overall	31.06	9.86	13.19	5.63	28.26	9.25	13.25	5.66	**–0.29**	0.01
Female	32.05	9.60	13.68	5.97	29.92	9.44	14.60	6.36	**–0.22**	0.15
Male	31.00	9.88	13.16	5.66	28.08	9.16	13.09	5.54	**–0.31**	–0.01
Hispanic White	33.18	9.66	13.78	6.41	30.34	8.84	13.62	5.71	**–0.31**	–0.03
Non-Hispanic Asian	33.56	9.02	13.64	4.83	29.86	9.45	14.98	6.31	**–0.40**	0.23
Non-Hispanic Black	33.72	10.27	14.19	6.23	32.13	9.89	14.62	6.89	**–0.16**	0.07
Non-Hispanic White	28.66	9.18	12.60	4.94	24.97	7.77	12.22	4.62	**–0.43**	–0.08
English Proficiency: Yes	30.99	9.85	13.13	5.58	28.20	9.21	13.21	5.58	**–0.29**	0.01
English Proficiency: No	36.15	9.05	17.17	6.73	32.61	10.63	15.53	8.92	–0.36	–0.21

Summary & Conclusions for RQ4

- Calculators do not appear to differentially impact time spent by demographic subgroups
  - All subgroups completed AR more quickly when a calculator was available; the magnitude of the time spent difference was similar across subgroups
  - The impact of calculator availability on MK time spent was trivial to small for all subgroups

OPA
OFFICE OF PEOPLE ANALYTICS

Supplemental Analyses

- There were trivial to small differences between the No Calculator and Calculator conditions on some of the post-test questions. Participants in the Calculator condition reported feeling slightly more motivated and slightly less anxious than participants in the No Calculator condition.

Post-test Question	No Calculator			Calculator			
	M	*SD*	*n*	*M*	*SD*	*N*	*d*
How anxious were you while taking the two math tests today?	2.55	0.62	1,474	2.62	0.57	1,545	0.12
What was your motivation to answer questions correctly while taking these tests? (analysis sample)	1.44	0.50	1,485	1.39	0.49	1,557	-0.10
What was your motivation to answer questions correctly while taking these tests? (prior to data cleaning)	1.63	0.73	1,737	1.52	0.66	1,727	-0.16

OPA
OFFICE OF PEOPLE ANALYTICS

Note: For the anxiety question, the positive *d* value reflects higher anxiety in the No Calculator condition. For the motivation question, the negative *d* values reflect higher motivation in the Calculator condition.

General Conclusions and Implications

OPA
OFFICE OF PEOPLE ANALYTICS

Summary of Findings

- There is no discernible impact of allowing calculators on the factor structure or dimensionality of AR and MK
  - Parallel analysis, bifactor CFA analysis, and correlation analysis indicate no meaningful dimensionality differences between conditions for AR and MK
  - DTF results indicate some AR items are easier in Calculator condition
- Allowing calculators had no notable impact on item discrimination and subtest reliability

Summary of Findings (cont.)

- Some AR items were easier for participants in the Calculator condition than in the No Calculator condition; overall, AR scores were higher in the Calculator condition
  - Differences in AR test characteristic curves (TCCs) between conditions were minimal after using linear rescaling to account for the impact of calculators (overall impact of calculators on IRT parameters is primarily linear)
  - Scores of examinees who test with a calculator can be linked to the score scale of examinees who test without a calculator with a high degree of accuracy using linear transformations
    - **Note: this finding is likely limited to the specific, fixed linear forms used in this study, and we do not expect it to necessarily generalize to all P&P-ASVAB and CAT-ASVAB forms (see Limitations).**
- MK items tended not to be impacted by allowing calculators; overall, MK scores were not significantly different between conditions

Summary of Findings (cont.)

- The impact of allowing calculators is similar across demographic subgroups
  - Mean differences between the No Calculator and Calculator conditions were comparable across subgroups for both subtests
  - Where there were apparent differences across subgroups in potential performance gains in the Calculator condition, the subgroup sample sizes were small (meaning that sampling error cannot be ruled out as an explanation for the pattern of results observed)
- All subgroups completed AR more quickly when a calculator was available; this difference was statistically significant for all subgroups except non-English proficient participants
  - The numbers of non-English proficient participants were small for both the No Calculator and Calculator conditions, so this finding should be interpreted with caution
- There were no significant mean differences in testing times between conditions for MK

Limitations

- Study included only 30 AR and 25 MK items
  - A very small subset of the total inventory (approximately 10,000) of AR and MK items
  - It is possible the impact of calculators on other fixed-length, linear forms composed of different subsets of AR and MK items could be stronger or weaker than the current results
  - Other subtests that could be affected by calculator use, such as MC and EI, were not included
- Use of a fixed-length, linear form limits our ability to infer impact in CAT-ASVAB administrations, or even on other fixed-length, linear forms that may include a different mix of calculator-sensitive items
  - It seems reasonable to assume there will be a range across examinees in the number of calculator-sensitive items administered (i.e., some examinees might see significantly more calculator-sensitive items than other examinees)

Limitations (cont.)

- Use of a fixed-length, linear form limits our ability to infer impact in CAT-ASVAB administrations, or even on other fixed-length, linear forms that may include a different mix of calculator-sensitive items (cont.)
  - If calculators are permitted on the ASVAB, it will be important to account for the variability in calculator sensitivity across items to minimize the possibility that any given applicant could be advantaged or disadvantaged based on the number of calculator-sensitive items received
  - It would be inappropriate to apply a single scaling constant to all applicants provided with a calculator if some applicants receive fewer calculator-sensitive items than others
- All AR and MK item parameters, regardless of P&P or CAT format, would need to be rescaled based on a linkage of parameter estimates derived from larger samples of both examinees and items
- This rescaling would involve a universal scale transformation for item parameters on all forms, such that all item parameters for a given subtest would be adjusted via the same linear transformation, not form-specific transformations

Limitations (cont.)

- The P&P-ASVAB and CAT-ASVAB could be impacted by this universal rescaling in different ways
  - P&P-ASVAB forms, although psychometrically parallel at the time of their design, may contain different numbers of calculator (in)sensitive items
    - Variation in form-level calculator sensitivity could result in forms producing scores impacted by systematic biases, even after the average effect of calculators is taken into account
    - Forms with more calculator-sensitive items would produce overestimated scores, while forms with fewer calculator-sensitive items would produce underestimated scores
  - CAT-ASVAB forms could also be impacted by residual errors in parameter estimates after item parameters are rescaled, as those errors would impact the efficiency with which the CAT algorithm selects items

Implications if Calculator Use Is Allowed on AR and MK

- Psychometric implications
  - An equating study will be necessary to maintain statutorily-required AFQT qualification rates
    - USC, Title 10, Sec 520, mandates how AFQT is to be applied for the purpose of enlistment (statute mandates a limitation on enlistment of applicants with an AFQT score between 10 and 30)
    - This implies an ability to accurately estimate aptitude—allowing use of calculators on the ASVAB could result in changing the definition of the AFQT scores
  - Calculator use would affect both the CAT and P&P formats and multiple administration purposes (AFCT, PiCAT, VTest, ETP, etc.)
    - Will have implications for score scale as forms are recycled for different purposes
  - Between AR and MK, approximately 10,000 items have been developed, calibrated, and scaled under no-calculator conditions
    - All item parameters will need to be rescaled (a complementary study suggests relying on SME judgments of impact would be insufficient)

Implications if Calculator Use Is Allowed on AR and MK (cont.)

- Psychometric implications (cont.)
  - The linear transformation constants used to convert theta estimates to standard scores are based on linking form-specific score distributions to the 1997 Profile of American Youth (PAY97) norms under no-calculator conditions
    - These constants will need to be adjusted to account for calculator effects on score distributions
  - New specifications for item development would be needed to guide item writing for use on future ASVAB administrations if calculators are allowed
  - A new testing time would need to be determined to account for possible changes in the amount of time needed to complete AR, MK, and the remainder of the ASVAB

Implications if Calculator Use Is Allowed on AR and MK (cont.)

- Psychometric implications (cont.)
  - Even if equated, many uncertainties persist
    - Impact(s) on validity: decades of validity evidence is based on ASVAB administered without the use of calculators
    - There is also a potential concern of accurately assessing the ability of examinees at the high-end of AR achievement
      - Calculators could create a ceiling effect on AR for higher ability applicants such that the AR subtest may no longer be able to accurately measure/assess the ability of examinees at the high end of the ability distribution
    - We have or will have only some knowledge (a snapshot based on 30 AR & 25 MK items) of psychometric impacts on *difficulty*, *dimensionality*, *response time*, *fairness*, *norms*, and *composite cut scores*

Implications if Calculator Use Is Allowed on AR and MK (cont.)

- Logistic considerations
  - Determining when and how to distribute and collect calculators during ASVAB administrations
  - Distributing and maintaining calculators (including for overseas testing)
  - Distributing and transporting calculators for ASVAB CEP administrations
  - Determining who will provide and maintain calculators for each Service for Armed Forces Classification Test (AFCT) administrations
  - Addressing test security concerns associated with monitoring the use of the approved device (including the possibility that individuals might attempt to alter their calculator to use as a recording device)
  - Creating training/guidance for Test Administrators
    - Including guidance on enforcement of approved calculator
    - Determining if/how to prevent calculator use on non-math subtests (e.g., MC, EI)

Implications if Calculator Use Is Allowed on AR and MK (cont.)

- Practical considerations
  - Given the parallelism between conditions' equated TCCs, allowing calculators could put some examinees at a disadvantage if they choose not to make full use of the calculators
    - Choosing not to (consistently) use a calculator could reduce examinees' expected rates of correct responses (but they would be evaluated relative to calculator users)
    - Examinees who prefer not to use a calculator would effectively test under no-calculator conditions but be scored according to calculator-based standards
    - Scores would reflect a function of both math ability and individual differences in calculator use

Thank you!

For more information
please contact:

Kevin Bradley
kbradley@humrro.org
703.706.5647

OPA
OFFICE OF PEOPLE ANALYTICS

Appendix

OPA
OFFICE OF PEOPLE ANALYTICS

Demographic Characteristics of Analysis Sample*

Demographic	No Calculator		Calculator		Total		FY 2023 Applicants/Accessions	
	n	%	*n*	%	*n*	%	*n*	%
Female	158	10.6	169	10.9	327	10.8	80,986	25.1
Male	1,216	81.9	1,307	83.9	2,523	82.9	237,604	73.5
Data Not Available	111	7.5	81	5.2	192	6.3	4,653	1.4
Hispanic White	291	19.6	363	23.3	654	21.5	80,348	24.9
Non-Hispanic Asian	50	3.4	68	4.4	118	3.9	17,406	5.4
Non-Hispanic Black	326	22.0	313	20.1	639	21.0	87,395	27.0
Non-Hispanic White	629	42.4	652	41.9	1,281	42.1	113,921	35.2
Other†	63	4.2	63	4.0	126	4.1	16,317	5.1
Data Not Available	126	8.5	98	6.3	224	7.4	7,856	2.4

OPA
OFFICE OF PEOPLE ANALYTICS

*Shippers from 59 of 65 MEPS participated in the study. Demographic characteristics of sample was similar between conditions.

†Participants who provided ethnicity information and identified as American Indian, Alaska Native, Native Hawaiian, or Other Pacific Islander, and/or identified as Hispanic Black or Hispanic Asian.

Analysis Sample

Service	No Calculator		Calculator		Total		FY 23 Applicants/Accessions	
	n	%	*N*	%	*n*	%	*n*	%
Army	517	34.8	534	34.3	1,051	34.5	165,358	51.2
Air Force	306	20.6	285	18.3	591	19.4	56,736	17.6
Marine Corps	224	15.1	304	19.5	528	17.4	46,935	14.5
Navy	275	18.5	329	21.1	604	19.9	46,199	14.3
Coast Guard	47	3.2	32	2.1	79	2.6	6,679	2.1
Space Force*	13	0.9	0	0.0	13	0.4		
Invalid/Missing	103	6.9	73	4.7	176	5.8	1,336	0.4
Total	1,485	100.0	1,557	100.0	3,042	100.0	323,243	100.0

*Due to Space Force service code not yet being consistently implemented in data system, Space Force applicants are included with Air Force.

Tab M

OPA

OFFICE OF PEOPLE ANALYTICS

Evaluation of Calculator Use on CAT-ASVAB

Glen Heinrich-Wallace

Human Resources Research Organization

Briefing presented to the DACMPT

January 22, 2025

Briefing Agenda

- Define scope and context for study
  - Specify conditions
- Present results
- Pros and cons for calculator use on CAT-ASVAB
- Questions for the DAC

OPA
OFFICE OF PEOPLE ANALYTICS

Background

OPA
OFFICE OF PEOPLE ANALYTICS

Scope of Current Study

- The previous presentation ("Update on Calculator Impact Study;" Bradley, 2025) demonstrates what we might expect to happen with fixed-length, linear forms, but what could happen with CAT-ASVAB remains an open question
- In this study, we aim to evaluate what might happen to CAT-ASVAB composite score distributions after AR and MK item parameters are rescaled to account for the impact of calculators on latent ability distributions
  - Assumption: The results from the Impact Study generalize to CAT-ASVAB
- We used the simulation pipeline infrastructure described in the June 2024 meeting of the DACMPT ("An Evaluation of Calibration Method and Sample Size on the Reliability of New CAT-ASVAB Forms;" Heinrich-Wallace, 2024)
  - This allows us to evaluate consistency between a reference (i.e., unmodified) condition and different experimental conditions

Context for the Current Study: Nature of Available Data

- The only data we have are from the Impact Study
  - These data have a small sample size (30 items for Arithmetic Reasoning [AR], 25 items for Mathematics Knowledge [MK])
    - Under-representation of the universe of items
    - Not all items are expected to have equal calculator sensitivity
    - MK alone has 40+ taxonomies and 200+ identified enemy item groups
  - The Impact Study evaluated fixed-length, linear forms, which are constructed differently from CAT forms
  - CAT, by definition, adaptively selects items from the form and has explicit content balancing for only two subtests (AO, GS)
    - Due to the "greedy" selection algorithm, discrimination plays a larger role than content area in item selection
- This study evaluates what *might* happen after a formal linking study is completed to rescale existing CAT-ASVAB AR and MK item parameters onto a metric that is compatible with calculators *if that study's findings converge with the Impact Study*

Simulating Empirically-based Error

- Because of the characteristics of the Impact Study data, instead of focusing on a single condition, we evaluate a range of counterfactuals, each of which answers what we can expect would happen if different types of error were introduced
- To generalize from the available data, we fit a 3D Gaussian copula to the Impact Study's item parameter data and sampled values from the copula; specifically, we:
  - Converted *a* and *c* parameters to the normal metric for 1) the Impact Study data and 2) the generating parameters used in the simulation pipeline
  - Fit the copula to residuals between without-calculator parameters and equated with-calculator parameters from the Impact Study
  - Added these residuals to the transformed generating parameters
  - Transformed the altered *a* and *c* parameters back to their natural metrics
  - Estimated new composites for the holdout sample from Heinrich-Wallace (2024)
- Several conditions modify the *b* parameters deflections to address plausible scenarios for how the universe of items may differ from our sample in terms of calculator sensitivity

Research Questions

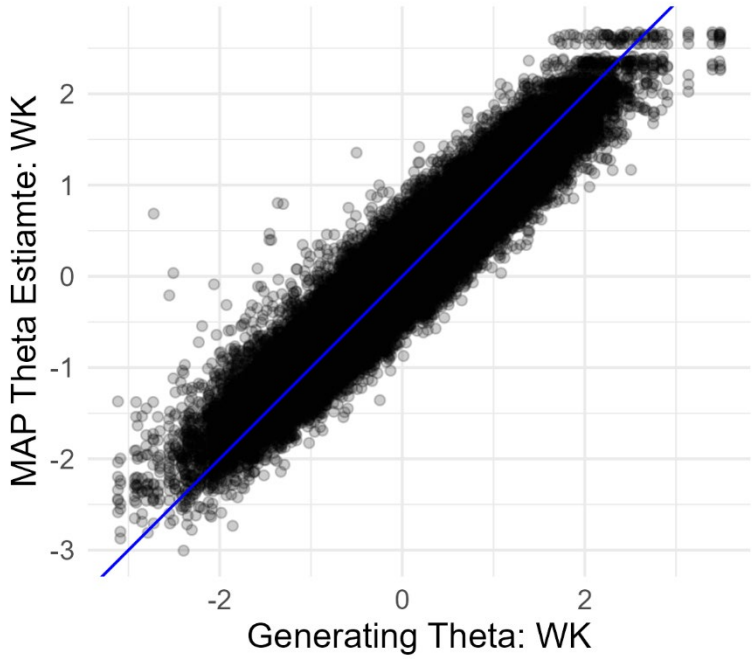
- How do empirically informed, copula-based deflections to item parameter estimates affect composite score distributions for CAT-ASVAB?
- How do biased difficulty parameter deflections affect composite score distributions for CAT-ASVAB?
- If effects are present, which composites and which ranges of those score distribution are most affected?

Bottom Line Up Front (BLUF)

- Across all conditions, measurement precision decreases relative to the test/retest baseline
  - This is expected because all manipulations introduce additional error into the parameter estimates, which increases measurement error and decreases precision
- In general, there is more measurement error for higher-ability simulees, and these simulees are more likely to be under-classified
- Because the classification composites for different Services place different weights on AR and MK, the impact of calculator use on composite precision varies across Services

Conditions (Part 1)

- Test (Condition 0)
  - Consists of running the final stage of the simulation pipeline from Heinrich-Wallace (2024) to compute composite scores for the holdout sample; we evaluate 10 replications (700,000 cases per composite per condition)
- All other conditions are evaluated relative to the test condition
  - This is conceptually similar to decision consistency (comparing two estimated scores)
  - In this case, decision consistency is preferable to decision accuracy (comparing an estimated and a generating score) because all composites are based on Bayesian modal estimate theta-hats, which are subject to shrinkage



OPA
OFFICE OF PEOPLE ANALYTICS

Conditions (Part 2)

- Retest (Condition 1)
  - The same as the Test condition, but with a different random seed
- Random Error (Condition 2)
  - *a*, *b*, and *c* parameters have copula-based deflections based on the Impact Study data
- Alternating Tail-Sampled Error (Conditions 3–7)
  - *a* and *c* parameter deflections are the same as Condition 2, but the *b* parameter deflections are sampled from the top and bottom 5% of copula-based deflections
  - Different proportions of items (3/15, 6/15, 9/15, 12/15, and 15/15) have the manipulation while the remaining items have no manipulation
  - These conditions evaluate counterfactuals where different proportions of items have higher or lower sensitivity to calculators than the average items included in the Impact Study

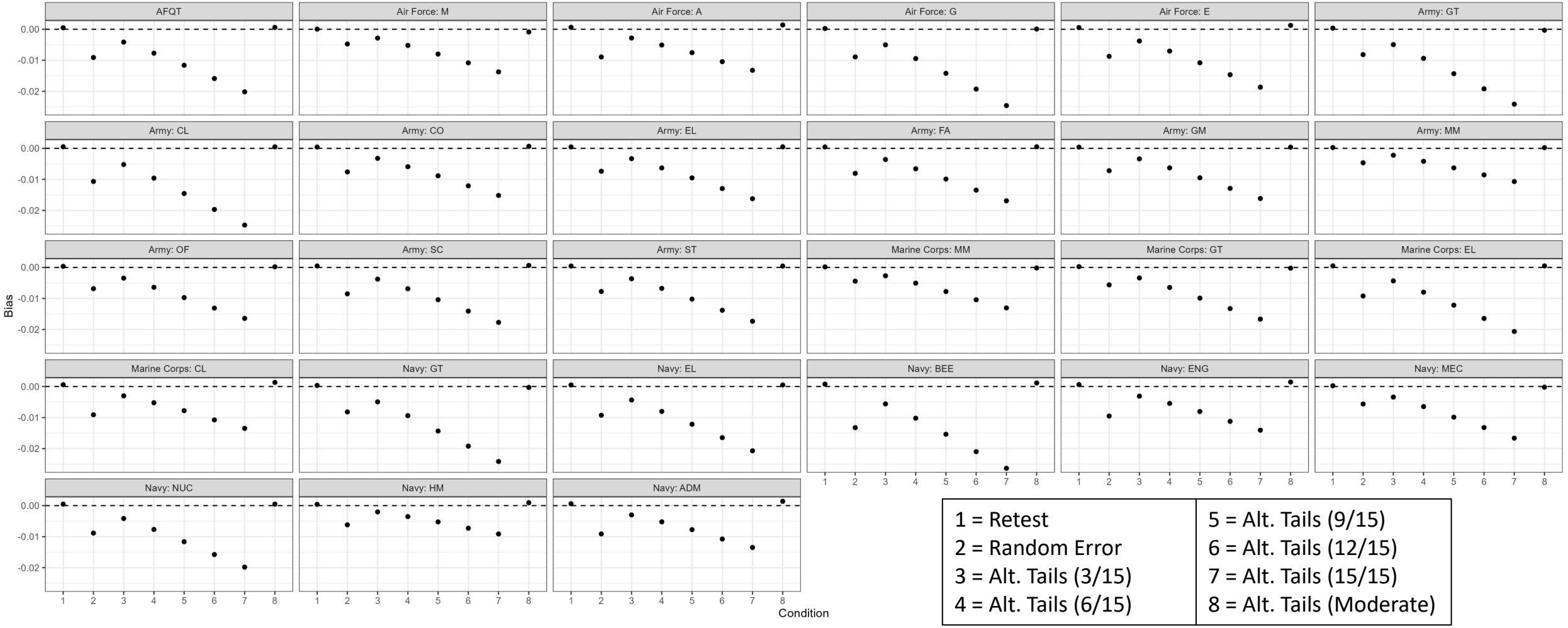
Conditions (Part 3)

- Alternating Tail-Sampled Error, Moderate (Condition 8)
  - Same as Condition 7 (15/15) but all *b* parameter deflections are halved
  - Assesses the same counterfactual as Condition 7 (15/15 items are manipulated), but items varied less in their calculator sensitivity
- Systematic Error in *b* Parameters (Condition 9)
  - Shows the effect of systematic error on composite scores
    - The largest simulated deflection for *b* parameters is added (which was negative) to the difficulty of each item, indicative of an item that is more calculator sensitive than the average error from the Impact Study sample of items
  - Emphasizes the importance of equating (which removes systematic error)
  - Proof of concept that the pipeline is working properly
    - We can simulate extreme results

Results

OPA
OFFICE OF PEOPLE ANALYTICS

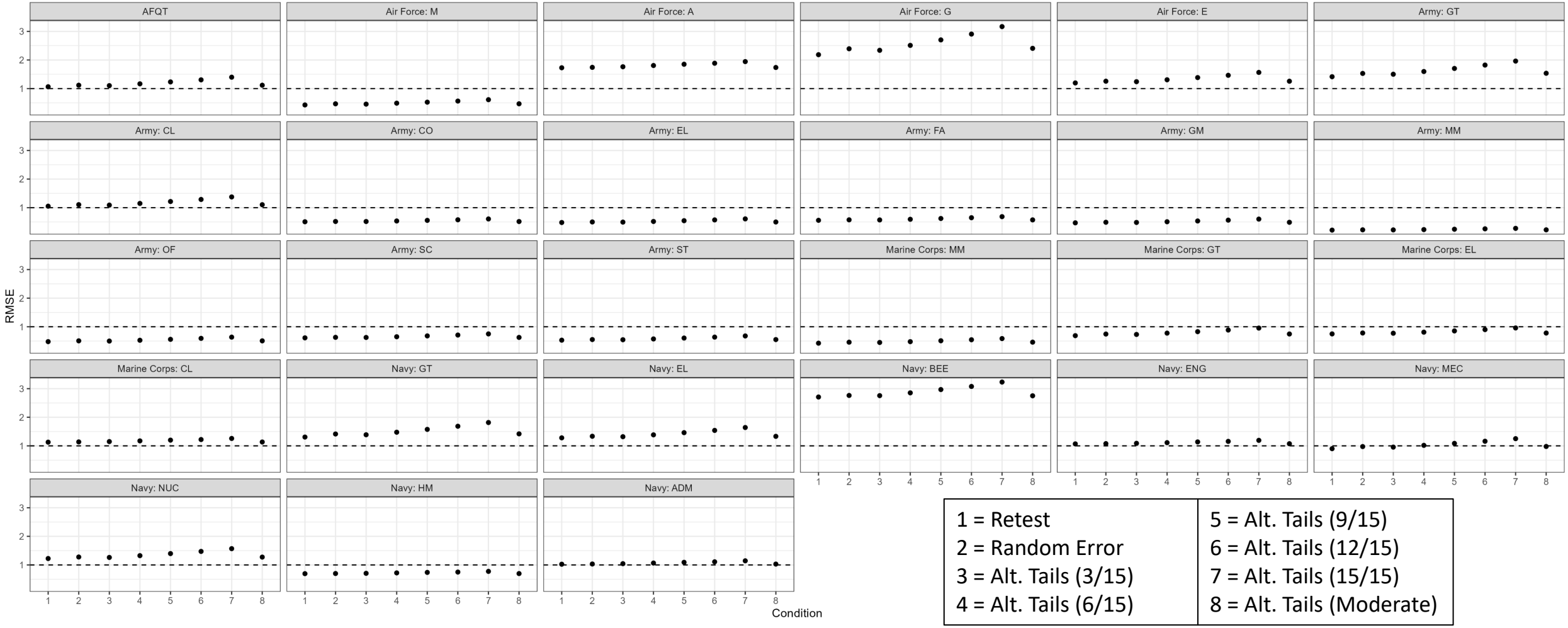
Bias per Composite and Condition



1 = Retest	5 = Alt. Tails (9/15)
2 = Random Error	6 = Alt. Tails (12/15)
3 = Alt. Tails (3/15)	7 = Alt. Tails (15/15)
4 = Alt. Tails (6/15)	8 = Alt. Tails (Moderate)

OPA
OFFICE OF PEOPLE ANALYTICS

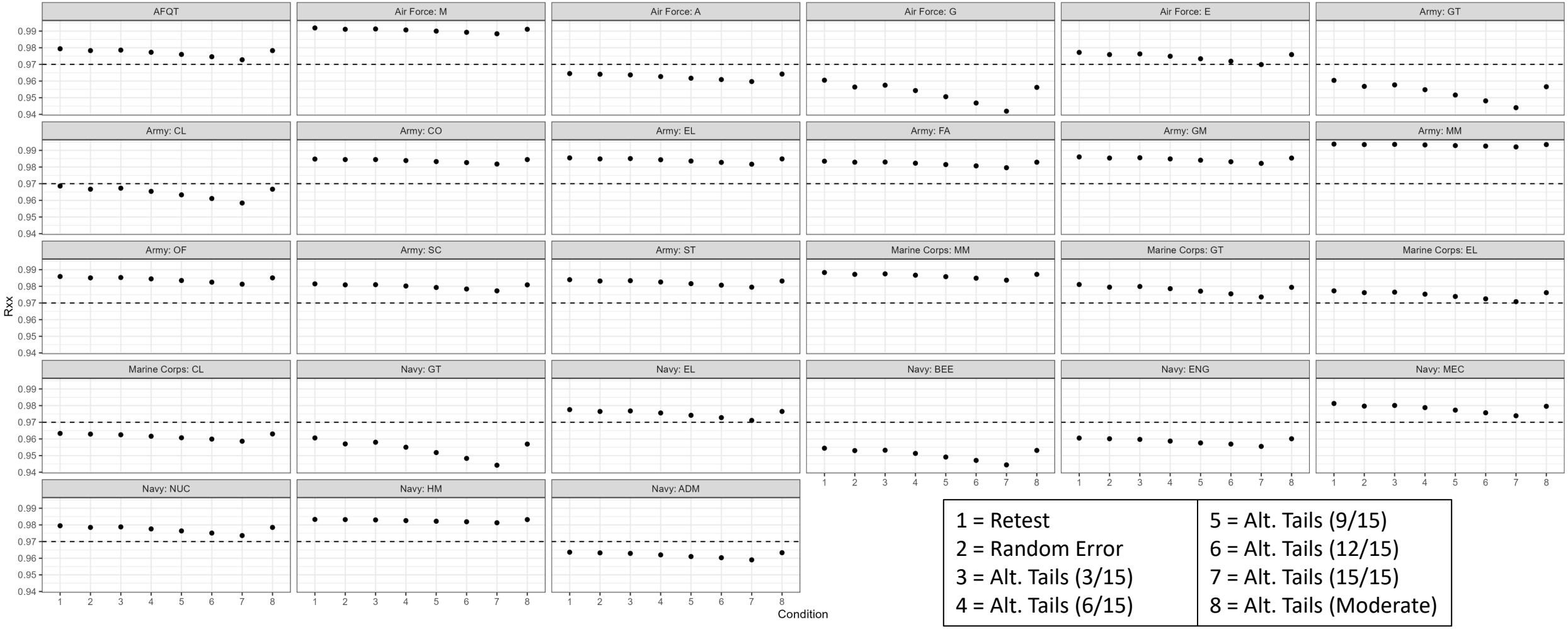
Root Mean Square Error (RMSE) per Composite and Condition



1 = Retest	5 = Alt. Tails (9/15)
2 = Random Error	6 = Alt. Tails (12/15)
3 = Alt. Tails (3/15)	7 = Alt. Tails (15/15)
4 = Alt. Tails (6/15)	8 = Alt. Tails (Moderate)

OPA
OFFICE OF PEOPLE ANALYTICS

R_{xx} per Composite Between Test and Focal Condition



1 = Retest	5 = Alt. Tails (9/15)
2 = Random Error	6 = Alt. Tails (12/15)
3 = Alt. Tails (3/15)	7 = Alt. Tails (15/15)
4 = Alt. Tails (6/15)	8 = Alt. Tails (Moderate)

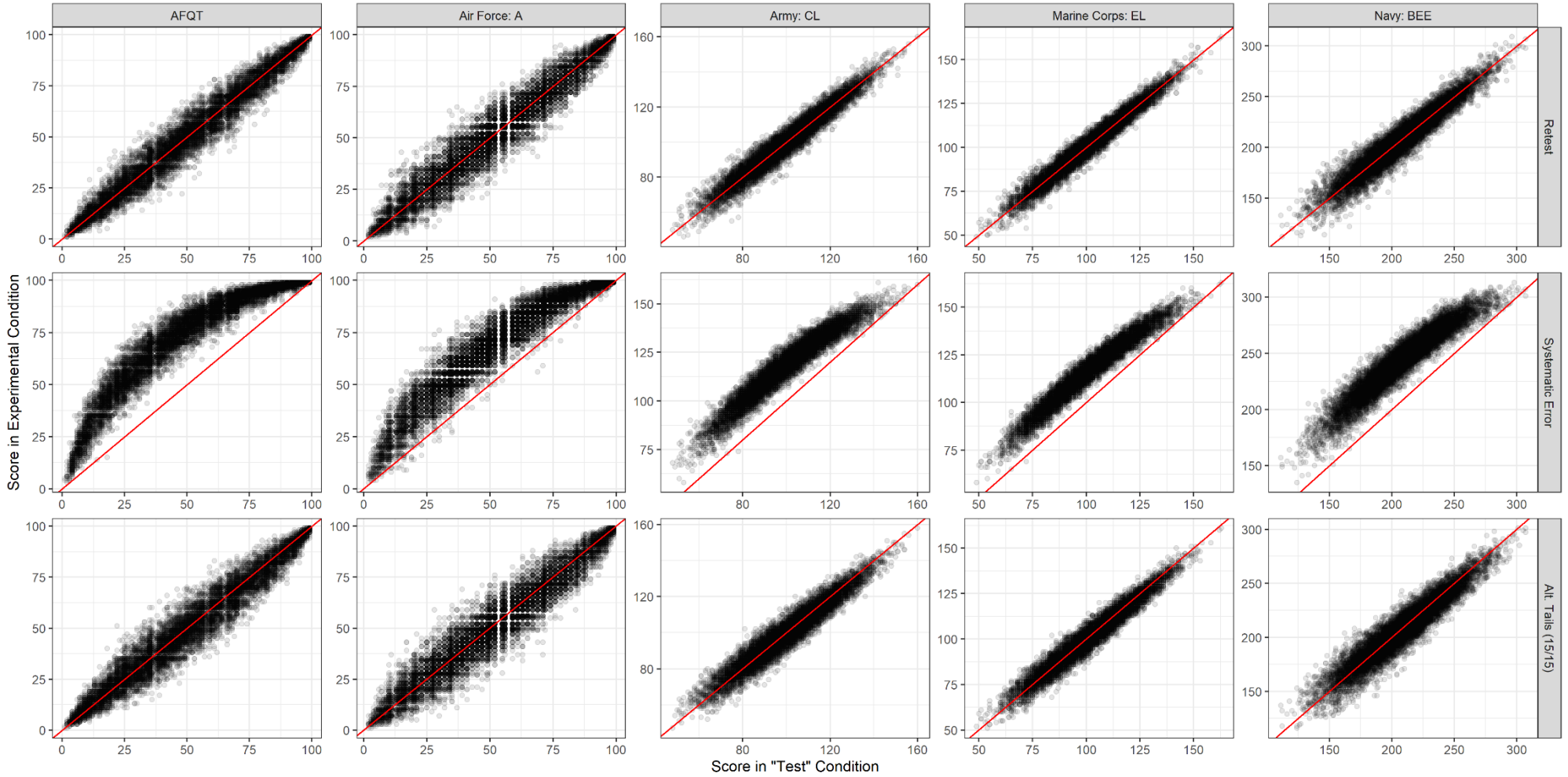
OPA
OFFICE OF PEOPLE ANALYTICS

Proportion of Total Composite Weight Attributable to AR + MK

Service	Composite	Proportion AR+MK
AFQT	AFQT	0.50
Air Force	A	0.50
	G	0.50
	E	0.50
	M	0.20
Army	CL	0.59
	GT	0.50
	SC	0.44
	FA	0.42
	ST	0.40
	CO	0.39
	EL	0.38
	GM	0.37
	OF	0.36
	MM	0.25

Service	Composite	Proportion AR+MK
Marine Corps	EL	0.50
	CL	0.50
	GT	0.33
	MM	0.25
Navy	BEE	0.75
	GT	0.50
	EL	0.50
	ENG	0.50
	NUC	0.50
	ADM	0.50
	MEC	0.33
	HM	0.33

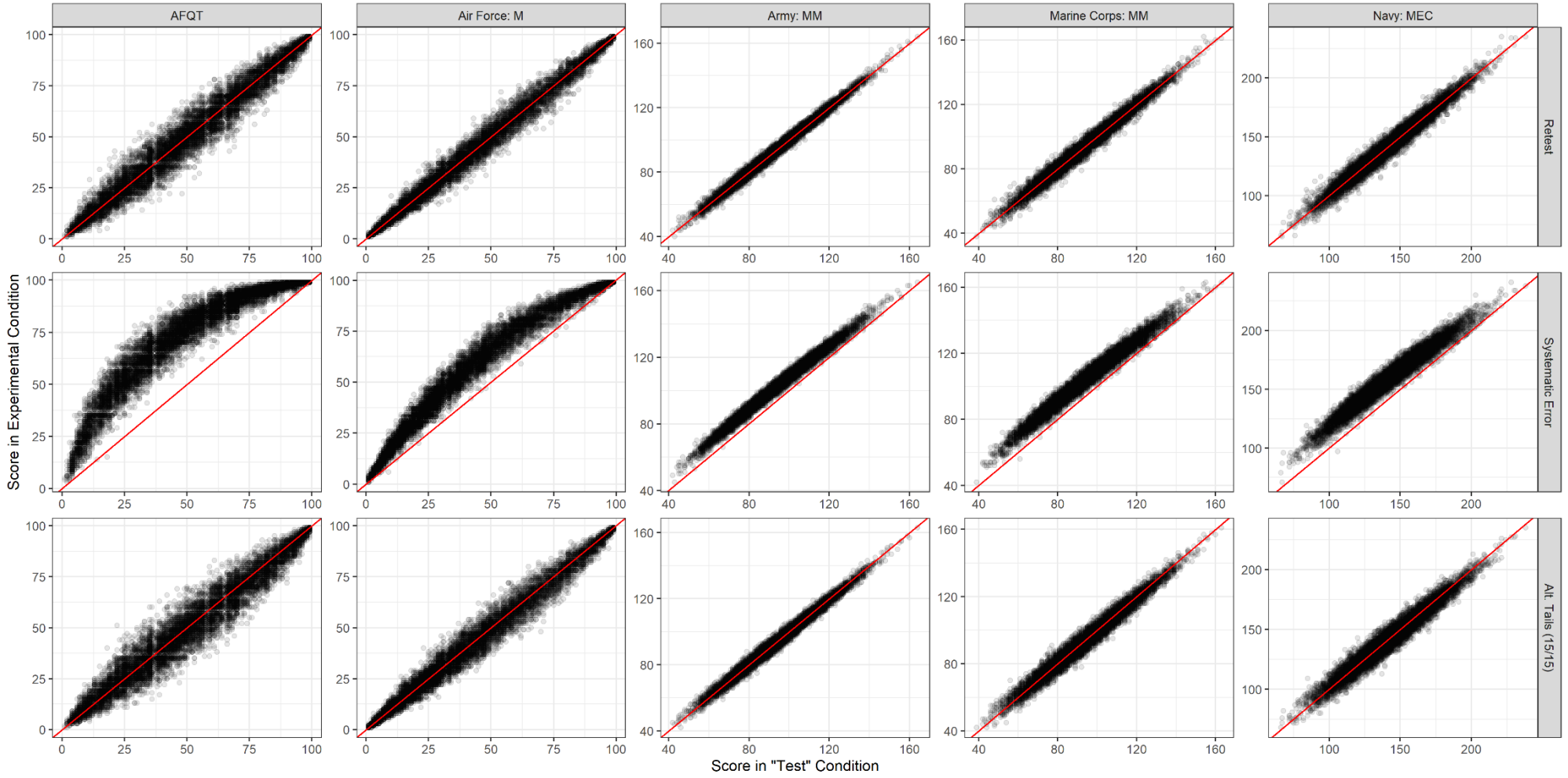
Scatterplots of Experimental Conditions vs. Test Condition for AFQT and the Most Math-Heavy Composites per Service



OPA
OFFICE OF PEOPLE ANALYTICS

Note: Each plot shows a random sample of 10,000 cases

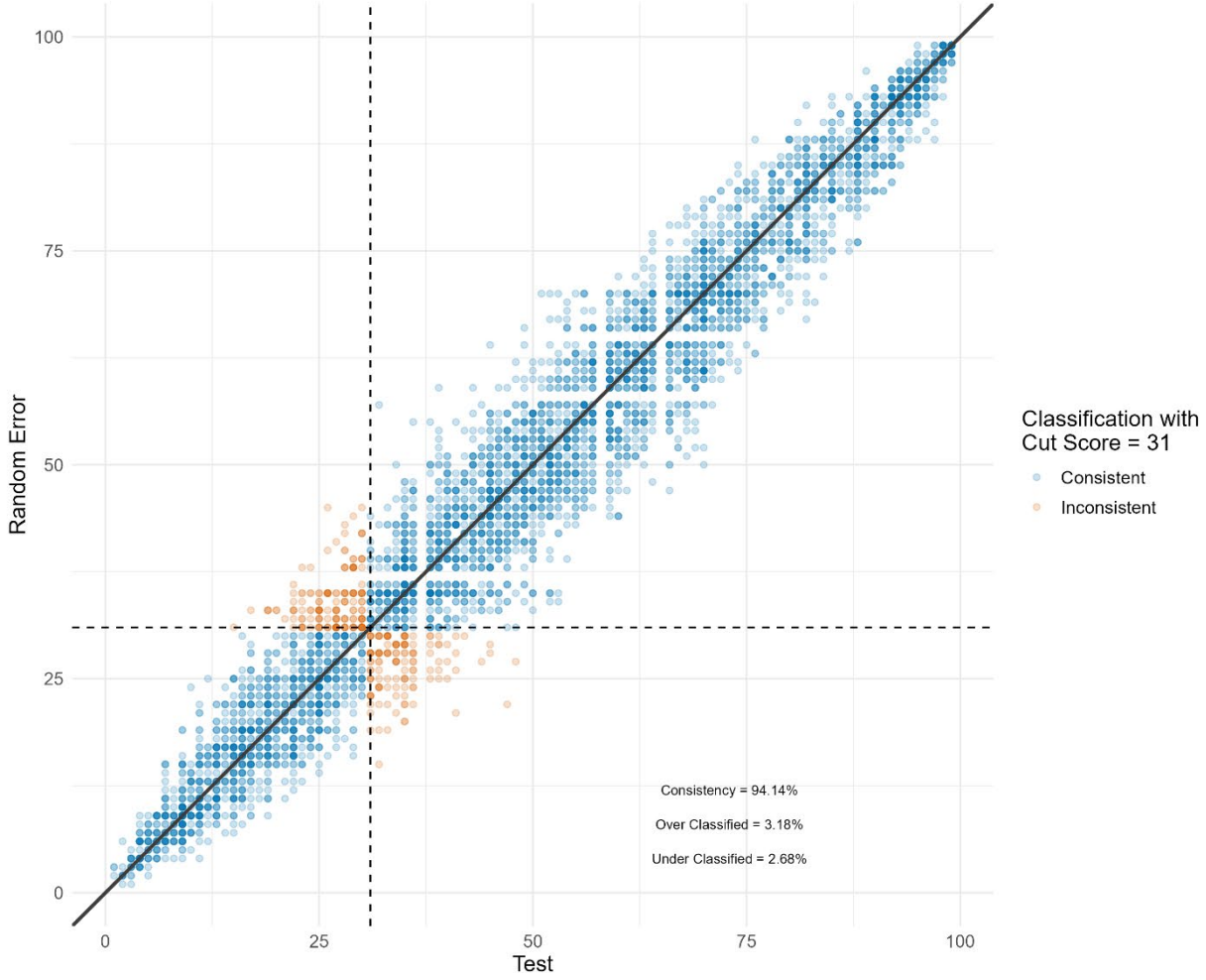
Scatterplots of Experimental Conditions vs. Test Condition for AFQT and the Least Math-Heavy Composites per Service

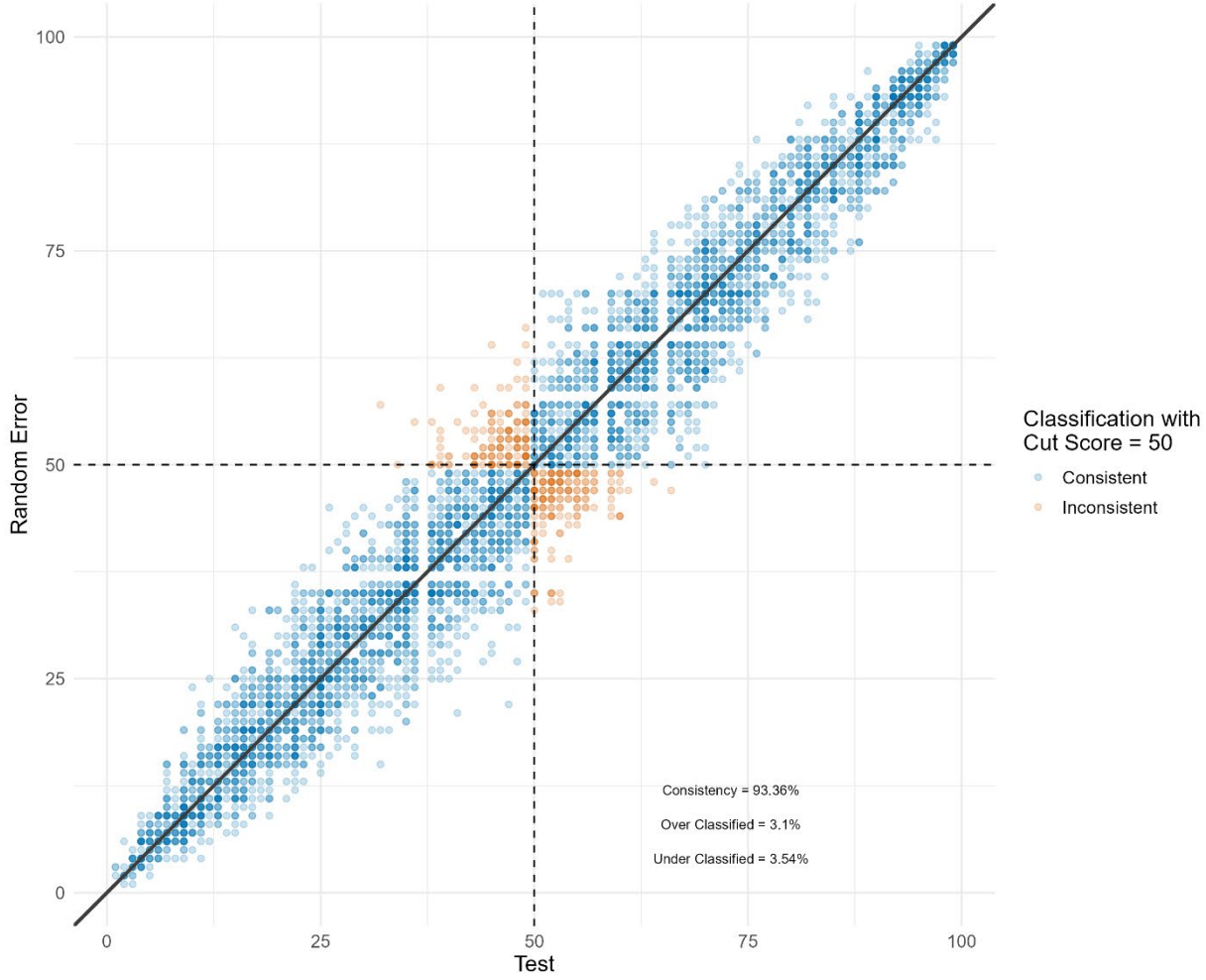


OPA
OFFICE OF PEOPLE ANALYTICS

Note: Each plot shows a random sample of 10,000 cases

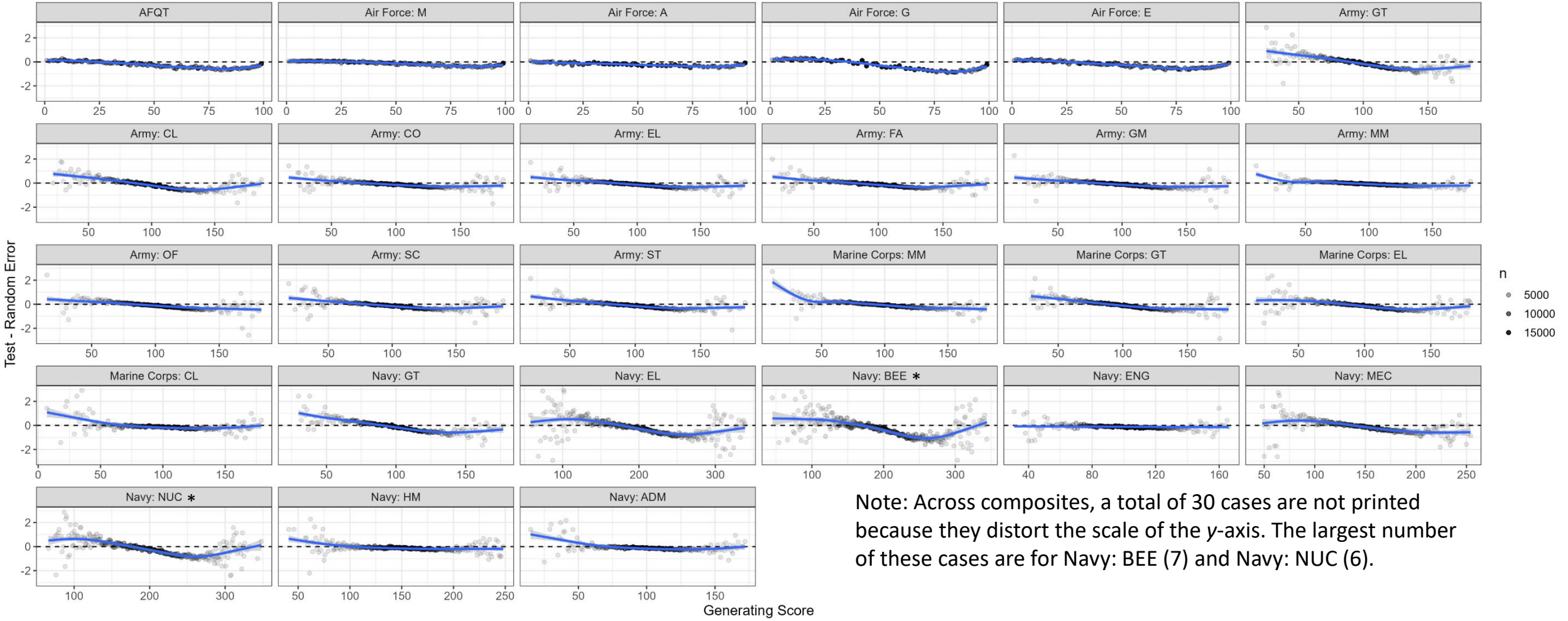
AFQT Classification for Test/Random Error for Two Cut Scores





OPA
OFFICE OF PEOPLE ANALYTICS

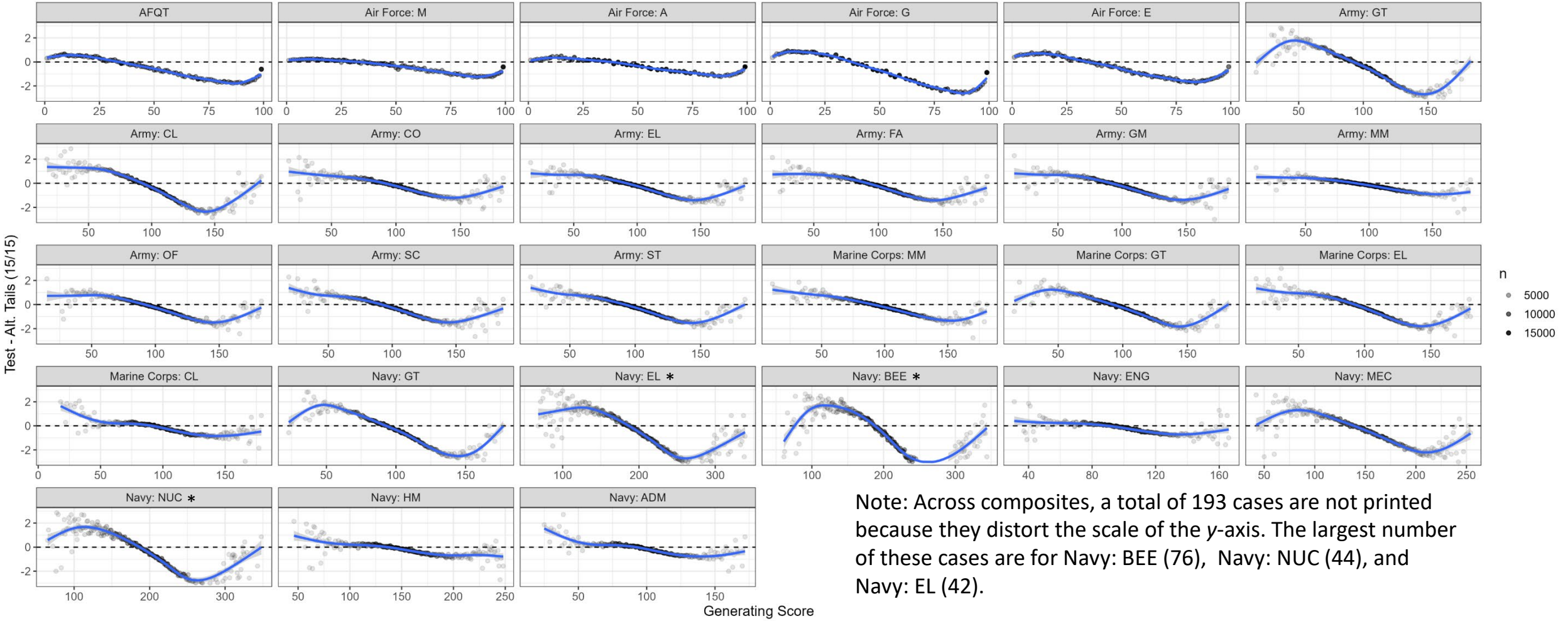
Mean Score Conditional Bias for All Composites: Random Error



Note: Across composites, a total of 30 cases are not printed because they distort the scale of the *y*-axis. The largest number of these cases are for Navy: BEE (7) and Navy: NUC (6).

OPA
OFFICE OF PEOPLE ANALYTICS

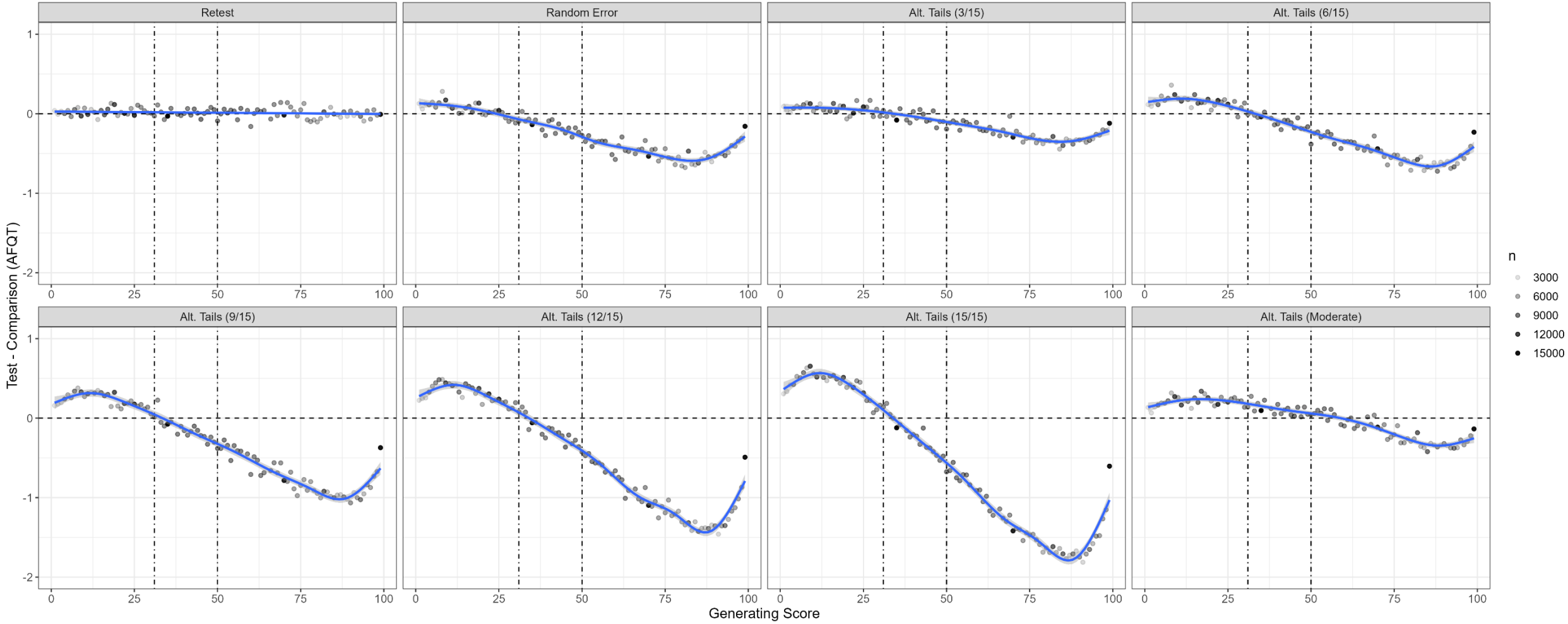
Mean Score Conditional Bias for All Composites: Alternating Tails 15/15



Note: Across composites, a total of 193 cases are not printed because they distort the scale of the *y*-axis. The largest number of these cases are for Navy: BEE (76), Navy: NUC (44), and Navy: EL (42).

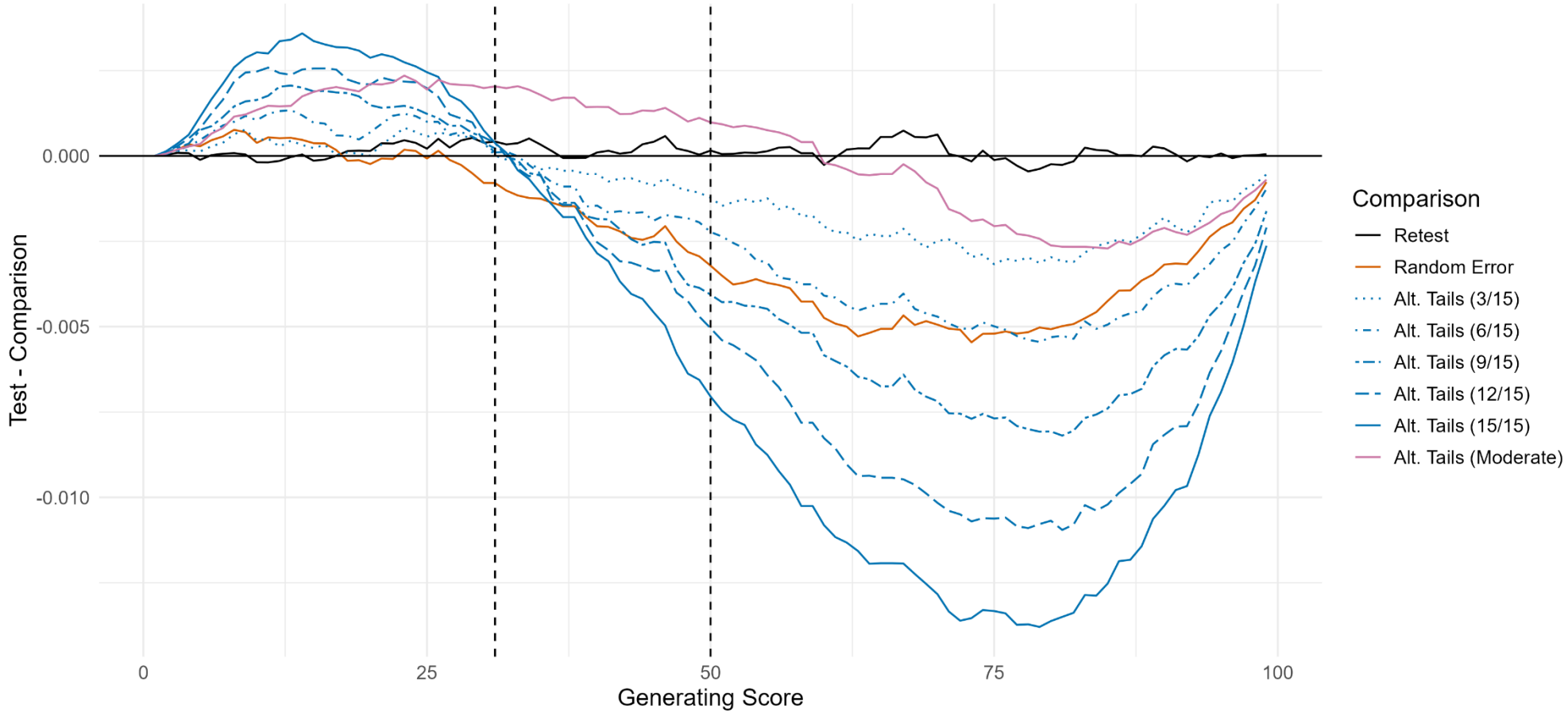
OPA
OFFICE OF PEOPLE ANALYTICS

Mean Score Conditional Bias for AFQT across Conditions



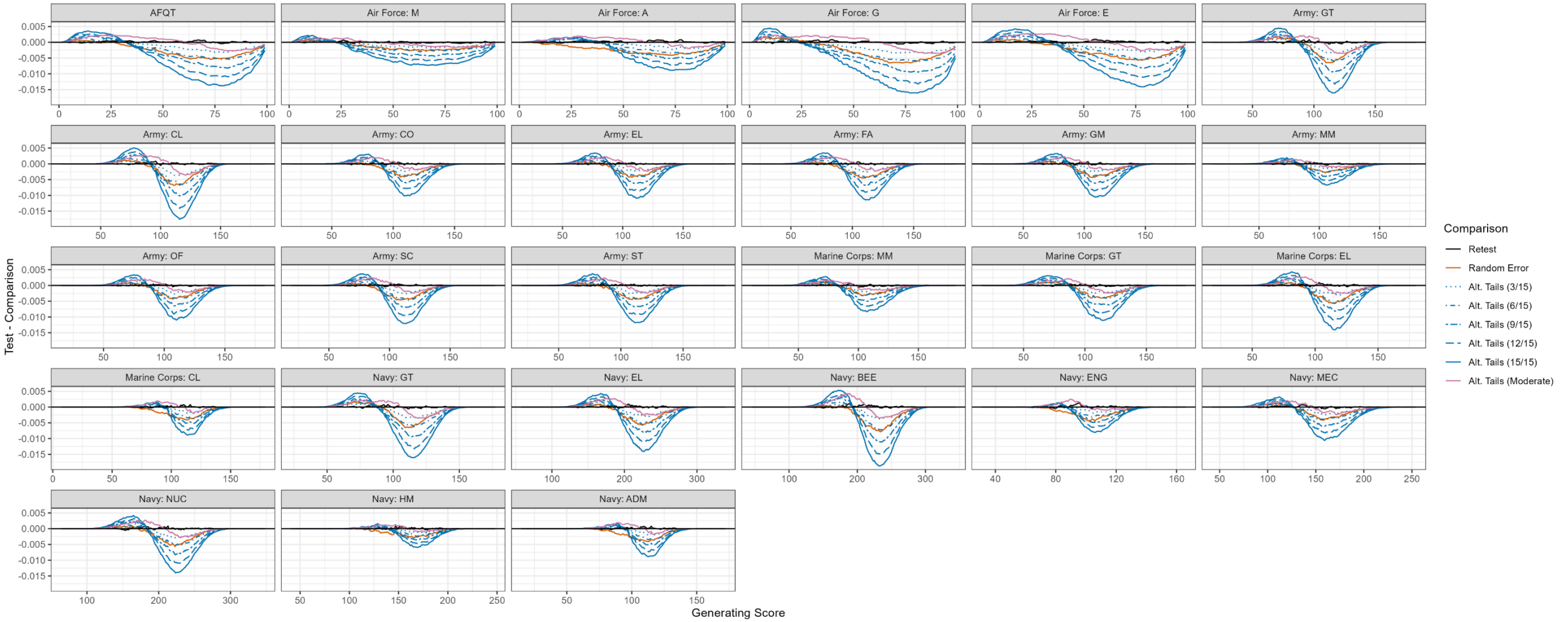
OPA
OFFICE OF PEOPLE ANALYTICS

Qualification Rate Differences per Condition for AFQT



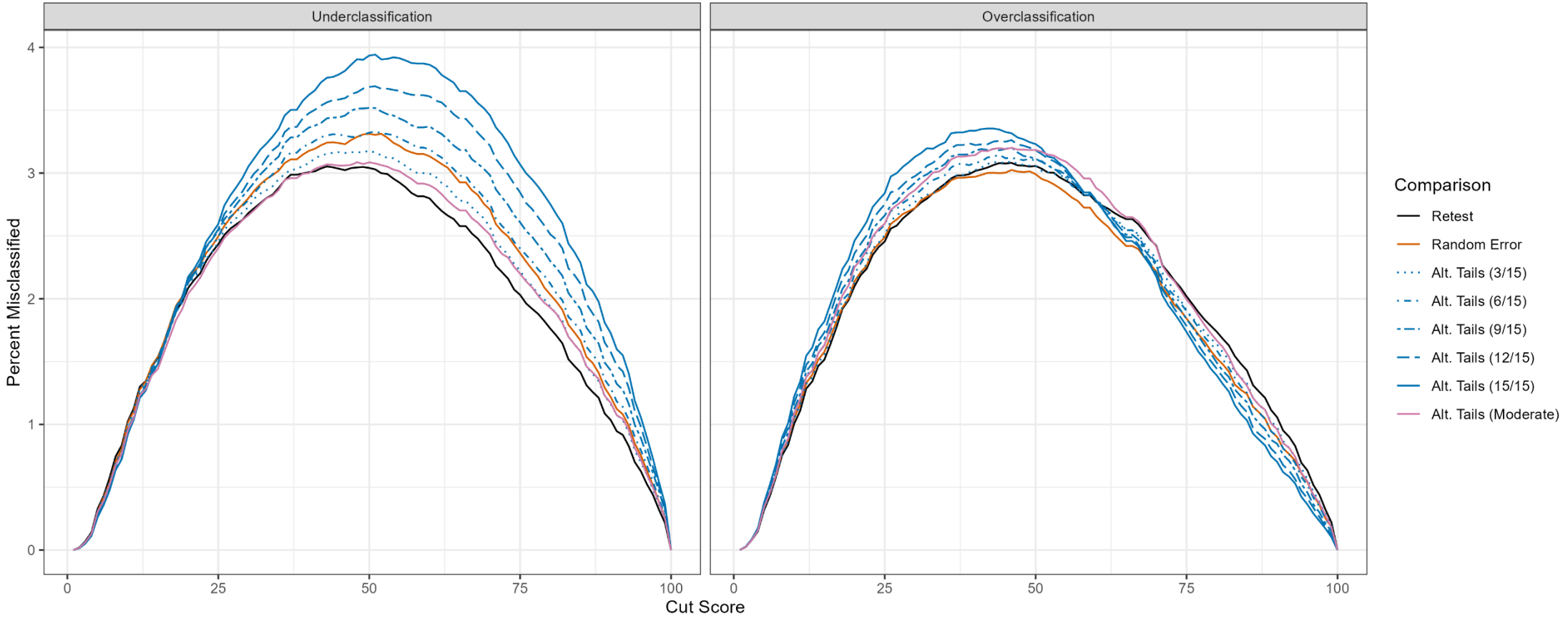
OPA
OFFICE OF PEOPLE ANALYTICS

Qualification Rate Differences per Composite and Condition



OPA
OFFICE OF PEOPLE ANALYTICS

AFQT Misclassification by Type and Condition



OPA
OFFICE OF PEOPLE ANALYTICS

Discussion

- Across bias, RMSE, reliability, mean score conditional bias, and qualification rate differences, in all conditions, calculator error introduces the same pattern of effects while the degree of these effects depends on the condition
- Pattern
  - Low-ability simulees have inflated scores while moderate-to-high ability simulees have deflated scores, with a larger effect for high-ability simulees
  - For AFQT, there is very little conditional bias at the IIIB cut score (31 on the percentile AFQT scale) across conditions
- Degree
  - Linear on proportion of items with manipulation, Random Error most like Alternating Tail Error (6/15)
  - The effect varies across composites and is predicted by the proportion of the composite that is contributed by AR and MK (see slide 16)
    - The most affected composite is Navy: BEE

Thank You!

For more information, please contact:

Glen Heinrich-Wallace
gheinrich-wallace@humrro.org
glen.heinrich-wallace.ctr@mail.mil

OPA
OFFICE OF PEOPLE ANALYTICS

References

- Bradley, K. (2025, January 22). Update on calculator impact study [Presentation]. DACMPT, El Paso, TX, United States.
- Heinrich-Wallace, G. (2024, June 12). An evaluation of calibration method and sample size on the reliability of new CAT-ASVAB forms [Presentation]. DACMPT, Monterey, CA, United States.

OPA
OFFICE OF PEOPLE ANALYTICS

Tab N

OPA

OFFICE OF PEOPLE ANALYTICS

Update on Math/Calculator Needs and Requirements Assessment

Monica Gribben

Human Resources Research Organization

Briefing presented to the DACMPT

January 22, 2025

Purpose

- Conduct a needs and requirements assessment to determine *whether a test assessing math content with a calculator is warranted* and, if so, use findings to inform what the taxonomy/blueprint would be.

Bottom Line Up Front (BLUF)

- Based on the sample of training courses and occupations in the needs assessment, there are no types of math where calculator use is a *prerequisite* for successful performance in training or on the job *across all types of occupations*.
  - The relatively *few* types of math where calculator use is a *prerequisite* for successful performance are primarily limited to three clusters of occupations:
    - Logistics and Administration
    - Science and Engineering
    - Medical
- The target sample was selected purposefully to include a range of occupations, including some with intensive math requirements. Due to limited participation in specific occupational areas and in some Services, the sample is not as robust as planned. We are working with the Services to augment the sample.

Procedures

- Administered Needs Assessment on HumRRO platform from June 2024 – October 2024
  - Types of math needed in training and role of calculators
  - Types of math needed on-the-job and role of calculators
- Met with MAPWG technical and policy reps to identify training staff and occupational managers across Services to receive the online Needs Assessment
  - Based on the 2022 Training Relevance Survey sample, the Needs Assessment sample includes training courses and occupations covering a variety of content, including some with intensive math requirements (e.g., Air Force Precision Measurement Equipment Laboratory 2P0X1)
  - In September, additional training courses and occupations were added for more representation in some job clusters

Procedures (cont.)

- Averaged responses from same training course or occupation to equally weight each training course/occupation represented
- Clustered responses into eight areas to summarize data:
  - Electrical
  - Infantry and Combat
  - Information Technology
  - Intelligence
  - Logistics and Administration
  - Mechanical
  - Medical
  - Science and Engineering

Results: Training Needs Assessment Responses

Service	Electrical	Infantry and Combat	Information Technology	Intelligence	Logistics and Administration	Mechanical	Medical	Science and Engineering	Overall
Air Force & Space Force	3 (3)			3 (4)	4 (4)	2 (10)	4 (4)	3 (4)	**19 (29)**
Army	7 (8)			1 (4)		11 (21)	2 (5)	2 (2)	**23 (40)**
Marine Corps	1 (3)	3 (3)				1 (1)			**5 (7)**
Navy	3 (5)		1 (2)	3 (5)		4 (6)		2 (9)	**13 (27)**
Overall	**14 (19)**	**3 (3)**	**1 (2)**	**7 (13)**	**4 (4)**	**18 (38)**	**6 (9)**	**7 (15)**	**60 (103)**

Response data include the number of training courses represented; number of respondents are in parentheses.

OPA
OFFICE OF PEOPLE ANALYTICS

Results: On-The-Job Needs Assessment Responses

Service	Electrical	Infantry and Combat	Information Technology	Intelligence	Logistics and Administration	Mechanical	Medical	Science and Engineering	Overall
Air Force				9 (11)	3 (3)		3 (3)	2 (2)	**17** **(19)**
Army	2 (2)			1 (1)		5 (5)	2 (4)	2 (2)	**12** **(14)**
Marine Corps	1 (2)	1 (1)				1 (1)			**3** **(4)**
Navy	10 (11)		2 (2)	2 (2)		14 (14)		3 (3)	**31** **(32)**
Space Force									**0**
Overall	**13** **(15)**	**1** **(1)**	**2** **(2)**	**12** **(14)**	**3** **(3)**	**20** **(20)**	**5** **(7)**	**7** **(7)**	**63** **(69)**

Response data include the number of training courses represented; number of respondents are in parentheses.

OPA
OFFICE OF PEOPLE ANALYTICS

Rating Scale

- Scaled responses for **with a calculator**
  - 0 = No, they do not do this type of math or yes, they **only** do this type of math **without** a calculator
  - 1 = Yes, they must be able to do this type of math **with a calculator**, but those who enter training (their first job) knowing how to do this type of math **do not perform better** than those who do not
  - 2 = Yes, they must be able to do this type of math **with a calculator**, and those who enter training (their first job) knowing how to do this type of math **do perform better** than those who do not

Results Interpretation

- Participants rated this type of math as 0 <= and <= 1, on average. In general, this type of math is not needed in training (on the job).
- Participants rated this type of math as 1 < and < 1.5, on average. Being able to do this type of math with a calculator is, in general, not a *prerequisite* for successful performance in training (on the job).
- Participants rated this type of math >=1.5, on average. Doing this type of math with a calculator is, in general, a *prerequisite* for successful performance in training (on the job).

Results for Math with a Calculator: Training (Arithmetic Reasoning)

Clusters circled in green highlight types of jobs where doing some types of math with a calculator is a *prerequisite* for successful performance in training.

	ELECTRICAL	INFANTRY AND COMBAT	INFORMATION TECHNOLOGY	INTELLIGENCE	LOGISTICS AND ADMINISTRATION	MECHANICAL	MEDICAL	SCIENCE AND ENGINEERING	OVERALL	Overall (without N < 5)
N survey participants:	19	3	2	13	4	38	9	15	103	94
N survey courses:	14	3	1	7	4	18	6	7	60	52
AR-1	0.69	0.33	0.00	0.43	1.00	1.19	1.00	1.17	0.73	0.90
AR-2	0.71	0.33	0.00	0.43	1.50	1.42	1.17	1.77	0.92	1.10
AR-3	0.62	0.33	0.00	0.43	1.25	1.31	1.00	1.79	0.84	1.03
AR-4	0.67	0.33	0.00	0.50	0.75	1.23	1.00	1.93	0.80	1.07
AR-5	0.64	0.00	0.00	0.50	0.75	1.44	0.83	1.24	0.68	0.93
AR-6	0.54	0.00	0.00	0.50	1.00	1.35	0.83	1.71	0.74	0.99
AR-7	0.21	0.00	0.00	0.32	1.75	1.07	0.67	1.45	0.68	0.74
AR-8	0.21	0.00	0.00	0.75	1.50	0.71	0.40	1.21	0.60	0.66
AR-9	0.21	0.00	0.00	0.75	0.50	1.17	0.17	1.31	0.51	0.72
AR-10	0.07	0.67	0.00	0.75	0.25	1.01	0.17	1.38	0.54	0.68
AR-11	0.21	0.00	0.00	0.29	0.00	0.81	0.83	1.57	0.46	0.74
AR-12	0.36	0.00	0.00	0.43	0.50	0.99	0.83	0.74	0.48	0.67

Legend:

<= 1
1 < and < 1.5
>=1.5

OPA
OFFICE OF PEOPLE ANALYTICS

Results for Math with a Calculator: Training (Mathematical Knowledge)

	ELECTRICAL	INFANTRY AND COMBAT	INFORMATION TECHNOLOGY	INTELLIGENCE	LOGISTICS AND ADMINISTRATION	MECHANICAL	MEDICAL	SCIENCE AND ENGINEERING	OVERALL	Overall (without N < 5)
N survey participants:	19	3	2	13	4	38	9	15	103	94
N survey courses:	14	3	1	7	4	18	6	7	60	52
MK-1	0.33	0.00	0.00	0.29	1.00	0.79	0.67	0.79	0.48	0.57
MK-2	0.36	0.00	0.00	0.29	1.25	0.73	0.83	0.86	0.54	0.61
MK-3	0.67	0.33	0.00	0.14	0.50	0.52	0.67	0.64	0.43	0.53
MK-4	0.14	0.33	0.00	0.17	0.50	0.58	0.33	0.64	0.34	0.37
MK-5	0.14	0.00	0.00	0.21	1.00	0.55	0.00	0.36	0.28	0.25
MK-6	0.21	0.33	0.00	0.17	0.50	0.64	0.33	0.50	0.34	0.37
MK-7	0.14	0.00	0.00	0.25	1.00	0.49	0.33	0.43	0.33	0.33
MK-8	0.21	0.00	0.00	0.50	1.50	0.86	0.95	1.07	0.64	0.72
MK-9	0.19	0.00	0.00	0.42	0.50	0.90	0.95	1.71	0.58	0.83
MK-10	0.57	0.33	0.00	0.33	1.00	0.91	0.83	1.36	0.67	0.80
MK-11	0.83	0.00	0.00	0.50	0.50	0.75	0.50	1.21	0.54	0.76
MK-12	0.50	0.33	0.00	0.25	0.50	0.93	0.67	0.93	0.51	0.66
MK-13	0.19	0.33	0.00	0.42	1.00	0.77	0.67	1.21	0.57	0.65
MK-14	0.14	0.00	0.00	0.00	0.50	0.59	0.00	1.07	0.29	0.36
MK-15	0.29	0.00	0.00	0.17	0.50	0.55	0.00	0.07	0.20	0.22
MK-16	0.57	0.50	0.00	0.33	1.50	1.06	0.50	1.50	0.74	0.79
MK-17	0.14	0.00	0.00	0.17	0.50	0.38	0.00	0.07	0.16	0.15
MK-18	0.71	0.50	0.00	0.42	1.00	0.63	0.78	1.14	0.65	0.74
MK-19	0.85	0.50	0.00	0.58	1.50	0.70	0.83	1.86	0.85	0.96
MK-20	0.69	0.50	0.00	0.58	1.50	0.59	0.83	1.79	0.81	0.90
MK-21	0.23	0.50	0.00	0.33	0.50	0.60	0.33	1.07	0.44	0.51
MK-22	0.36	0.50	0.00	0.25	1.00	0.55	0.83	1.07	0.57	0.61
MK-23	0.23	0.50	0.00	0.25	0.50	0.60	0.95	0.93	0.50	0.59
MK-24	0.00	0.50	0.00	0.33	0.50	0.39	0.00	0.29	0.25	0.20
MK-25	0.27	0.50	0.00	0.50	0.50	0.40	0.00	0.64	0.35	0.36
MK-26	0.38	0.50	0.00	0.17	0.00	0.44	0.17	1.00	0.33	0.43
MK-27	0.31	0.50	0.00	0.25	0.00	0.97	0.00	0.93	0.37	0.49
MK-28	0.54	0.00	0.00	0.08	0.00	1.05	0.17	1.14	0.37	0.60
MK-29	0.00	0.00	0.00	0.08	0.00	0.36	0.00	0.21	0.08	0.13
MK-30	0.31	0.50	0.00	0.08	0.00	0.56	0.00	0.64	0.26	0.32
MK-31	0.08	0.00	0.00	0.33	0.00	0.77	0.17	0.68	0.25	0.41
MK-32	0.08	0.00	0.00	0.50	0.00	0.94	0.17	1.07	0.35	0.55
MK-33	0.23	0.00	0.00	0.25	0.00	0.75	0.60	1.14	0.37	0.59
MK-34	0.62	0.00	0.00	0.50	1.00	0.95	0.60	1.50	0.65	0.83

Legend:

<= 1
1 < and < 1.5
>=1.5

OPA
OFFICE OF PEOPLE ANALYTICS

Results for Math with a Calculator: Training (Additional Types of Math)

Additional types of math are from a taxonomy of math generated by Waugh et al. (2015). The additional types of math are those that experts rated as not included in the AR and MK blueprints.

	ELECTRICAL	INFANTRY AND COMBAT	INFORMATION TECHNOLOGY	INTELLIGENCE	LOGISTICS AND ADMINISTRATION	MECHANICAL	MEDICAL	SCIENCE AND ENGINEERING	OVERALL	Overall (without N < 5)
N survey participants:	19	3	2	13	4	38	9	15	103	94
N survey courses:	14	3	1	7	4	18	6	7	60	52
AM-1	0.23	0.00	0.00	0.50	1.00	0.93	0.40	0.50	0.44	0.51
AM-2	0.54	0.00	0.00	0.17	0.50	0.35	0.33	0.86	0.34	0.45
AM-3	0.08	0.50	0.00	0.42	0.00	0.31	0.17	0.29	0.22	0.25
AM-4	0.46	0.50	0.00	0.58	0.50	0.85	0.95	1.21	0.63	0.81
AM-5	0.31	0.00	0.00	0.25	1.00	0.50	0.33	0.14	0.32	0.31
AM-6	0.08	0.00	0.00	0.00	0.00	0.17	0.00	0.07	0.04	0.06
AM-7	0.08	0.00	0.00	0.33	0.00	0.06	0.17	0.07	0.09	0.14
AM-8	0.38	0.00	0.00	0.42	0.50	0.34	0.00	0.36	0.25	0.30
AM-9	0.39	0.00	0.00	0.17	0.50	0.63	0.67	0.43	0.35	0.46
AM-10	0.15	0.50	0.00	0.00	0.00	0.36	0.50	0.43	0.24	0.29
AM-11	0.08	0.00	0.00	0.33	0.50	0.11	0.33	0.14	0.19	0.20
AM-12	0.08	0.00	0.00	0.42	1.50	0.64	0.50	0.21	0.42	0.37
AM-13	0.08	0.00	0.00	0.08	0.50	0.53	0.33	0.50	0.25	0.30
AM-14	0.00	0.00	0.00	0.17	0.50	0.22	0.00	0.36	0.16	0.15
AM-15	0.31	0.00	0.00	0.42	1.00	0.57	0.83	0.67	0.48	0.56
AM-16	0.23	0.00	0.00	0.00	0.50	0.44	0.50	0.42	0.26	0.32
AM-17	0.08	0.50	0.00	0.00	0.00	0.32	0.50	0.08	0.18	0.20
AM-18	0.08	0.00	0.00	0.17	0.50	0.15	0.00	0.42	0.16	0.16
AM-19	0.15	0.00	0.00	0.25	0.00	0.25	0.17	0.67	0.19	0.30
AM-20	0.15	0.00	0.00	0.17	1.00	0.28	0.33	0.00	0.24	0.19
AM-21	0.00	0.00	0.00	0.33	1.00	0.12	0.33	0.00	0.22	0.16
AM-22	0.00	0.00	0.00	0.33	1.50	0.15	0.33	0.00	0.29	0.16
AM-23	0.00	0.00	0.00	0.42	1.00	0.26	0.33	0.00	0.25	0.20
AM-24	0.00	0.00	0.00	0.42	1.00	0.00	0.33	0.00	0.22	0.15
AM-25	0.00	0.00	0.00	0.17	1.00	0.00	0.33	0.08	0.20	0.12
AM-26	0.00	0.00	0.00	0.17	1.50	0.11	0.33	0.08	0.27	0.14
AM-27	0.00	0.00	0.00	0.17	1.50	0.00	0.33	0.42	0.30	0.18
AM-28	0.00	0.00	0.00	0.58	1.00	0.11	0.33	0.00	0.25	0.20
AM-29	0.00	0.00	0.00	0.58	1.50	0.17	0.33	0.00	0.32	0.22
AM-30	0.00	0.00	0.00	0.25	0.00	0.33	0.00	0.62	0.15	0.24
AM-31	0.15	0.00	0.00	0.25	0.00	0.14	0.00	1.00	0.19	0.31
AM-32	0.08	0.00	0.00	0.58	0.50	0.11	0.50	0.25	0.25	0.30
AM-33	0.23	0.00	0.00	0.17	0.00	0.11	0.00	0.58	0.14	0.22
AM-34	0.08	0.00	0.00	0.58	0.50	0.12	0.00	0.54	0.23	0.26
AM-35	0.31	0.00	0.00	0.25	0.00	0.58	0.17	1.08	0.30	0.48
AM-36	0.00	0.00	0.00	0.42	0.00	0.17	0.00	0.67	0.16	0.25
AM-37	0.00	0.00	0.00	0.25	0.00	0.17	0.17	0.42	0.13	0.20
AM-38	0.00	0.00	0.00	0.08	0.00	0.17	0.17	0.42	0.10	0.17
AM-39	0.00	0.00	0.00	0.17	0.00	0.29	0.17	0.42	0.13	0.21
AM-40	0.00	0.00	0.00	0.25	0.00	0.31	0.00	0.42	0.12	0.20
AM-41	0.28	0.00	0.00	0.25	0.00	0.64	0.17	0.67	0.25	0.40

Legend:

<= 1
1 < and < 1.5
>=1.5

OPA
OFFICE OF PEOPLE ANALYTICS

Results for Math with a Calculator: On-the-Job (Arithmetic Reasoning)

Infantry and Combat results were based on 1 job as of November 12. This sample size is not suitable for decision making.

Clusters circled in green highlight types of jobs where doing some types of math with a calculator is a *prerequisite* for successful performance on the job.

	ELECTRICAL	INFANTRY AND COMBAT	INFORMATION TECHNOLOGY	INTELLIGENCE	LOGISTICS AND ADMINISTRATION	MECHANICAL	MEDICAL	SCIENCE AND ENGINEERING	OVERALL	Overall (without N < 5)
N survey participants:	15	1	2	14	3	20	7	7	69	63
N survey occupations:	13	1	2	12	3	20	5	7	63	57
AR-1	0.77	2.00	0.50	0.58	2.00	0.90	1.20	1.00	1.12	0.89
AR-2	0.92	2.00	0.50	1.17	1.33	1.00	1.00	1.29	1.15	1.08
AR-3	0.92	2.00	1.00	0.96	2.00	0.70	1.00	0.71	1.16	0.86
AR-4	0.77	2.00	1.00	1.17	1.33	0.80	1.20	0.86	1.14	0.96
AR-5	0.92	0.00	1.00	1.00	1.33	0.85	1.80	0.86	0.97	1.09
AR-6	0.77	0.00	1.00	1.00	1.33	0.65	1.00	0.86	0.83	0.86
AR-7	0.69	0.00	0.50	0.83	1.33	1.00	1.00	0.83	0.77	0.87
AR-8	0.38	0.00	0.50	1.08	0.67	1.10	1.40	0.67	0.72	0.93
AR-9	0.46	0.00	0.00	0.58	0.00	0.70	0.60	1.00	0.42	0.67
AR-10	0.77	0.00	0.50	0.58	0.00	0.75	0.60	1.50	0.59	0.84
AR-11	0.23	0.00	0.00	0.50	0.00	0.80	1.00	1.17	0.46	0.74
AR-12	0.85	0.00	0.50	0.83	0.67	0.90	1.40	0.33	0.68	0.86

Legend:

<= 1
1 < and < 1.5
>=1.5

OPA
OFFICE OF PEOPLE ANALYTICS

Results for Math with a Calculator: On-the-Job (Mathematical Knowledge)

	ELECTRICAL	INFANTRY AND COMBAT	INFORMATION TECHNOLOGY	INTELLIGENCE	LOGISTICS AND ADMINISTRATION	MECHANICAL	MEDICAL	SCIENCE AND ENGINEERING	OVERALL	Overall (without N < 5)
N survey participants:	15	1	2	14	3	20	7	7	69	63
N survey occupations:	13	1	2	12	3	20	5	7	63	57
MK-1	0.46	0.00	1.00	0.67	0.00	0.68	0.20	0.67	0.46	0.54
MK-2	0.46	0.00	1.00	0.75	0.00	0.55	0.60	1.00	0.54	0.67
MK-3	0.31	0.00	0.00	0.58	0.67	0.55	1.20	0.33	0.46	0.59
MK-4	0.31	0.00	0.00	0.50	0.67	0.60	0.40	0.33	0.35	0.43
MK-5	0.38	0.00	0.00	0.25	0.67	0.55	0.00	0.33	0.27	0.30
MK-6	0.38	0.00	1.00	1.00	0.67	0.55	1.00	0.67	0.66	0.72
MK-7	0.00	0.00	0.00	0.21	0.67	0.60	0.80	0.33	0.33	0.39
MK-8	0.85	0.00	1.00	1.25	0.67	0.90	1.20	0.33	0.78	0.91
MK-9	0.62	1.00	0.00	1.00	0.67	0.70	0.60	1.17	0.72	0.82
MK-10	0.69	0.00	0.00	0.92	1.33	0.80	1.20	1.67	0.83	1.06
MK-11	0.15	0.00	0.00	0.71	1.33	0.70	1.00	0.67	0.57	0.65
MK-12	0.38	0.00	0.50	0.75	1.33	0.75	0.90	0.67	0.66	0.69
MK-13	0.15	2.00	0.00	0.50	0.67	0.53	0.60	1.00	0.68	0.56
MK-14	0.15	0.00	0.50	0.33	1.33	0.53	1.00	0.67	0.56	0.54
MK-15	0.23	0.00	0.00	0.42	0.67	0.42	0.40	0.33	0.31	0.36
MK-16	0.54	0.00	0.00	1.29	1.00	0.85	1.60	1.17	0.81	1.09
MK-17	0.08	0.00	0.00	0.67	0.67	0.30	0.40	0.33	0.31	0.36
MK-18	0.92	0.00	0.00	0.88	0.67	0.90	1.20	1.33	0.74	1.05
MK-19	0.62	1.00	0.00	1.00	0.67	0.95	1.20	1.33	0.85	1.02
MK-20	0.92	2.00	0.00	1.08	0.67	0.85	0.80	1.00	0.92	0.93
MK-21	0.23	0.00	0.00	0.96	0.67	0.65	0.40	0.67	0.45	0.58
MK-22	0.46	0.00	0.00	0.33	0.67	0.65	1.00	0.67	0.47	0.62
MK-23	0.08	0.00	0.00	0.33	0.00	0.65	1.00	1.33	0.42	0.68
MK-24	0.15	0.00	0.00	0.42	0.67	0.40	0.20	0.50	0.29	0.33
MK-25	0.54	0.00	0.50	0.67	0.00	0.84	0.40	1.17	0.52	0.72
MK-26	0.15	0.00	0.50	0.42	0.00	0.60	0.00	0.83	0.31	0.40
MK-27	0.38	2.00	0.00	0.58	0.00	0.95	0.40	0.83	0.64	0.63
MK-28	0.69	2.00	0.00	0.50	0.67	0.80	0.80	0.83	0.79	0.72
MK-29	0.38	0.00	0.00	0.08	0.00	0.50	0.40	0.67	0.25	0.41
MK-30	0.46	0.00	0.50	0.42	0.00	0.65	0.40	1.00	0.43	0.59
MK-31	0.38	0.00	0.00	0.67	0.00	0.75	0.40	1.67	0.48	0.77
MK-32	0.54	0.00	0.00	0.50	0.00	0.85	0.60	1.20	0.46	0.74
MK-33	0.23	0.00	0.00	0.25	0.00	0.85	0.60	1.67	0.45	0.72
MK-34	0.77	2.00	0.00	1.00	0.67	0.90	1.00	1.17	0.94	0.97

Legend:

<= 1
1 < and < 1.5
>=1.5

OPA
OFFICE OF PEOPLE ANALYTICS

Results for Math with a Calculator: On-the-Job (Additional Types of Math)

Additional types of math are from a taxonomy of math generated by Waugh et al. (2015). The additional types of math are those that experts rated as not included in the AR and MK blueprints.

	ELECTRICAL	INFANTRY AND COMBAT	INFORMATION TECHNOLOGY	INTELLIGENCE	LOGISTICS AND ADMINISTRATION	MECHANICAL	MEDICAL	SCIENCE AND ENGINEERING	OVERALL	Overall (without N < 5)
N survey participants:	15	1	2	14	3	20	7	7	69	63
N survey occupations:	13	1	2	12	3	20	5	7	63	57
AM-1	0.38	2.00	0.50	0.79	0.67	0.65	0.60	0.67	0.78	0.62
AM-2	0.62	0.00	0.00	0.58	0.00	0.30	0.80	0.33	0.33	0.53
AM-3	0.38	0.00	0.00	0.46	0.00	0.30	0.50	0.17	0.23	0.36
AM-4	0.62	2.00	0.50	0.67	0.67	0.60	1.00	1.00	0.88	0.78
AM-5	0.15	0.00	0.50	0.58	0.67	0.30	0.80	0.67	0.46	0.50
AM-6	0.15	2.00	0.00	0.25	0.67	0.20	0.40	0.00	0.46	0.20
AM-7	0.00	0.00	0.00	0.58	0.00	0.25	0.40	0.33	0.20	0.31
AM-8	0.23	0.00	0.00	0.50	0.67	0.25	1.00	0.83	0.44	0.56
AM-9	0.42	0.00	0.00	0.33	0.67	0.40	0.40	0.33	0.32	0.38
AM-10	0.17	0.00	0.00	0.17	0.67	0.30	0.80	0.67	0.35	0.42
AM-11	0.17	0.00	0.00	0.17	0.67	0.40	0.80	0.67	0.36	0.44
AM-12	0.50	0.00	0.00	0.58	0.67	0.55	1.00	0.67	0.50	0.66
AM-13	0.17	0.00	0.50	0.25	0.67	0.30	0.40	1.00	0.41	0.42
AM-14	0.00	0.00	0.00	0.17	0.67	0.20	0.00	0.67	0.21	0.21
AM-15	0.50	0.00	0.00	0.42	0.67	0.60	0.80	1.17	0.52	0.70
AM-16	0.33	0.00	0.50	0.33	0.67	0.30	0.80	0.33	0.41	0.42
AM-17	0.42	0.00	0.00	0.67	0.67	0.20	0.80	0.33	0.39	0.48
AM-18	0.17	0.00	0.00	0.17	0.00	0.20	0.00	0.33	0.11	0.17
AM-19	0.25	0.00	0.50	0.42	0.00	0.50	0.40	0.50	0.32	0.41
AM-20	0.00	0.00	0.00	0.42	0.67	0.40	0.40	0.17	0.26	0.28
AM-21	0.17	0.00	0.00	0.58	0.67	0.40	0.40	0.33	0.32	0.38
AM-22	0.00	0.00	0.50	0.25	0.67	0.30	0.40	0.00	0.26	0.19
AM-23	0.25	0.00	0.50	1.08	0.67	0.40	0.20	0.67	0.47	0.52
AM-24	0.08	2.00	0.00	0.83	0.67	0.30	0.80	0.67	0.67	0.54
AM-25	0.08	0.00	0.00	0.25	0.67	0.30	0.60	0.33	0.28	0.31
AM-26	0.25	0.00	0.00	0.83	0.67	0.50	0.80	0.67	0.47	0.61
AM-27	0.17	0.00	0.00	0.75	0.67	0.30	0.40	0.33	0.33	0.39
AM-28	0.17	0.00	0.00	0.83	1.00	0.40	0.60	0.00	0.38	0.40
AM-29	0.25	2.00	0.50	0.83	1.00	0.50	0.40	0.67	0.77	0.53
AM-30	0.33	2.00	0.00	0.50	0.00	0.40	0.00	0.17	0.42	0.28
AM-31	0.25	0.00	0.00	0.42	0.00	0.40	0.00	0.17	0.16	0.25
AM-32	0.08	0.00	0.00	0.42	0.00	0.20	0.40	0.17	0.16	0.25
AM-33	0.42	0.00	0.00	0.27	0.00	0.30	0.00	0.67	0.21	0.33
AM-34	0.50	0.00	0.00	0.17	0.00	0.30	0.60	0.17	0.22	0.35
AM-35	0.42	0.00	0.50	0.42	0.00	0.55	0.00	1.33	0.40	0.54
AM-36	0.25	0.00	0.00	0.08	0.00	0.20	0.00	0.67	0.15	0.24
AM-37	0.08	0.00	0.00	0.42	0.00	0.20	0.00	0.50	0.15	0.24
AM-38	0.00	0.00	0.00	0.25	0.00	0.20	0.00	0.33	0.10	0.16
AM-39	0.42	0.00	0.00	0.25	0.00	0.40	0.40	0.50	0.25	0.39
AM-40	0.08	0.00	0.00	0.25	0.00	0.53	0.40	0.17	0.18	0.29
AM-41	0.33	0.00	0.00	0.42	0.00	0.95	1.40	1.17	0.53	0.85

Legend:

<= 1
1 < and < 1.5
>=1.5

OPA
OFFICE OF PEOPLE ANALYTICS

Conclusion

- Based on the sample of training courses and occupations in the needs assessment, there are relatively *few* types of math where calculator use is a *prerequisite* for successful performance in training or on the job (green-shaded cells).
  - Furthermore, the relatively *few* types of math where calculator use is a *prerequisite* for successful performance are primarily limited to three clusters:
    - Logistics and Administration
    - Science and Engineering
    - Medical
- There are some other types of math where calculators are used in training or on the job, but calculator use is generally *not a prerequisite* for success (yellow-shaded cells).
- The target sample was selected purposefully to include a range of occupations and math requirements. Due to limited participation in specific occupational areas and in some Services, the sample is not as robust as planned. We are working with the Services to augment the sample.

Questions for the DAC

OPA
OFFICE OF PEOPLE ANALYTICS

Questions for the DAC

- Does the DAC have recommendations to address any of the implications identified in the calculator impact study?
- Are there other complications resulting from calculator error that could affect CAT tests, specifically, that were not addressed in the simulation study?
- Based on the results of the Needs Assessment, does the DAC believe the results support the need for a special purpose test that assesses math with a calculator for use in classification?

OPA
OFFICE OF PEOPLE ANALYTICS

Thank you!

For more information
please contact:

Monica Gribben
mgribben@humrro.org
703.706.5690

OPA
OFFICE OF PEOPLE ANALYTICS

Tab O

OPA

OFFICE OF PEOPLE ANALYTICS

Refinement of the Joint-Service TAPAS Instrument

Dan Putka

Human Resources Research Organization

Briefing presented to the DACMPT

January 23, 2025

Briefing Agenda

- Joint-Service (JS) TAPAS Background Refresh
  - JS TAPAS Composites, Instrument, and Development Phases
- Recap of Preliminary Phase 1 JS TAPAS Composite Recommendations
- FY24 Research to Inform Phase 1 JS TAPAS Revisions
- Finalizing the Phase 1 JS TAPAS Design
- Next Steps for JS TAPAS
  - Operations and Maintenance (O&M) Track (FY25)
  - R&D Track (FY25-26)
- Questions for the DAC

OPA
OFFICE OF PEOPLE ANALYTICS

Joint-Service TAPAS Background Refresh

OPA
OFFICE OF PEOPLE ANALYTICS

Joint-Service TAPAS Mission

1. Develop a composite for military compatibility
   - Designed to predict alignment with military core values — various forms of misconduct
   - DoD directive that applies to enlisted personnel

2. Develop a composite for enlisted selection
   - Designed to predict first-term enlisted job performance
   - Expand qualified applicant pool without compromising valued outcomes

3. Develop a Joint-Service TAPAS instrument

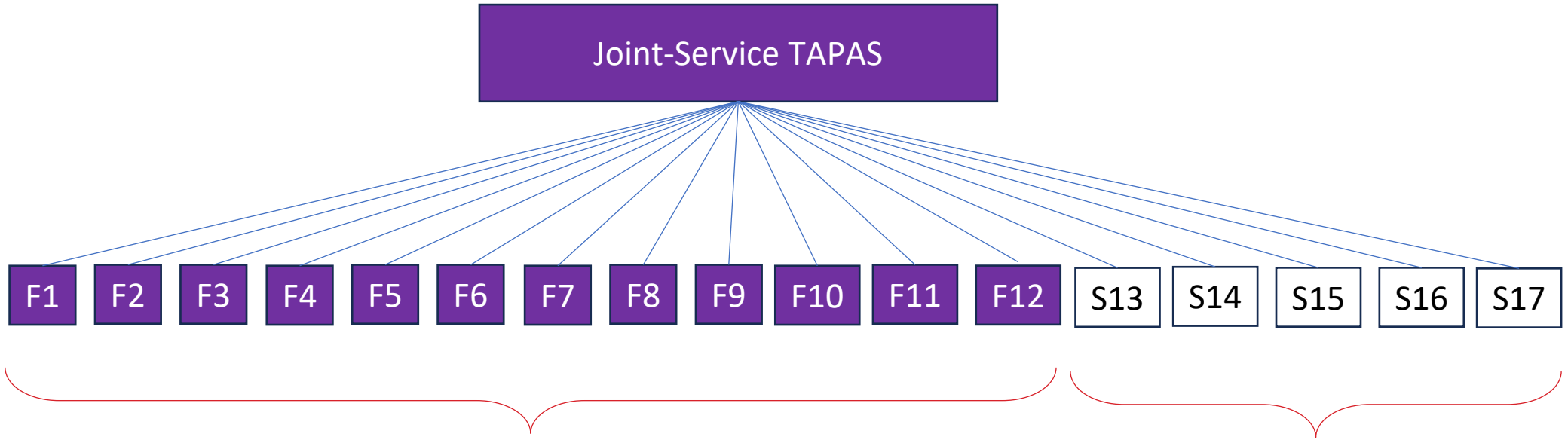
OPA
OFFICE OF PEOPLE ANALYTICS

Joint-Service (JS) TAPAS Concept

- The JS TAPAS “instrument” is modular and will include:
  - A common core of facets that support scoring of the military compatibility (MC) and enlisted (ENL) composites
  - Service-specific facets to support Service-specific use cases

OPA
OFFICE OF PEOPLE ANALYTICS

Joint-Service TAPAS Concept

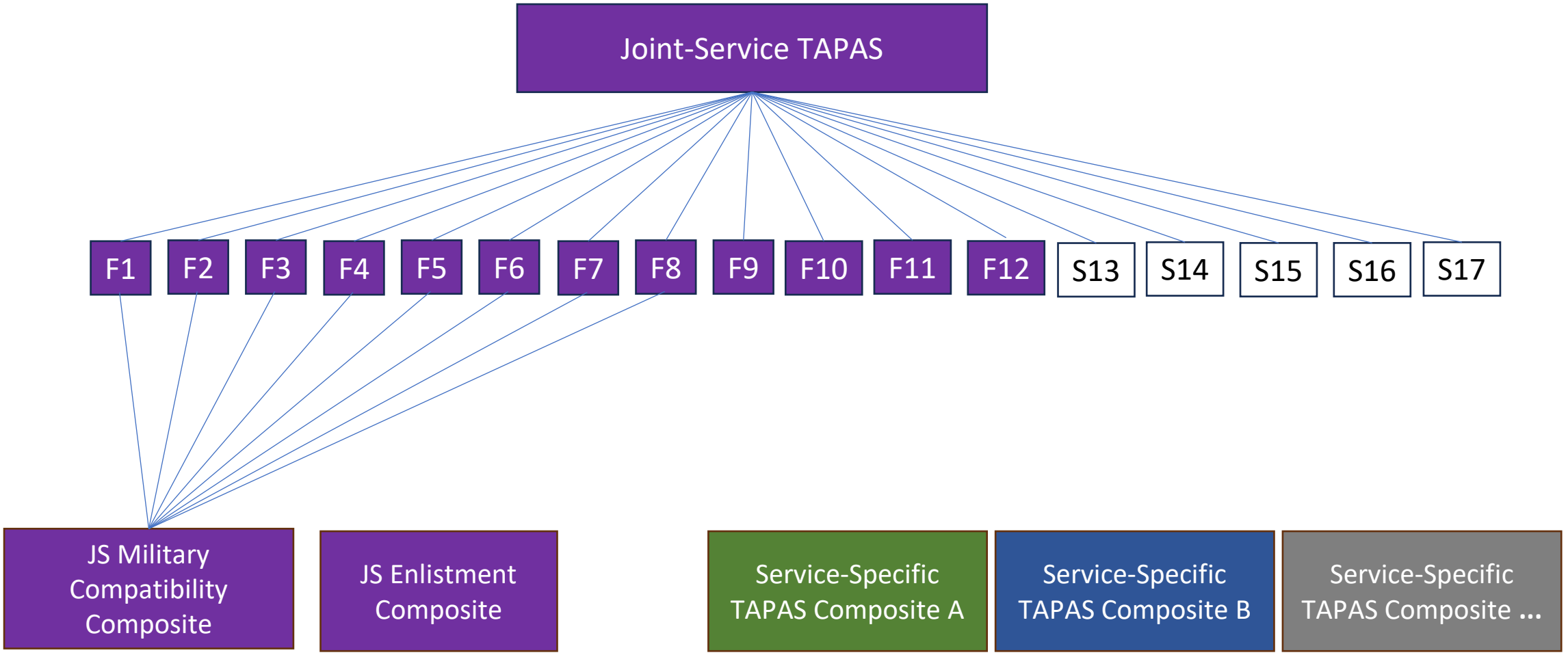


JS MC and ENL composite facets shared across all TAPAS versions

Slots reserved for Service-specific facets

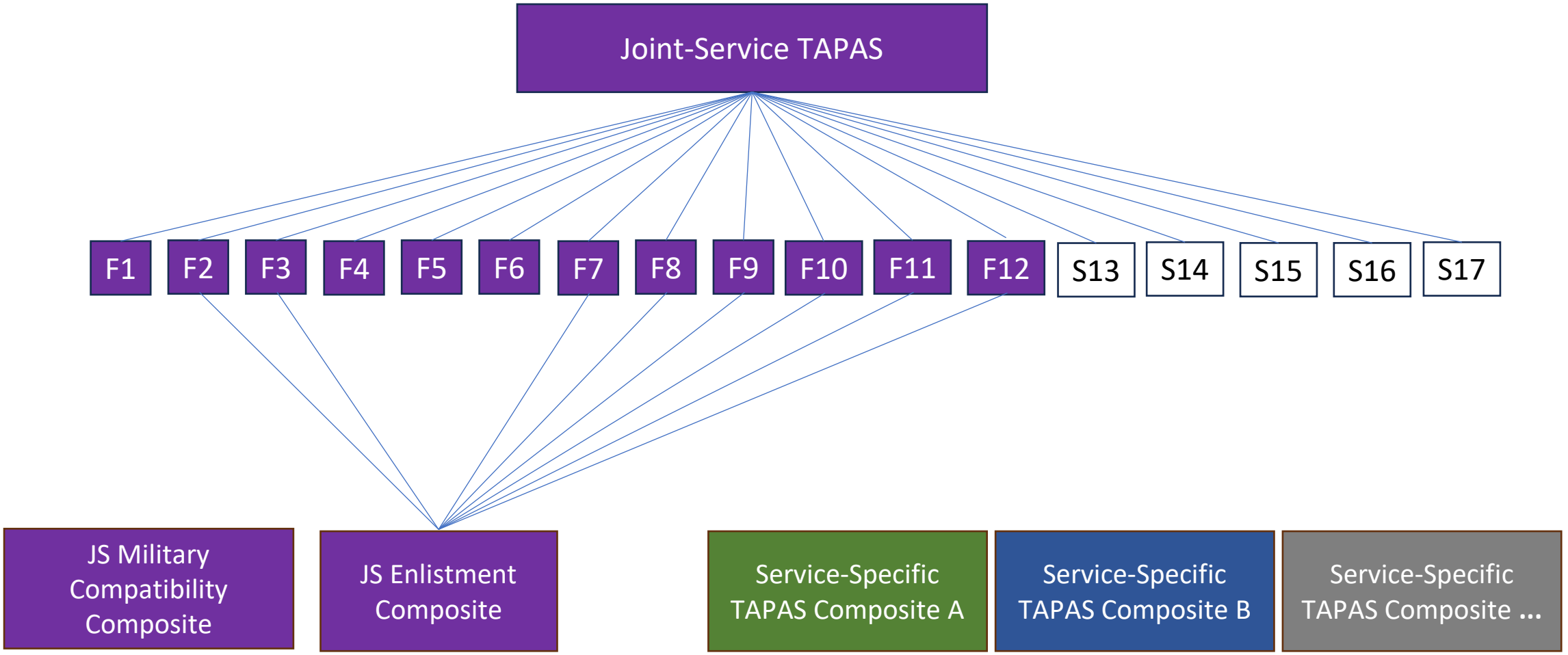
OPA
OFFICE OF PEOPLE ANALYTICS

Joint-Service TAPAS Concept



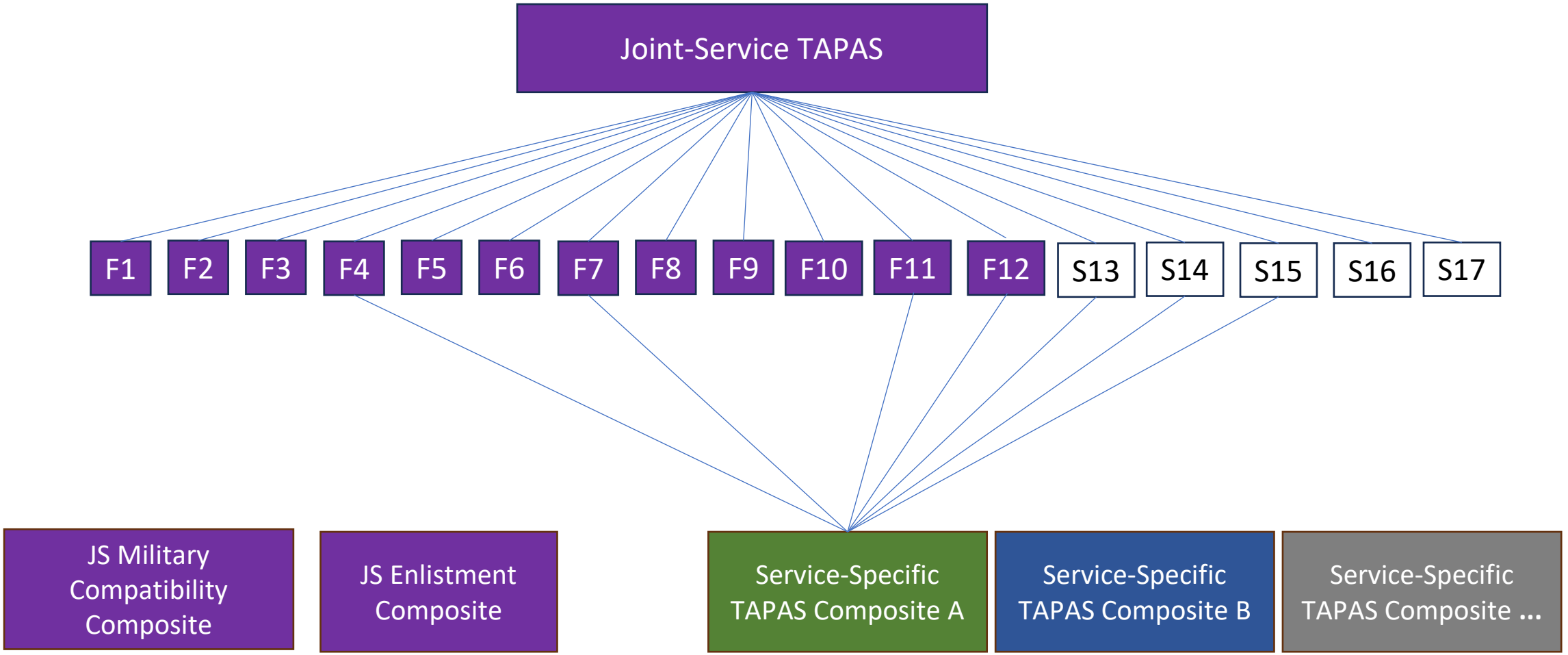
OPA
OFFICE OF PEOPLE ANALYTICS

Joint-Service TAPAS Concept



OPA
OFFICE OF PEOPLE ANALYTICS

Joint-Service TAPAS Concept



OPA
OFFICE OF PEOPLE ANALYTICS

Phased Development Approach

- **Phase 0** JS TAPAS instrument and composites
  - FY23 work designed to address immediate OSD tasking
  - Features **interim** MC and ENL composites
  - Added facets to USAF and USMC TAPAS needed for scoring of interim MC composite
  - Implemented at MEPS in September 2024
    - Phase 0 MC and ENL composites scored but not used for operational decision making

OPA
OFFICE OF PEOPLE ANALYTICS

Phased Development Approach

- **Phase 1** JS TAPAS instrument and composites
  - **Preliminary** recommendations for Phase 1 composites (facets, weighting) made based on FY23 research
  - **Refined** recommendations for Phase 1 instrument (design, JS facet set) made based on FY24 research
  - Content development and psychometric work to occur in FY25
    - **Refined** composition and facet weighting Phase 1 MC and ENL composites
    - Updating of TAPAS statements pools
    - Calibrating TAPAS statement pools with a joint-Service sample
    - Develop provisional joint-Service norms for JS and SS facets
  - IT work to enable implementation at MEPS sometime in FY27 (TBD)

Phased Development Approach

- **Phase 2**: Evaluation and refinement of Phase 1 JS composites for operational decision making
  - Update joint-Service norms for JS and SS facets
    - Informed by FY27 applicant data and subsequent evaluation work
  - Revisit composition and weights for each Phase 1 composite and adjust as needed
  - Establish an evidentiary base for use of final Phase 2 composites for enlistment and military compatibility-related screening decisions (e.g., criterion-related validity study for enlistment composite)

FY23	Foundational research to inform Phase 0 and preliminary recommendations for Phase 1 JS TAPAS composites	
FY24	Follow-up research to inform refined recommendations for Phase 1 JS TAPAS instrument (design, facet set)	IT work to support implementation of Phase 0 JS TAPAS at MEPS
FY25	Development work to support Phase 1 JS TAPAS instrument and refined composites + JS TAPAS R&D	Began administering Phase 0 JS TAPAS at MEPS
FY26	JS TAPAS R&D (continued) Evaluation and refinement of Phase 1 JS TAPAS composites	IT work to support implementation of Phase 1 JS TAPAS at MEPS
FY27		Begin administering Phase 1 JS TAPAS at MEPS
FY28		

Begin operational use of TAPAS composites based on FY26–FY28 work

OPA
OFFICE OF PEOPLE ANALYTICS

Recap of Preliminary Phase 1 JS TAPAS Composite Recommendations

OPA
OFFICE OF PEOPLE ANALYTICS

Military Compatibility (MC) Composite – Focal Criterion

- Focal criterion reflects 10 categories of misconduct
  - Violent Behavior
  - Sexual Violence/ Assault
  - Sexual Harassment
  - Harassment and Non-Violent Abuse
  - Disclosing Classified or Sensitive Information
  - Rebellious/Extremist Behavior
  - Unethical Behavior
  - Vandalism/Sabotage
  - Theft
  - Production Deviance
- Informed by literature and expert review
  - Counterproductive work behavior (CWB) literature (e.g., Spector et al., 2006)
  - Uniform Code of Military Justice
  - OPA/PERSEREC reports
  - DoD instruction 1304.26

OPA
OFFICE OF PEOPLE ANALYTICS

Preliminary Phase 1 MC Composite Recommendations

- Subject matter experts (SMEs) evaluated conceptual and empirical evidence of alignment between TAPAS facets and 10 categories of misconduct
  - Rated alignment as strong, moderate, or weak
- Reached consensus on facet composition and weighting for a preliminary Phase 1 MC composite *[facets withheld for test security]*
  - See June 2023 DACMPT slides for more details

OPA
OFFICE OF PEOPLE ANALYTICS

Enlistment Composite – Focal Criterion

- Focal criterion reflects first-term enlisted job performance composite
- Based on performance dimensions from Russell et al. (2023)[1] taxonomy
- Captured "overall performance" policy from Service stakeholders
  - Task Performance, Decision Making, Problem Solving, and Innovation (m = 17.0)
  - Organizational Support (m = 12.8)
  - Support for Peers (m = 10.2)
  - Conscientious Initiative (m = 10.2)
  - Communication (m = 9.8)
  - Adjusting to Stressful Situations (m = 9.2)
  - Physical Performance (m = 9.2)
  - Safety and Security Consciousness (m = 8.2)
  - Initiating Structure for Self and Others (m = 8.0)
  - Counterproductive Work Behavior (m = 5.4)

Note. Parenthetical values reflect distribution of 100 points across dimensions.

[1]Russell, T., Allen, M., Ford, L., Carretta, T., & Kirkendall, C. (2023). Development of a performance taxonomy for entry-level military occupations. *Military Psychology*, *35*(4), 283-294. https://doi.org/10.1080/08995605.2022.2050163.

OPA
OFFICE OF PEOPLE ANALYTICS

Preliminary Phase 1 ENL Composite Recommendations

- Gathered archival and SME data to support development and validation
  - Developed regression-weighted composite based on mix of archival and SME-estimated correlations
  - See June 2023 DACMPT slides for more details
- Identified subset of facets for predicting first-term enlisted job performance based on regression models *[facets withheld for sensitivity]*

OPA
OFFICE OF PEOPLE ANALYTICS

FY24 Research to Inform Phase 1 JS TAPAS Revisions

OPA
OFFICE OF PEOPLE ANALYTICS

Focus of FY24 Research

- Largely focused on refining preliminary Phase 1 JS TAPAS recommendations and identifying needs for FY25 development work
  - Conducted multiple research efforts pertinent to evaluating TAPAS facets and their statement pools
  - Engaged in multiple rounds of discussion with OSD and Services to arrive at an agreed-upon set of JS facets and JS instrument design/configuration
    - Factoring in FY23 recommendations *AND* FY24 research results
  - Established plans for recalibration of TAPAS statements with a joint-Service sample

OPA
OFFICE OF PEOPLE ANALYTICS

Outline of FY24 Research Activities

- Retranslation of facet statements
- Bias and sensitivity review of facet statements
- Susceptibility of facet statements to transient error
- Revisiting marginal IRT reliability of facet scores
- Equivalence of facet scores across TAPAS versions
- Composite shortening analyses

The above research provided additional perspectives on the functioning of TAPAS facets beyond what was known when preliminary Phase 1 composites recommendations were made in FY23.

OPA
OFFICE OF PEOPLE ANALYTICS

Retranslation of Facet Statements

- Purpose
  - Evaluate whether TAPAS statements are clear indicators of their intended facets
- Method
  - Leveraged natural language processing (NLP) methods to identify items most in need of review by SMEs (*n* = 482 out of 1,200+ statements in DoD TAPAS statement pool)
    - Focused on statements that were more semantically similar to statements of another facet rather than their intended facet
  - Eight psychologist SMEs independently indicated which facet each statement primarily measured
    - At least 6 of 8 (75%) SMEs had to agree on the facet a statement was designed to measure for it to be considered “translated” to that facet

Retranslation of Facet Statements

Target Facet	% of the Target Facet Statements Retranslated to		
	Target Facet	Non-Target Facet	No Clear Translation
Physical Conditioning	100.0	0.0	0.0
Team Orientation	100.0	0.0	0.0
Tolerance	100.0	0.0	0.0
Order	96.0	0.0	4.0
Sociability	95.8	2.1	2.1
Non-Delinquency	95.7	2.2	2.2
Army Self Efficacy	95.6	4.4	0.0
Selflessness	94.4	0.0	5.6
Cooperation	93.5	0.0	6.5
Dominance	92.2	2.0	5.9
Persistence	88.9	0.0	11.1
Even Tempered	84.9	1.9	13.2
Intellectual Efficiency	84.1	4.5	11.4
Courage	78.6	3.6	17.9
Self Control	76.2	7.1	16.7
Commitment to Serve	75.0	13.5	11.5
Virtue	74.0	6.0	20.0
Adjustment	73.6	3.8	22.6
Humility	71.1	6.7	22.2
Optimism	70.2	12.8	17.0
Achievement	70.2	12.3	17.5
Self Efficacy	69.6	2.2	28.3
Responsibility	65.9	2.4	31.7
Situational Awareness	64.6	6.3	29.2
Attention Seeking	61.7	14.9	23.4

Key Findings

- Facets varied in the % of statements translated by SMEs into their target facet, with some facets (e.g., Physical Conditioning) exhibiting perfect retranslation and others (e.g., Attention Seeking) exhibiting relatively poor retranslation (61.7%)

Recommendations for FY25

- Have humans retranslate remainder of statements in pool
- Move statements to proper facet as needed and recalibrate
- Revise statements so they have a clear translation and recalibrate

Note. Statements not flagged for retranslation by the NLP methods for rating by SMEs were considered as translated to their target facet for purposes of the percentages in this table. Facets are sorted in descending order based on the percentage of statements in their pool that was successfully retranslated to their target facet. Cells are color coded to facilitate interpretation. Green/red indicates better/poorer retranslation results.

Bias and Sensitivity Review of Facet Statements

- Purpose
  - Identify TAPAS statements that may be problematic from a bias or sensitivity perspective
- Method
  - Each statement was evaluated by two external SMEs with expertise in bias and sensitivity review (a total of four external SMEs participated in this exercise)
    - Five categories of biased-sensitive language considered (see next slide)
  - Statements flagged by at least one external SME underwent a second round of review by three internal experts who indicated whether statements should be revised or dropped, and reason(s) for doing so

OPA
OFFICE OF PEOPLE ANALYTICS

Bias and Sensitivity Categories

1. **Unfamiliar Term:** The item uses simple, familiar terms that most people can understand and avoids unnecessarily complex or obscure language. For example, if an item is attempting to describe something that is uninteresting, it would be more appropriate to use words such as *boring*, *uninteresting*, or *dull* than to use words such as *jejune*, *pedestrian*, or *humdrum*.
2. **Colloquial:** The item avoids informal and figurative expressions such as colloquialisms ("wicked good"), slang ("nuts"), idioms ("break a leg"), aphorisms ("*when it rains, it pours*"), and technical jargon ("masthead"). The meaning of such terms may not be clear to examinees from a wide variety of backgrounds. Instead, clearly describe the concept of interest in a way that can be reasonably considered comprehensible to all examinees. For example, a more appropriate phrasing of the item "I am easily thrown off" would be "I am easily distracted."
3. **Unfamiliar Situation:** The item avoids situations, contexts, behaviors, and/or other content that will likely not be familiar or accessible to, or feasible for, examinees from a wide variety of cultural, social, and economic backgrounds. For example, "I regularly attend opera performances" would be a less appropriate measure of individuals' artistic interests than "I enjoy listening to classical music."
4. **Controversial Language:** The item avoids language that could be reasonably considered controversial, inflammatory, offensive, insensitive, or otherwise likely to distract examinees by inducing strong emotional reactions (e.g., anger, distress, sadness). This includes avoiding invoking potentially upsetting or controversial topics and concepts (e.g., abortion, colonialism, death, extreme pain, religion, sexuality, violence, illegal activities) explicitly *or* implicitly. For example, the item "I always choose the master bedroom" does not directly concern the controversial topic of slavery, but the historical association of slavery with the term "master bedroom" means it would be inappropriate to phrase the item in this way.
5. **Discrimination**: The item avoids explicit *or* implicit reference to groups that could potentially be discriminated against. Such groups include those related to characteristics such as age, appearance (e.g., attractiveness, height, weight), citizenship status, culture, disability, ethnicity, sex, national or regional origin, native or primary language, political beliefs, race, religion (or its absence), sexual orientation, socioeconomic status. For example, the item "I often feel gypped by my friends" indirectly refers to "gypsy," a derogatory term sometimes applied to the Romani people.

Bias and Sensitivity of Facet Statements

Target Facet	% Fair	% Revise	% Drop	Reason for Revision/Drop				
				% Unfamiliar Term	% Colloquial	% Unfamiliar Situation	% Controversial Language	% Discrimination
Situational Awareness	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Dominance	98.0	2.0	0.0	2.0	0.0	0.0	0.0	2.0
Army Self Efficacy	97.8	2.2	0.0	0.0	2.2	0.0	0.0	0.0
Team Orientation	94.3	5.7	0.0	1.9	3.8	0.0	0.0	1.9
Humility	91.1	8.9	0.0	6.7	0.0	2.2	0.0	2.2
Intellectual Efficiency	90.9	9.1	0.0	2.3	4.5	0.0	0.0	0.0
Selflessness	90.7	9.3	0.0	7.4	5.6	1.9	0.0	3.7
Cooperation	89.1	10.9	0.0	6.5	10.9	0.0	0.0	0.0
Commitment to Serve	88.5	11.5	0.0	1.9	11.5	1.9	0.0	0.0
Virtue	88.0	8.0	4.0	6.0	12.0	2.0	0.0	0.0
Achievement	87.7	7.0	5.3	3.5	10.5	0.0	1.8	0.0
Tolerance	86.4	9.1	4.5	2.3	9.1	9.1	0.0	2.3
Courage	85.7	7.1	7.1	1.8	8.9	1.8	3.6	0.0
Sociability	83.3	16.7	0.0	10.4	16.7	0.0	0.0	0.0
Responsibility	82.9	17.1	0.0	2.4	14.6	4.9	0.0	0.0
Even Tempered	81.1	17.0	1.9	1.9	18.9	0.0	0.0	0.0
Adjustment	81.1	18.9	0.0	5.7	17.0	0.0	0.0	0.0
Self Control	81.0	16.7	2.4	4.8	14.3	2.4	2.4	0.0
Self Efficacy	80.4	19.6	0.0	8.7	10.9	4.3	0.0	2.2
Non-Delinquency	80.4	8.7	10.9	2.2	17.4	2.2	0.0	4.3
Physical Conditioning	79.6	13.0	7.4	1.9	11.1	0.0	5.6	3.7
Optimism	78.7	17.0	4.3	8.5	19.1	0.0	0.0	0.0
Attention Seeking	72.3	23.4	4.3	4.3	27.7	4.3	0.0	0.0
Order	70.0	26.0	4.0	20.0	12.0	6.0	0.0	0.0
Persistence	62.2	37.8	0.0	6.7	28.9	0.0	2.2	0.0

Key Findings

- Almost all facets had statements that were flagged for one or more reasons, though **most flags were related to use of unfamiliar/colloquial terms** rather than use of controversial or discriminatory language

Recommendations for FY25

- Have internal experts review remainder of pool
- Write new statements to replace drops and calibrate
- Revise statements flagged for revision and recalibrate

Note. Facets are sorted in descending order of the percentage of statements in their pool deemed fair by external and internal experts. Green/red indicates higher/lower percentages of facet statements deemed fair. Percentages under Reasons for Revision/Drop columns reflect the percentages of all statements in the facet's statement pool flagged by internal SMEs for the given reason (a statement could be flagged for more than one reason).

Susceptibility of Facet Statements to Transient Error

- Goal
  - Evaluate TAPAS statements for susceptibility to transient error variance
- Method
  - Eight psychologist SMEs independently rated each statement on the following scale:

    "Please rate how much you think applicants' responses to the following statements would be influenced by their psychological/physical state at the time of testing (e.g., based on their mood, how they physically feel, etc.), using a scale of 1 (not at all influenced), 2 (slightly influenced), 3 (moderately influenced), and 4 (very influenced)"

Susceptibility of Facet Statements to Transient Error

Target Facet	Percentage of Statements with Mean Ratings in the Given Range				
	1.0 to 1.5	1.6 to 2.0	2.1 to 2.5	2.6 to 3.5	3.6 to 4.0
Intellectual Efficiency	100.0	0.0	0.0	0.0	0.0
Order	100.0	0.0	0.0	0.0	0.0
Team Orientation	98.1	1.9	0.0	0.0	0.0
Tolerance	97.7	2.3	0.0	0.0	0.0
Attention Seeking	95.7	4.3	0.0	0.0	0.0
Non-Delinquency	94.4	5.6	0.0	0.0	0.0
Selflessness	94.4	5.6	0.0	0.0	0.0
Responsibility	92.6	4.9	2.4	0.0	0.0
Situational Awareness	91.7	9.3	0.0	0.0	0.0
Persistence	88.9	11.1	0.0	0.0	0.0
Self Control	88.1	11.9	0.0	0.0	0.0
Virtue	88.0	12.0	0.0	0.0	0.0
Dominance	86.3	13.7	0.0	0.0	0.0
Courage	83.9	16.1	0.0	0.0	0.0
Sociability	83.3	16.7	0.0	0.0	0.0
Cooperation	82.6	17.4	0.0	0.0	0.0
Achievement	82.5	17.5	0.0	0.0	0.0
Humility	82.2	17.8	0.0	0.0	0.0
Commitment to Serve	73.1	23.1	3.8	0.0	0.0
Physical Conditioning	61.1	37.0	1.0	0.0	0.0
Even Tempered	56.6	35.8	7.5	0.0	0.0
Self Efficacy	50.0	41.3	8.7	0.0	0.0
Adjustment	38.9	37.7	22.6	0.0	0.0
Army Self Efficacy	31.1	60.0	8.8	0.0	0.0
Optimism	19.1	53.2	25.5	2.1	0.0

Key Findings

- Overall, SMEs viewed responses to TAPAS statements as NOT very susceptible to transient error (low mean ratings)
- Statements rated as slightly more susceptible were consistent with expectations, given affective elements associated with those facets (e.g., Optimism, Adjustment, Even Tempered)

Recommendations for FY25

- Revisit/revise statements with ratings 2.0 or greater, if deemed warranted, and recalibrate

Note. Scale points: 1 (not at all influenced), 2 (slightly influenced), 3 (moderately influenced), and 4 (very influenced). Facets are sorted in descending order of the percentage of statements in their pool that had mean ratings in the range of 1.0 to 1.5. Green indicates higher percentages of facet statements were deemed not susceptible to transient error, and red indicates lower percentages of facet statements were deemed not susceptible to transient error.

Revisiting Marginal IRT Reliability of Facet Scores

- Goal
  - Provide updated estimates of marginal IRT reliability of facet scores based on large, current sets of applicant data (or published data when not available)

- Method
  - Evaluated marginal IRT reliability of TAPAS facet scores for TAPAS versions used by Army, USAF, and USMC in 2021–2023 and that were current as of February 2024
  - Based on applicant records where no more than one TAPAS response check item was incorrect
    - Army *n* = 212,726: TAPAS taken between 11/30/21 – 12/1/23
    - USAF *n* = 108,063: TAPAS taken between 6/30/21 – 1/3/24
    - USMC *n* = 82,794: TAPAS taken between 6/27/21 – 12/29/23

OPA
OFFICE OF PEOPLE ANALYTICS

Revisiting Marginal IRT Reliability Estimates of Facet Scores

Facet	*n* TAPAS Versions	Estimate
Physical Conditioning	3	0.76
Sociability	3	0.76
Commitment to Serve*	2	0.75
Order	2	0.69
Dominance	3	0.68
Cooperation	1	0.68
Courage	1	0.67
Adjustment	2	0.65
Selflessness	2	0.65
Intellectual Efficiency	1	0.64
Attention Seeking	2	0.63
Team Orientation	2	0.63
Non-Delinquency	2	0.62
Tolerance	3	0.61
Even Tempered	2	0.61
Responsibility	1	0.60
Achievement	3	0.58
Persistence*	2	0.57
Situational Awareness	1	0.55
Self-Control	1	0.54
Virtue*	1	0.53
Optimism	3	0.49
Self-Efficacy*	1	0.46
Humility*	1	0.40

OPA
OFFICE OF PEOPLE ANALYTICS

Key Findings

- Facets exhibited relatively low to middling reliability (average estimates = .40 –.76) compared to suggested reliability standards for high-stakes testing (e.g., Lance et al., 2006; Nunnally, 1978)[1]
- Low levels of reliability suggest not using individual facet scores for decision making — composites would be more defensible

Recommendations for FY25

- Carefully examine statement pools for low reliability facets during FY25 content development (e.g., evidence of heterogeneity, multiple clear dimensions within a facet) and aim to bolster/refine statement pool for those facets

Note. Reliability estimates reflect simple point estimates or averages across TAPAS versions used by the Army, USAF, and USMC in the 2021–2023 timeframe. Green/red indicates relatively higher/lower levels of reliability. Facets with an asterisk are those for at least one reliability estimate sourced from Drasgow et al (2023) based on experimental Part 2 of the Army TAPAS versions.

[1]Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria. What did they really say? *Organizational Research Methods, 9(2),* 202–220. https://doi.org/10.1177/1094428105284.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Equivalence of Facet Scores Across TAPAS Versions

- Goal
  - *Start to evaluate* the comparability of facet scores from TAPAS versions that differed in their facet composition

- Method
  - Examined comparability of TAPAS facet intercorrelations across versions (e.g., Is the Facet A-B correlation the same across versions?)
  - Examined comparability of TAPAS facet — other variable correlations across versions (e.g., Is the Facet A-AFQT correlation the same across versions?)
  - Examined seven different versions of Army TAPAS used at MEPS over time that partially overlapped in their facet composition

A Note on Research Design Limitations

- Considered multiple potential approaches to examining equivalence…most of which were not feasible within the study timeframe
  - Administer multiple TAPAS versions to same respondents with facet composition systematically varied
  - Implement multigroup CFA based approaches to studying measurement invariance
    - Issues with applying factor analysis to partially ipsative data
  - IRT/item based-approaches
    - Examining item equivalence or understanding differences in scores based on items administered
  - Simulation based approaches
    - Identifying true thetas and running them through different TAPAS versions and to see how the observed thetas for a facet varied across versions
- Given time and feasibility — we adopted a simpler (albeit more limited) approach that focused only on similarity of TAPAS facet intercorrelations and TAPAS facet—other variable correlations (AFQT, 6-month attrition, 24-month attrition) across TAPAS versions that differed in their facet composition

Equivalence of Facet Scores Across TAPAS Versions

Key Findings

- TAPAS facet intercorrelations and TAPAS facet—other variable correlations were generally quite similar across versions, indicating facet mix may NOT have notable impact on a target facet's measurement
- When differences were found, they tended to be for TAPAS facet intercorrelations between TAPAS versions from different Army TAPAS development "stages"
  - Stage 2 (least use of cleaning/quality flags) → Stage 4 (most use of cleaning/quality flags)
  - Average absolute differences between same-facet correlations across versions
    - .014 WITHIN stages
    - .054 (Stage 2 vs. Stage 3 versions) and .047 (Stage 2 vs. Stage 4 versions)
- Between-stage differences in facet-intercorrelations didn't translate into differences in TAPAS facet-AFQT and TAPAS facet-attrition correlations

OPA
OFFICE OF PEOPLE ANALYTICS

Composite Shortening Analyses

- Goal
  - Evaluate the possibility of shortening preliminary Phase 1 MC and ENL composites
- Method
  - Performed best subsets regression using preliminary Phase 1 MC and ENL composites as criteria (separate models for each criterion) and the facets that contribute to those composites as initial predictors
    - Regressions based on facet intercorrelation matrices developed during the FY23 research
  - Identified what facets were consistently retained in models as the number of features in the predictor subset was reduced and the Multiple *R* achieved by those reduced models
- Key Findings
  - There appears to be room to shorten the preliminary Phase 1 MC and ENL composites and still achieve a very high correlation with the full versions of those composites

OPA
OFFICE OF PEOPLE ANALYTICS

Finalizing the Phase 1 JS TAPAS Instrument Design

OPA
OFFICE OF PEOPLE ANALYTICS

JS Instrument Design

- Only a limited number of facets can be administered as part of the JS TAPAS due to testing time constraints at MEPS and the cognitive load associated with the use of more facets
- Tradeoff between "number of facets" and "number of statements per facet"
  - More facets mean more flexibility to cover JS MC and ENL composites and Service-specific uses
  - Due to testing time constraints, more facets also means fewer items per facet, resulting in less reliable measurement
  - Greater number of statements per facet → higher marginal IRT reliability for facets
    - Drasgow et al (2023)[1] suggests 20 statements per facet
- Targeting no more than 17 facets for the JS TAPAS instrument — key decision point was how many facets to reserve for the Joint-Service facets vs. Service-specific facets

[1]Drasgow, F., Chernyshenko, O. S., Stark, S., Nye, C. D. (2023). *Tailored Adaptive Personality Assessment System (TAPAS): Pre-implementation documentation*. (AFRL-RH-WP-TR-2023-0014). Air Force Research Laboratory.

OPA
OFFICE OF PEOPLE ANALYTICS

Key Considerations for Identifying Joint-Service Facets

Facet-level considerations

- Use in, and importance to, preliminary FY23 recommendations for Phase 1 JS composites
- Use in, and importance to, Service-specific models/composites
- Performance along FY24 research metrics (i.e., retranslation, bias/sensitivity, IRT marginal reliability, transient error)
  - Deficiencies here have the potential to be addressed via subsequent FY25 development work
- Secondary consideration — relevance to outcomes perceived to be of broad interest across Services
  - First-term attrition
  - Enlisted leadership potential/emergence

Set-level considerations

- Balance in terms of personality construct mix
- More JS facets (better prediction of JS criteria + construct coverage) ← vs → fewer JS facets (more Service-specific slots)
- More facets overall (fewer statements per facet = lower facet reliability) ← vs → fewer facets overall (more statements per facet = higher facet reliability)

Finalizing the Set of JS Facets

- SMEs from HumRRO, DCG, and DTAC reviewed information for each facet given the facet-level and set-level considerations and developed recommendations for potential sets of facets to include in the Joint-Service set
- Goal was to identify a single set of facets that could be used to support scoring of refined Phase 1 JS MC and ENL composites
  - Different facets from the set would be used to score each composite OR all may be used for each composite but differentially weighted — **TBD during FY25 development work**
- Reviewed considerations, research findings, and recommendations with Service representatives and came to consensus on a set of 12 JS facets that would be included in the Phase 1 JS TAPAS
  - 5 additional facet "slots" reserved for Service-specific facets

OPA
OFFICE OF PEOPLE ANALYTICS

Next Steps for JS TAPAS

OPA
OFFICE OF PEOPLE ANALYTICS

Operations and Maintenance (O&M) Track (FY25)

- Preparing for implementation of Phase 1 composites
  - Statement pool development
    - Needs identified in FY24 research
  - Existing statement re-calibration and new statement calibration using a joint-Service sample
  - Finalizing composition and weighting of Phase 1 composites
  - Development of provisional joint-Service norms for facets
- IT work to enable implementation at MEPS sometime in FY27 (TBD)

OPA
OFFICE OF PEOPLE ANALYTICS

R&D Track (FY25–FY26)

R&D to evaluate/enhance TAPAS adjacent to the Phase 1 JS TAPAS O&M work

- Effects of practice and coaching on TAPAS
  - Practice effects on TAPAS
  - Coaching/large language model (LLM) informed response
- AI for non-cognitive assessment
  - Review potential role of AI in bringing efficiencies to non-cognitive assessment (e.g., statement development, statement parameter estimation)
- TAPAS and supervised machine learning (ML) for attrition prediction
  - Exploration of input/features below the facet level

OPA
OFFICE OF PEOPLE ANALYTICS

Questions for the DAC

OPA
OFFICE OF PEOPLE ANALYTICS

Questions for the DAC

1. Should we hold TAPAS and cognitive test scores used for high-stakes decision making to different reliability standards? What's the minimum level of reliability you believe is acceptable for defending use of TAPAS composite scores for making high-stakes selection decisions?

2. Narrowing the construct we aim to cover with a TAPAS facet can help ensure unidimensionality of a facet's statement pool (which should help with reliability issues), but doing so would make it harder to develop a statement pool of sufficient size for use in TAPAS. Thoughts on strategies to deal with this tradeoff?

3. If our research finds TAPAS is susceptible to coaching effects (e.g., elevation of scores on particular facets), what suggestions do you have for mitigating such effects?

Thank you!

For more information please contact:

Dan Putka
dputka@humrro.org
703.706.5640

OPA
OFFICE OF PEOPLE ANALYTICS

Tab P

OPA

OFFICE OF PEOPLE ANALYTICS

Adverse Impact of the ASVAB and Special Tests

Findings from the FY2023 Applicant Sample

Nicholas Howald

Human Resources Research Organization

Briefing presented to the DACMPT

January 23, 2025

Agenda

- Adverse Impact Background
- AI Analysis Findings
- Conclusions

OPA
OFFICE OF PEOPLE ANALYTICS

Adverse Impact Background

OPA
OFFICE OF PEOPLE ANALYTICS

What Is Adverse Impact?

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

Fairness and Adverse Impact

- Adverse impact does **not** mean a test is biased
- Evidence for validity and fairness of the Armed Services Vocational Aptitude Battery (ASVAB):
  - There is extensive evidence supporting the validity of AFQT scores in selection (Thacker et al., 2020)
  - A study by Putka et al. (2022) using five years of applicant data showed a lack of differential prediction for the AFQT in the vast majority of analyses
  - Item writers are given sensitivity and bias guidelines, and multiple HumRRO and DTAC editors review for these factors (Harber & Day, 2023)
  - Items are pretested and any that are flagged for differential item functioning (DIF) are reviewed by experts for evidence of bias; if biased content is found, the item is not used operationally (Reeder, 2023)

How Is Adverse Impact Assessed?

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

  "A selection rate for any race, sex, or ethnic group, which is less than four-fifths (80%) 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)***:

$$IR = \frac{SR_{Focal}}{SR_{Reference}}$$, where SR is the selection ratio

- Ideally, IR = 1, but the four-fifths rule leaves wiggle room

OPA
OFFICE OF PEOPLE ANALYTICS

How Is Adverse Impact Assessed?

- Statistical significance of the IR can be computed, as well as confidence intervals around the IR (Morris & Lobsenz, 2000):

$$Z_{IR} = \frac{\ln\frac{SR_{Foc}}{SR_{Ref}}}{\sqrt{\frac{1-SR_{Tot}}{SR_{Tot}}\left(\frac{1}{N_{Foc}}+\frac{1}{N_{Ref}}\right)}}$$, where SR = selection rate

- Z_{IR} is significant at α = .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

Adverse Impact Analyses for the ASVAB

- The four-fifths rule (80%) and accompanying statistics are applied to the Armed Forces Qualification Test (AFQT) by comparing qualification rates across the focal and reference groups of interest regarding:
  - 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)
  - AI is assessed using **initial test scores *only***
  - Significance testing is not necessarily useful in analyses with very large numbers of applicants (i.e., > 2,000)
- How should we assess AI for individual ASVAB and Special Tests, where no direct selection occurs?

Potential for Adverse Impact

- **Effect sizes (ES)** — standardized mean differences, commonly Cohen's *d*
  - ES can be plotted and classified with respect to Cohen's (1988) standards of evaluation
    Small ≥ 0.20; **Moderate** ≥ 0.50; **Large** ≥ 0.80
- Effect sizes are computed for all group comparisons as:

$$ES = \frac{\mu_{Reference} - \mu_{Focal}}{\sigma_p}$$

where:

$\mu_{Reference}$ is the mean score in the Reference (Majority) group.

μ_{Focal} is the mean score in the Focal (Minority) group

σ_p is the pooled standard deviation across the two groups

Note: Positive values indicate the impact favors the majority group (i.e., the minority group is impacted negatively).

Confidence Intervals around Effect Sizes

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

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

where:

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

- Confidence intervals provide a boundary around an ES point estimate
  - Small boundaries indicate a more precise ES estimate
  - Large boundaries indicate a more variable ES estimate

FY2023 AI Analyses Findings

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB Tests and Special Tests on ASVAB Platform

- ASVAB: Multiple-aptitude battery that measures developed abilities and helps predict future academic and occupational success in the military (all Services).
- Special Tests: Not part of the ASVAB but delivered on the ASVAB platform, developed to inform Service-specific classification efforts

ASVAB Tests

AFQT			
Verbal	**Math**	**Science/Technical**	**Spatial**
Paragraph Comprehension (PC)	Arithmetic Reasoning (AR)	General Science (GS)	Assembling Objects (AO)
Word Knowledge (WK)	Math Knowledge (MK)	Electronics Information (EI)	
		Mechanical Comprehension (MC)	
		Auto Information/ Shop Information (AS)	

Special Tests

Cyber Test (CT)	Coding Speed (CS)
Test of basic computer and information systems knowledge (all Services)	A speeded test of assigning code numbers to words (Navy only)

Current ASVAB AI Analyses

- Sample: FY2023 applicants
- Tests: ASVAB AFQT (IIIA+ and IIIB+), ASVAB Subtests, Cyber Test, and Coding Speed
- Group comparisons:

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
5	Non-Hispanic Whites	Non-White Hispanics

- The focal group is potentially disadvantaged relative to the reference group.
- All included groups represent > 2% of the applicant population.

OPA
OFFICE OF PEOPLE ANALYTICS

Current ASVAB AI Analyses

Data Cleaning

- Initial test record with valid score, name, and SSN
  - **ASVAB:** 248,434 (Total); 163,132 (CAT); 1,003 (P&P); 84,299 (Verified PiCAT)
  - **CT:** 60,230
  - **CS:** 41,145
- Remove duplicates across assessments (ASVAB ONLY): 247,779 (*n* = 655 removed)
- Timing: >2.5 SD below mean response time
  - **ASVAB:** 247,754 (*n* = 25 removed); **CT:** 60,210 (*n* = 20 removed)
- Timing: < 2 minutes to complete assessment
  - **CS:** 40,984 (*n* = 161 removed)
- Missing on all demographic variables (i.e., sex, race, and ethnicity)
  - **ASVAB:** 241,412 (*n* = 6,342 removed)
  - **CT:** 49,681 (*n* = 10,529 removed)
  - **CS:** 39,213 (*n* = 1,771 removed)

Current ASVAB AI Analyses

Sample Category	ASVAB *N*	ASVAB Percent	Cyber Test *N*	Cyber Test Percent	Coding Speed *N*	Coding Speed Percent
Males	183,108	76%	38,710	78%	28,597	73%
Females	58,304	24%	10,971	22%	10,616	27%
Non-Hispanic Whites	92,151	38%	19,492	39%	13,335	34%
Hispanic Whites	58,622	24%	12,182	25%	8,459	22%
Non-Hispanic Blacks	62,416	26%	10,580	21%	10,693	27%
Non-Hispanic Asians	12,836	5%	2,882	6%	2,702	7%
Non-White Hispanics	6,553	3%	1,846	4%	1,623	4%
Total	**241,412**	-	**49,681**	-	**39,213**	-

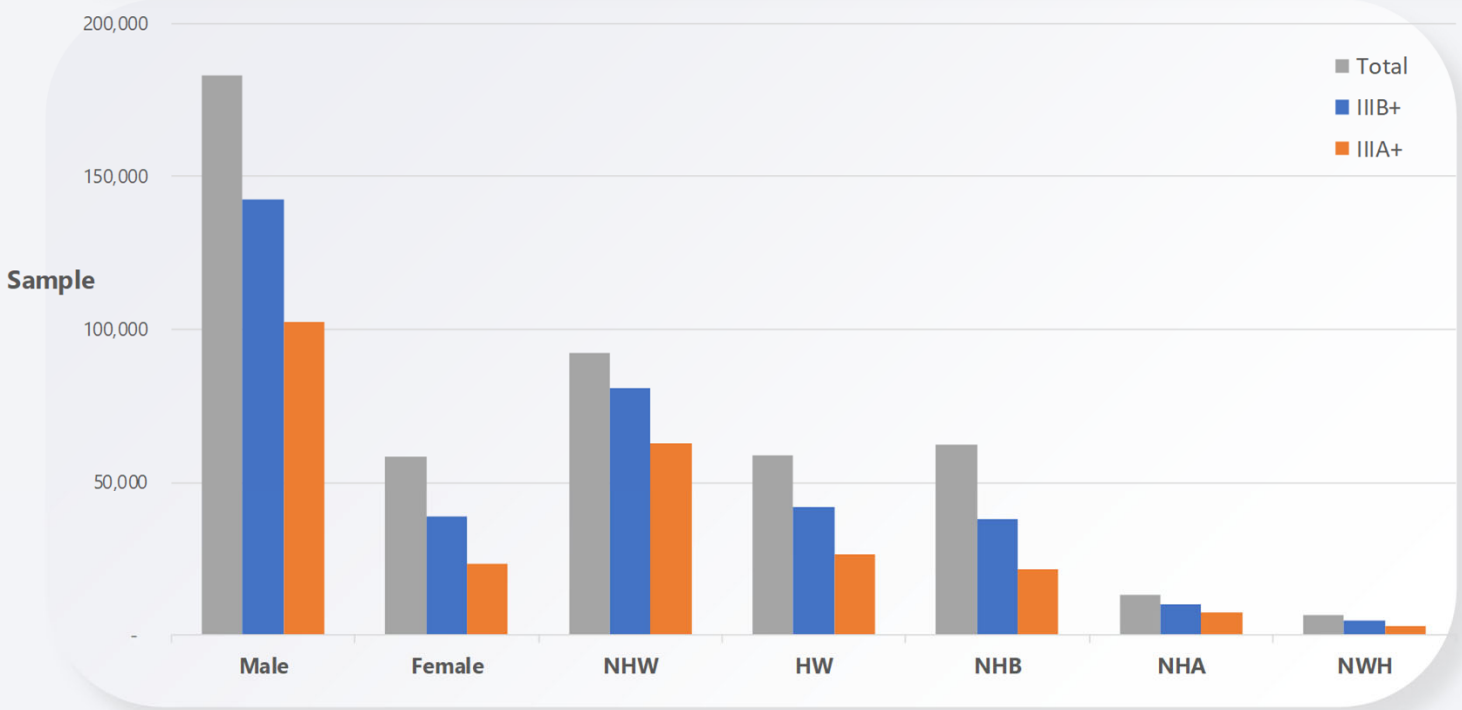
Note. Ethnicity does not add up to 100% due to missing data or other sample category values below the 2% threshold for some individuals. Some individuals in CT (*n* = 2,221; 4%) and CS (*n* = 5,193; 13%) did not have corresponding data in the ASVAB sample, as they had taken ASVAB during FY22 (Oct 1, 2021 – Sep 30, 2022).

OPA
OFFICE OF PEOPLE ANALYTICS

Adverse Impact Analysis Findings

OPA
OFFICE OF PEOPLE ANALYTICS

Adverse Impact Analysis Sample Sizes for FY23



NHW: Non-Hispanic White
HW: Hispanic White
NHB: Non-Hispanic Black
NHA: Non-Hispanic Asian
NWH: Non-White Hispanic

OPA
OFFICE OF PEOPLE ANALYTICS

Impact Ratios for IIIB+ and IIIA+

OPA
OFFICE OF PEOPLE ANALYTICS

Impact Ratios for AFQT Cutscores FY2023 IIIB+ and IIIA+

Adverse Impact Ratios

Threshold: IIIA+ | **IIIB+**

Comparison	Impact Ratio
Female vs Male	**0.85**
HW vs NHW	**0.81**
NHB vs NHW	**0.69**
NHA vs NHW	**0.88**
NWH vs NHW	**0.83**

Trivial | **Small** | **Mod** | **Large**

NHW: Non-Hispanic White
HW: Hispanic White
NHB: Non-Hispanic Black
NHA: Non-Hispanic Asian
NWH: Non-White Hispanic

OPA
OFFICE OF PEOPLE ANALYTICS

Impact Ratios for AFQT Cutscores FY2023 IIIB+ and IIIA+

Adverse Impact Ratios

Threshold

IIIA+ | IIIB+

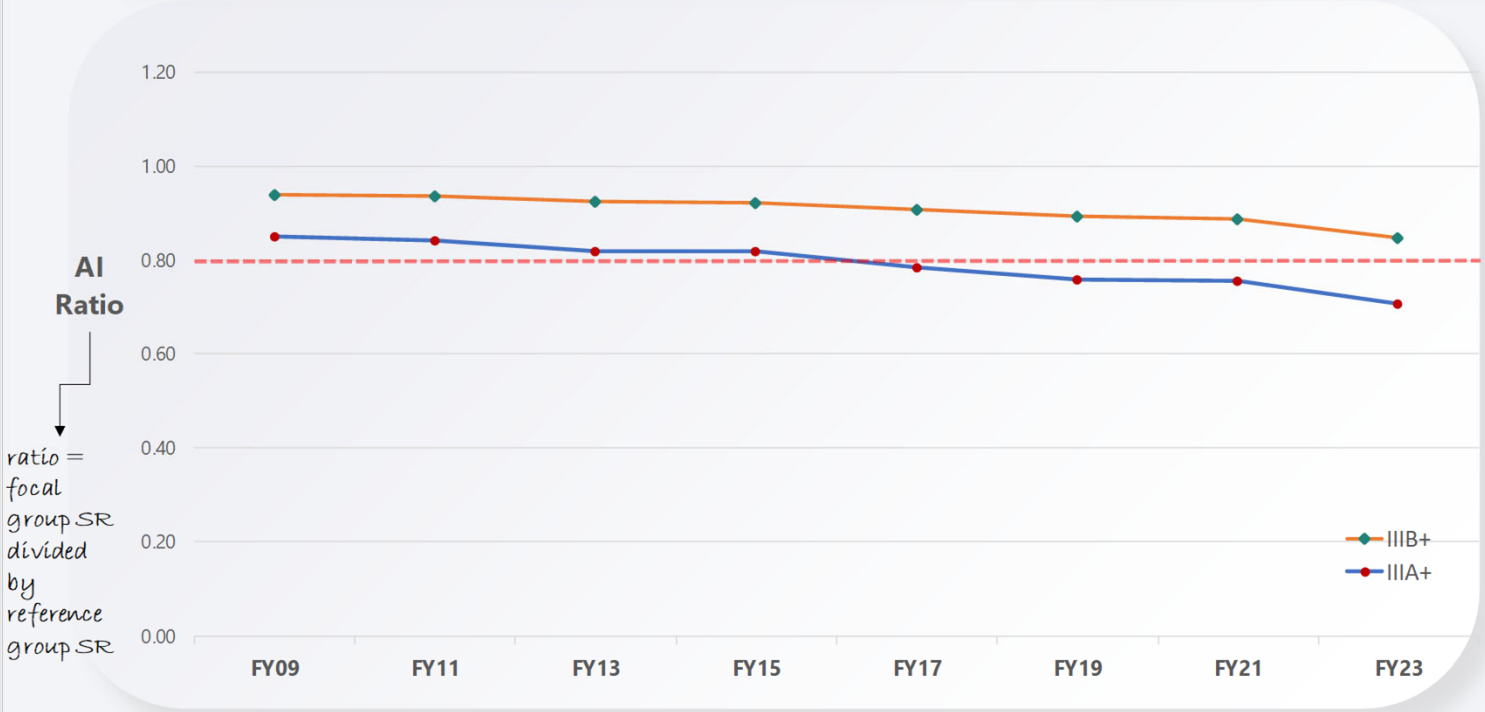
Comparison	Impact Ratio
Female vs Male	0.71
HW vs NHW	0.66
NHB vs NHW	0.50
NHA vs NHW	0.86
NWH vs NHW	0.69

Trivial | Small | Mod | Large

NHW: Non-Hispanic White
HW: Hispanic White
NHB: Non-Hispanic Black
NHA: Non-Hispanic Asian
NWH: Non-White Hispanic

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Impact Ratios for Odd-Numbered FYs 09-23



Adverse Impact Ratios

Comparison

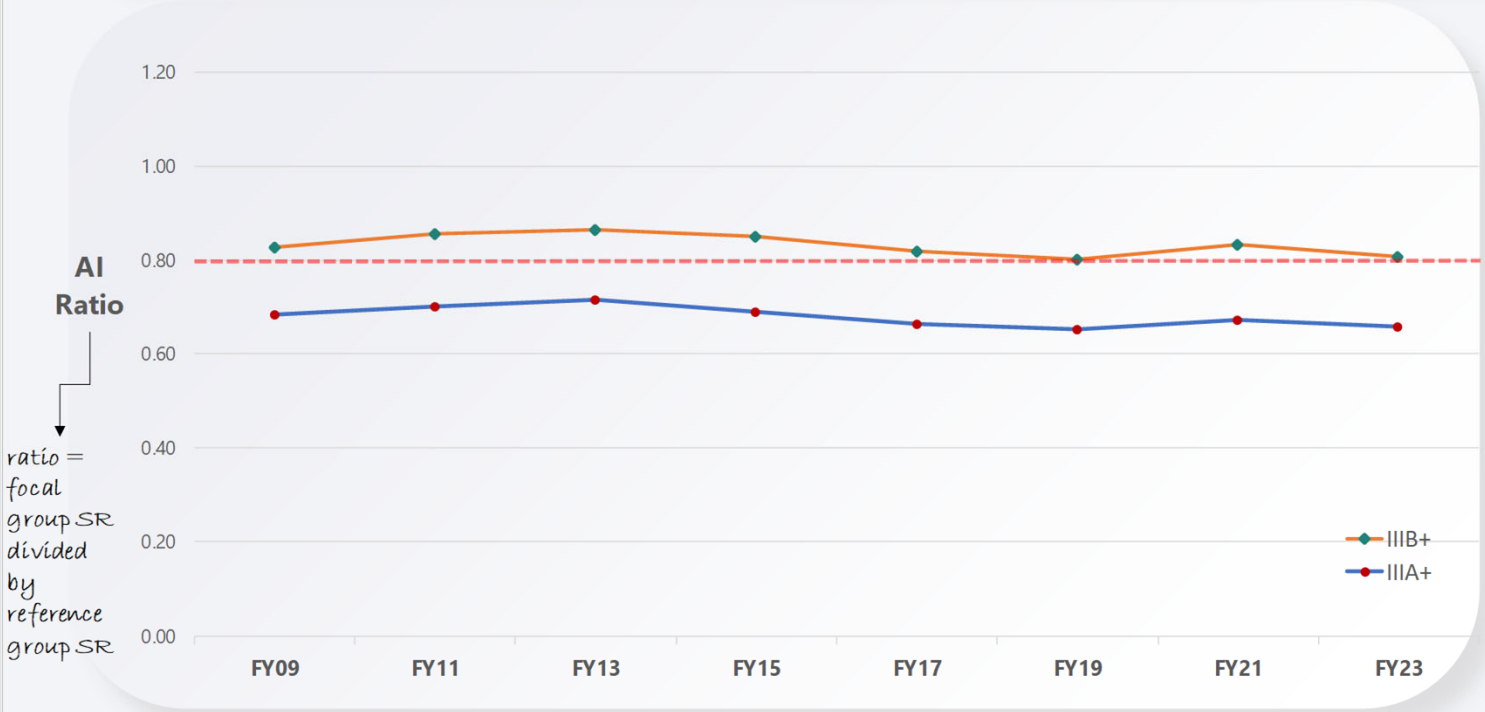
Female vs Male	HW vs NHW
NHA vs NHW	NHB vs NHW

Year	IIIB+	IIIA+
FY09	0.94	0.85
FY11	0.94	0.84
FY13	0.93	0.82
FY15	0.92	0.82
FY17	0.91	0.79
FY19	0.89	0.76
FY21	0.89	0.75
FY23	0.85	0.71

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Impact Ratios for Odd-Numbered FYs 09-23



Adverse Impact Ratios

Comparison

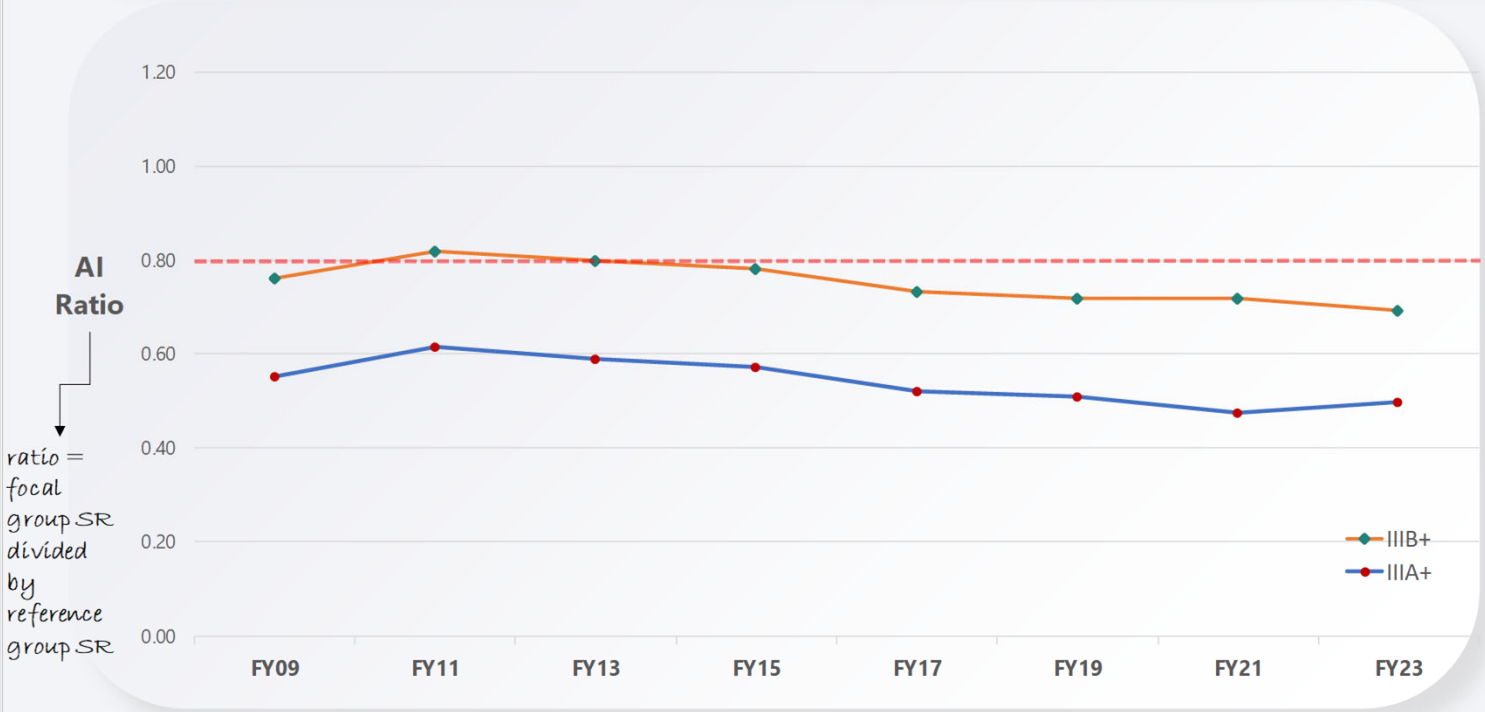
Female vs Male	**HW vs NHW**
NHA vs NHW	NHB vs NHW

Year	IIIB+	IIIA+
FY09	**0.83**	**0.68**
FY11	**0.86**	**0.70**
FY13	**0.87**	**0.72**
FY15	**0.85**	**0.69**
FY17	**0.82**	**0.66**
FY19	**0.80**	**0.65**
FY21	**0.83**	**0.67**
FY23	**0.81**	**0.66**

Trivial **Small** **Mod** **Large**

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Impact Ratios for Odd-Numbered FYs 09-23



Adverse Impact Ratios

Comparison

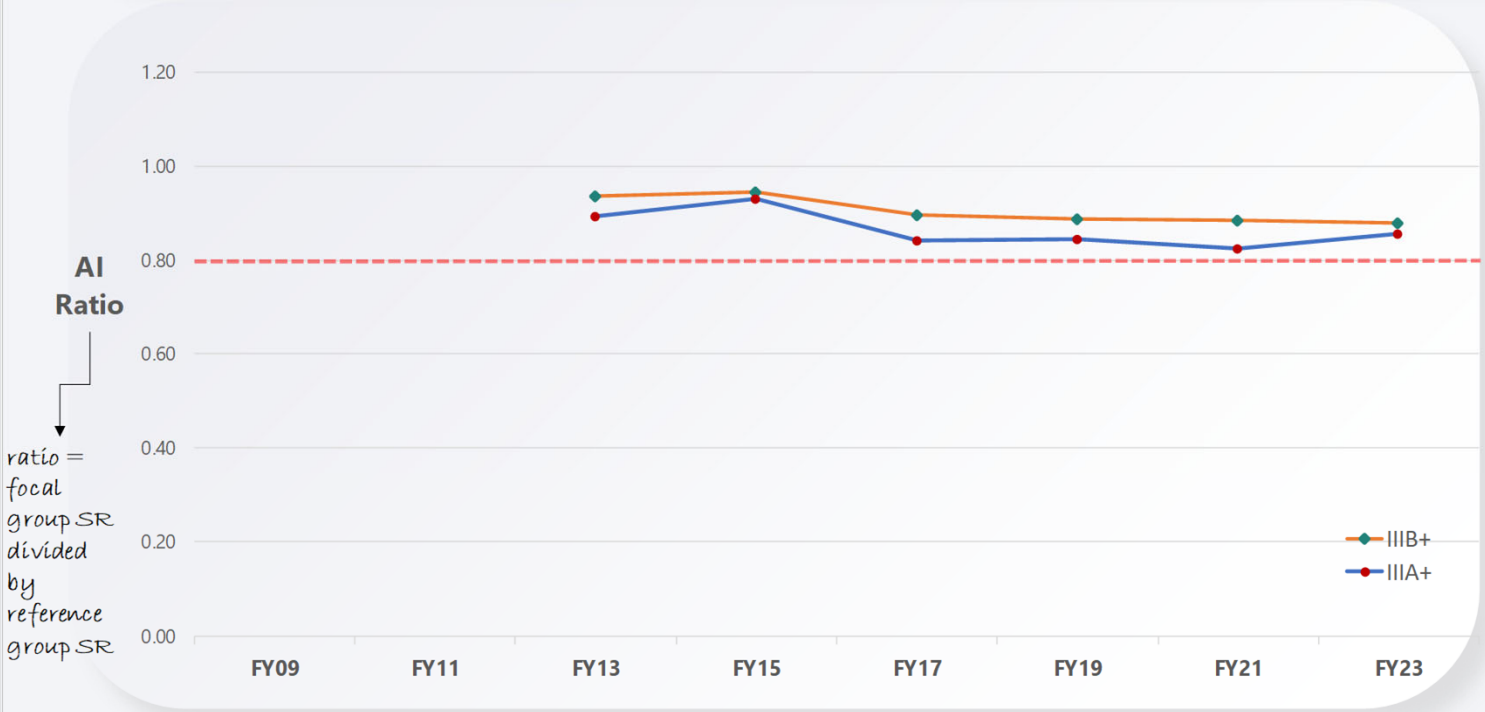
Female vs Male	HW vs NHW
NHA vs NHW	**NHB vs NHW**

Year	IIIB+	IIIA+
FY09	**0.76**	**0.55**
FY11	**0.82**	**0.62**
FY13	**0.80**	**0.59**
FY15	**0.78**	**0.57**
FY17	**0.73**	**0.52**
FY19	**0.72**	**0.51**
FY21	**0.72**	**0.47**
FY23	**0.69**	**0.50**

Trivial **Small** **Mod** **Large**

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Impact Ratios for Odd-Numbered FYs 09-23



Adverse Impact Ratios

Comparison

Female vs Male	HW vs NHW
NHA vs NHW	NHB vs NHW

Year	IIIB+	IIIA+
FY09		
FY11		
FY13	**0.93**	**0.89**
FY15	**0.95**	**0.93**
FY17	**0.90**	**0.84**
FY19	**0.89**	**0.84**
FY21	**0.88**	**0.82**
FY23	**0.88**	**0.86**

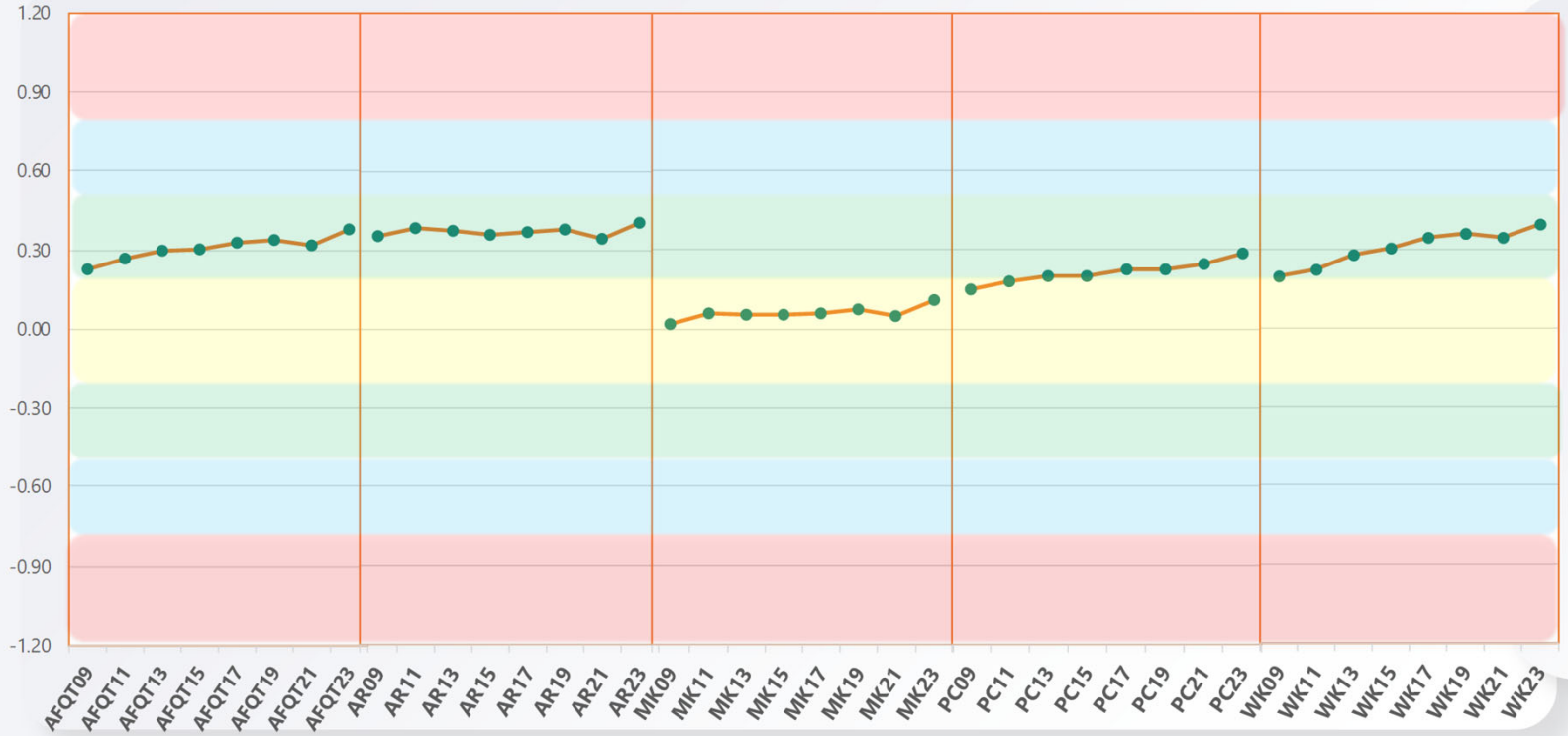
Trivial **Small** **Mod** **Large**

OPA
OFFICE OF PEOPLE ANALYTICS

Effect Sizes over Time

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes for Odd-Numbered FY 09-23 (AFQT Tests/Scores)



Effect Sizes

Comparison

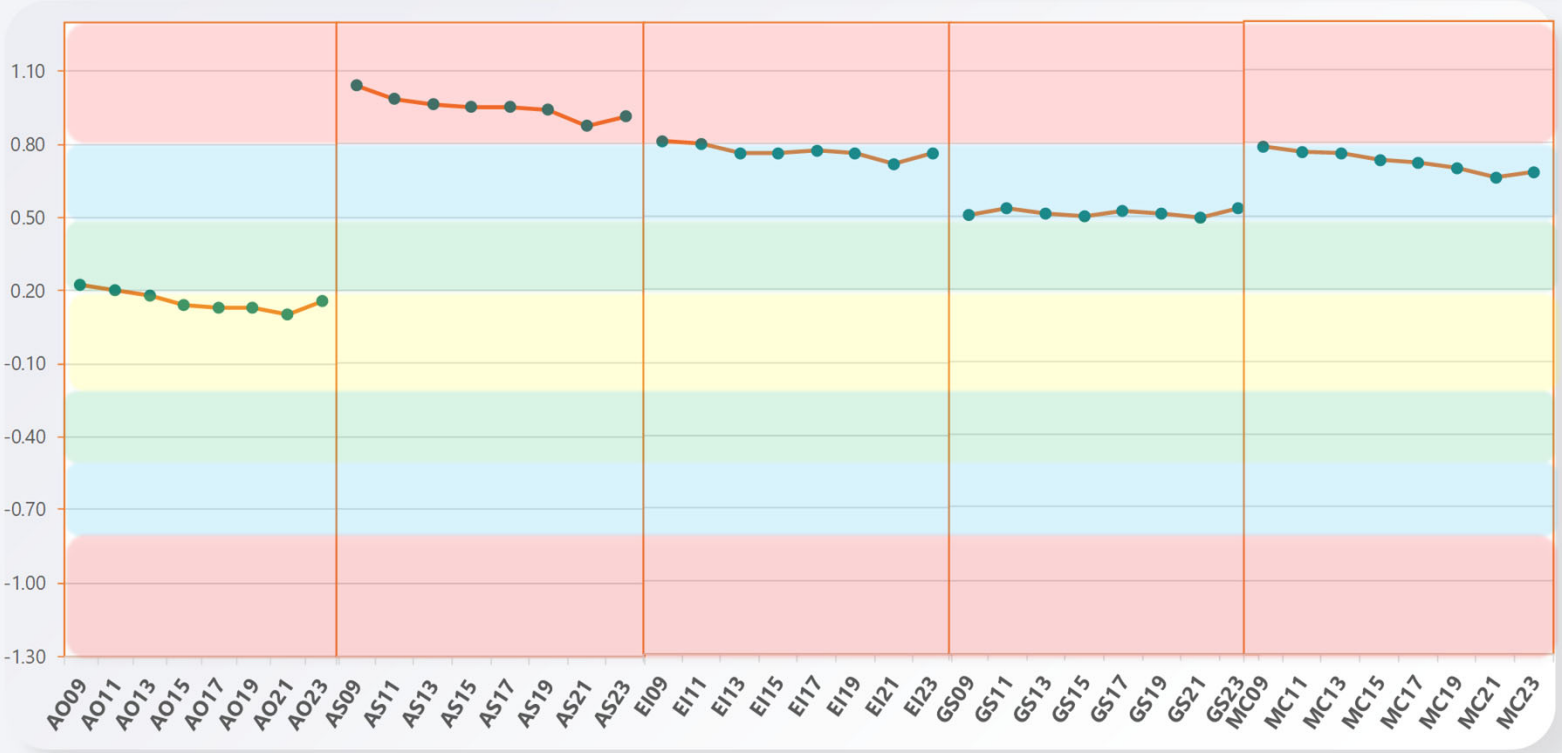
Female vs Male	HW vs NHW
NHA vs NHW	NHB vs NHW

Year	AFQT	AR	MK	PC	WK
FY09	0.23	0.35	0.02	0.15	0.20
FY11	0.27	0.38	0.06	0.18	0.23
FY13	0.30	0.37	0.05	0.20	0.28
FY15	0.30	0.36	0.05	0.20	0.31
FY17	0.33	0.37	0.06	0.23	0.35
FY19	0.34	0.38	0.07	0.23	0.36
FY21	0.32	0.34	0.05	0.25	0.35
FY23	0.38	0.40	0.11	0.29	0.40

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes for Odd-Numbered FY 09-23 (non-AFQT Tests)



Effect Sizes

Comparison

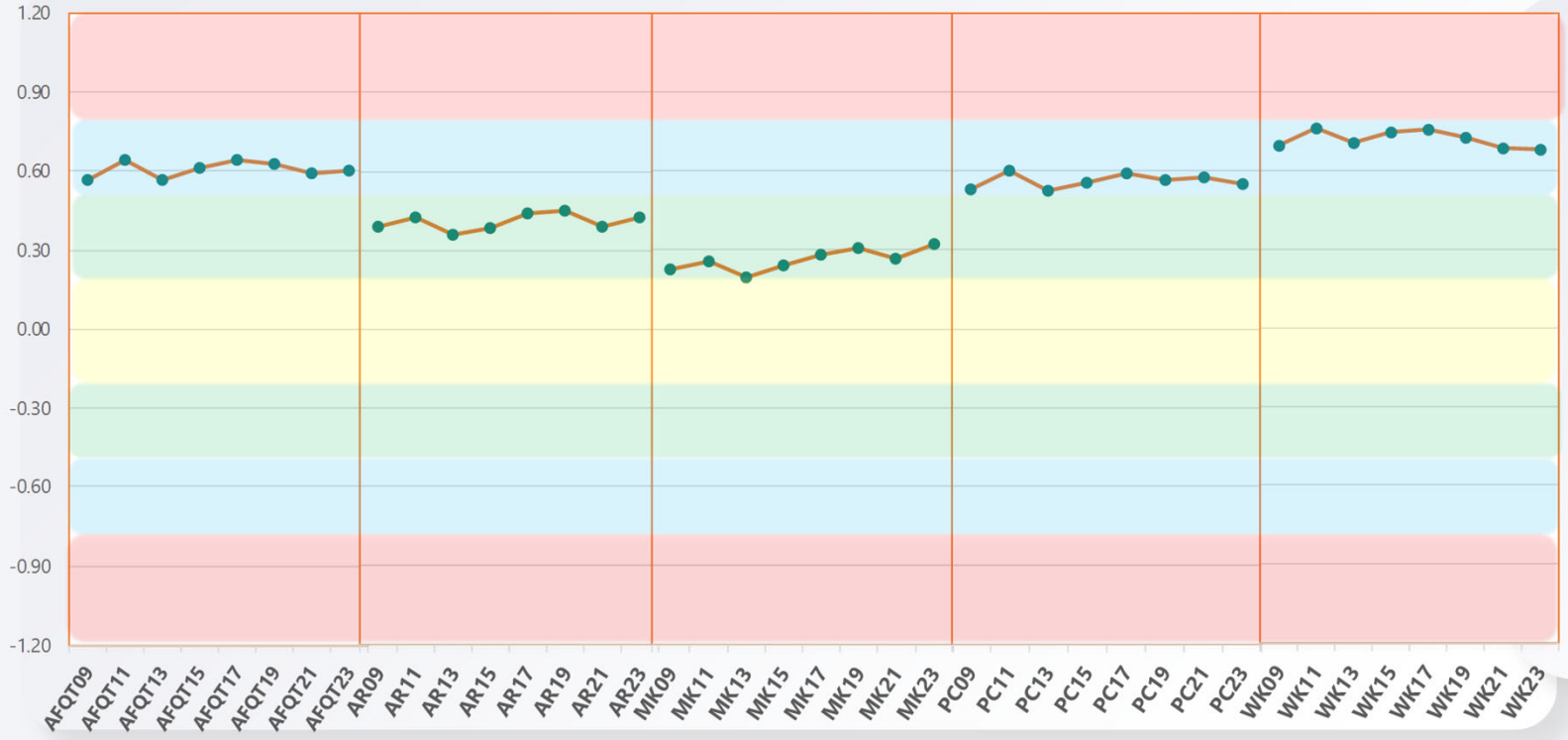
Female vs Male | HW vs NHW | NHA vs NHW | NHB vs NHW

Year	AO	AS	EI	GS	MC
FY09	0.22	1.04	0.81	0.51	0.79
FY11	0.20	0.98	0.80	0.53	0.77
FY13	0.18	0.96	0.76	0.51	0.76
FY15	0.14	0.95	0.76	0.50	0.73
FY17	0.13	0.95	0.77	0.53	0.72
FY19	0.13	0.94	0.76	0.51	0.70
FY21	0.10	0.87	0.72	0.50	0.66
FY23	0.16	0.91	0.76	0.53	0.68

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes for Odd-Numbered FY 09-23 (AFQT Tests/Scores)



Effect Sizes

Comparison

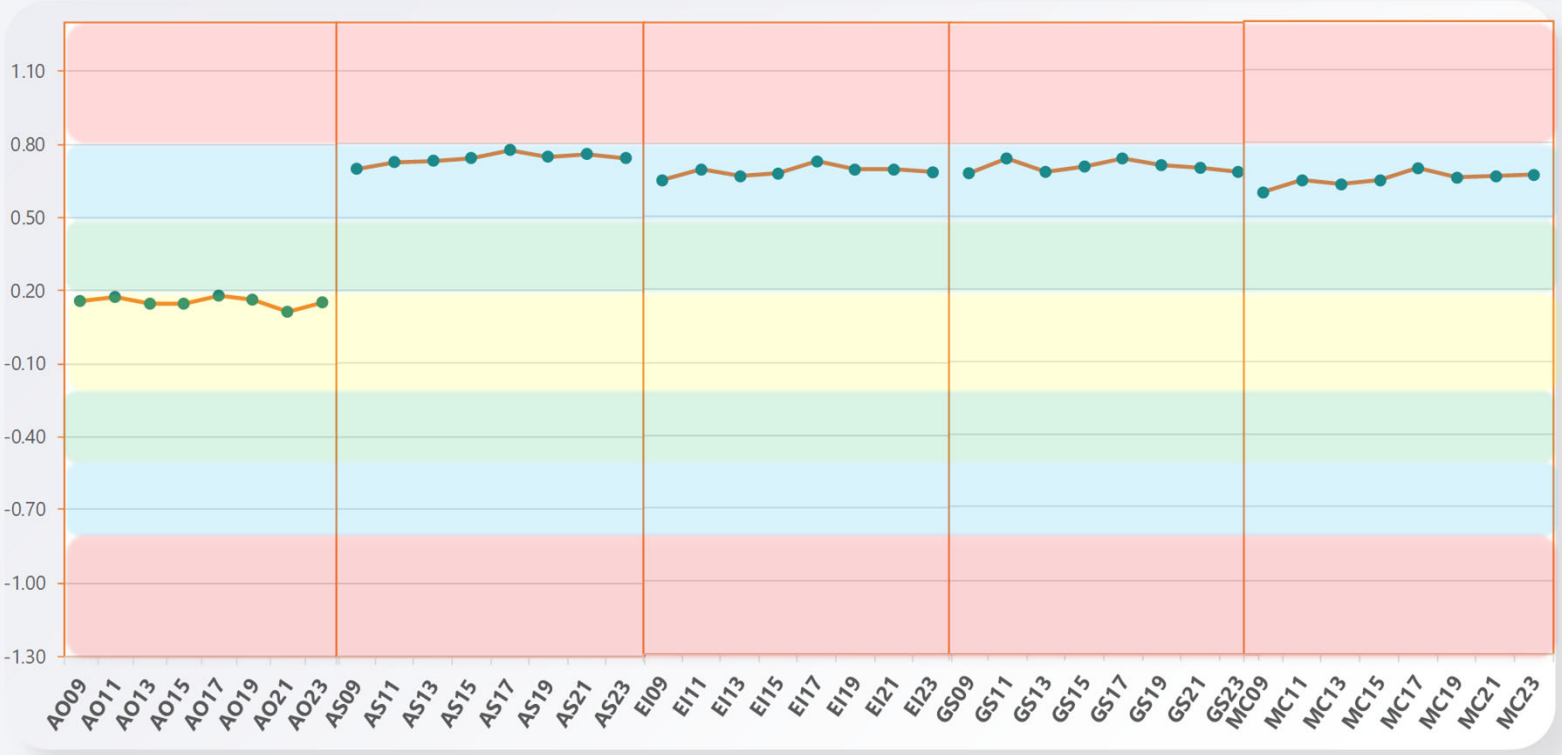
Female vs Male	**HW vs NHW**
NHA vs NHW	NHB vs NHW

Year	AFQT	AR	MK	PC	WK
FY09	0.57	0.39	0.23	0.53	0.70
FY11	0.64	0.42	0.25	0.60	0.76
FY13	0.57	0.36	0.20	0.53	0.71
FY15	0.61	0.39	0.24	0.56	0.75
FY17	0.64	0.44	0.28	0.59	0.76
FY19	0.63	0.45	0.30	0.57	0.73
FY21	0.59	0.39	0.26	0.58	0.69
FY23	0.60	0.42	0.32	0.55	0.68

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes for Odd-Numbered FY 09-23 (non-AFQT Tests)



Effect Sizes

Comparison

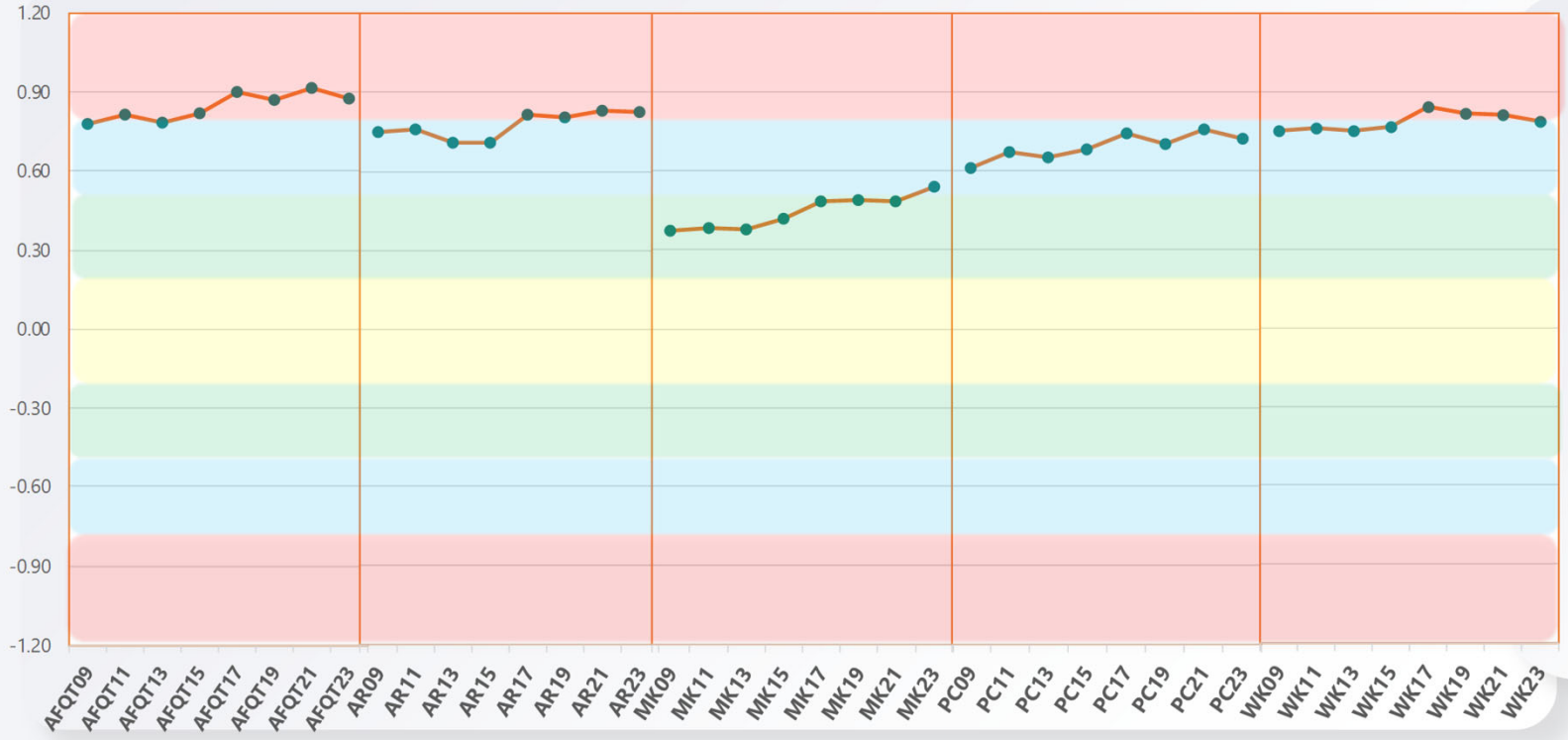
Female vs Male	**HW vs NHW**
NHA vs NHW	NHB vs NHW

Year	AO	AS	EI	GS	MC
FY09	0.16	0.70	0.65	0.68	0.60
FY11	0.17	0.73	0.70	0.74	0.65
FY13	0.14	0.73	0.67	0.68	0.63
FY15	0.15	0.74	0.68	0.71	0.65
FY17	0.18	0.77	0.73	0.74	0.70
FY19	0.16	0.75	0.70	0.71	0.66
FY21	0.11	0.76	0.70	0.70	0.66
FY23	0.15	0.74	0.68	0.68	0.67

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes for Odd-Numbered FY 09-23 (AFQT Tests/Scores)



Effect Sizes

Comparison

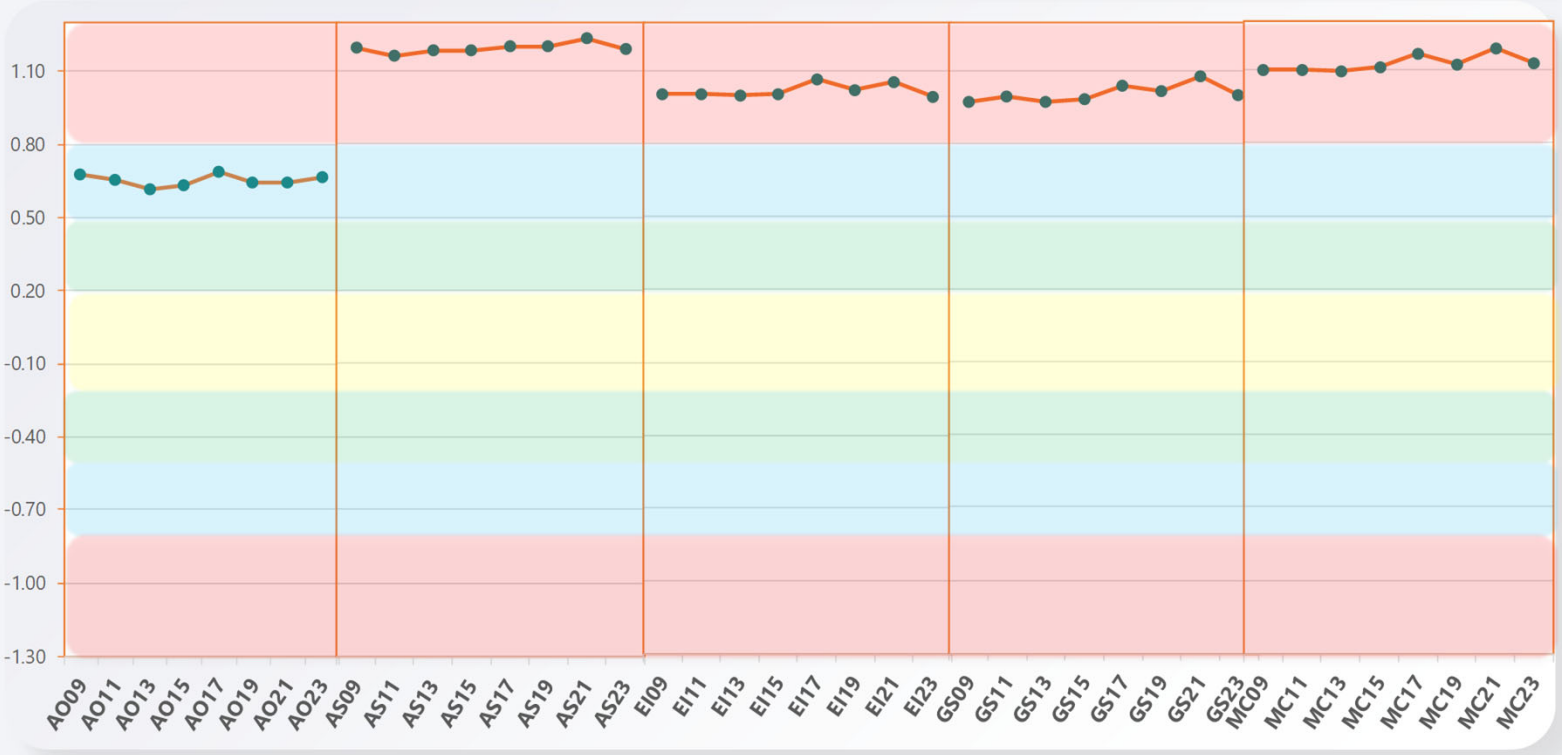
Female vs Male	HW vs NHW
NHA vs NHW	**NHB vs NHW**

Year	AFQT	AR	MK	PC	WK
FY09	0.78	0.75	0.38	0.61	0.75
FY11	0.82	0.76	0.38	0.67	0.76
FY13	0.78	0.71	0.38	0.65	0.76
FY15	0.82	0.71	0.42	0.68	0.77
FY17	0.90	0.82	0.49	0.75	0.85
FY19	0.87	0.80	0.49	0.70	0.82
FY21	0.91	0.83	0.48	0.76	0.81
FY23	0.87	0.83	0.54	0.73	0.79

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes for Odd-Numbered FY 09-23 (non-AFQT Tests)



Effect Sizes

Comparison

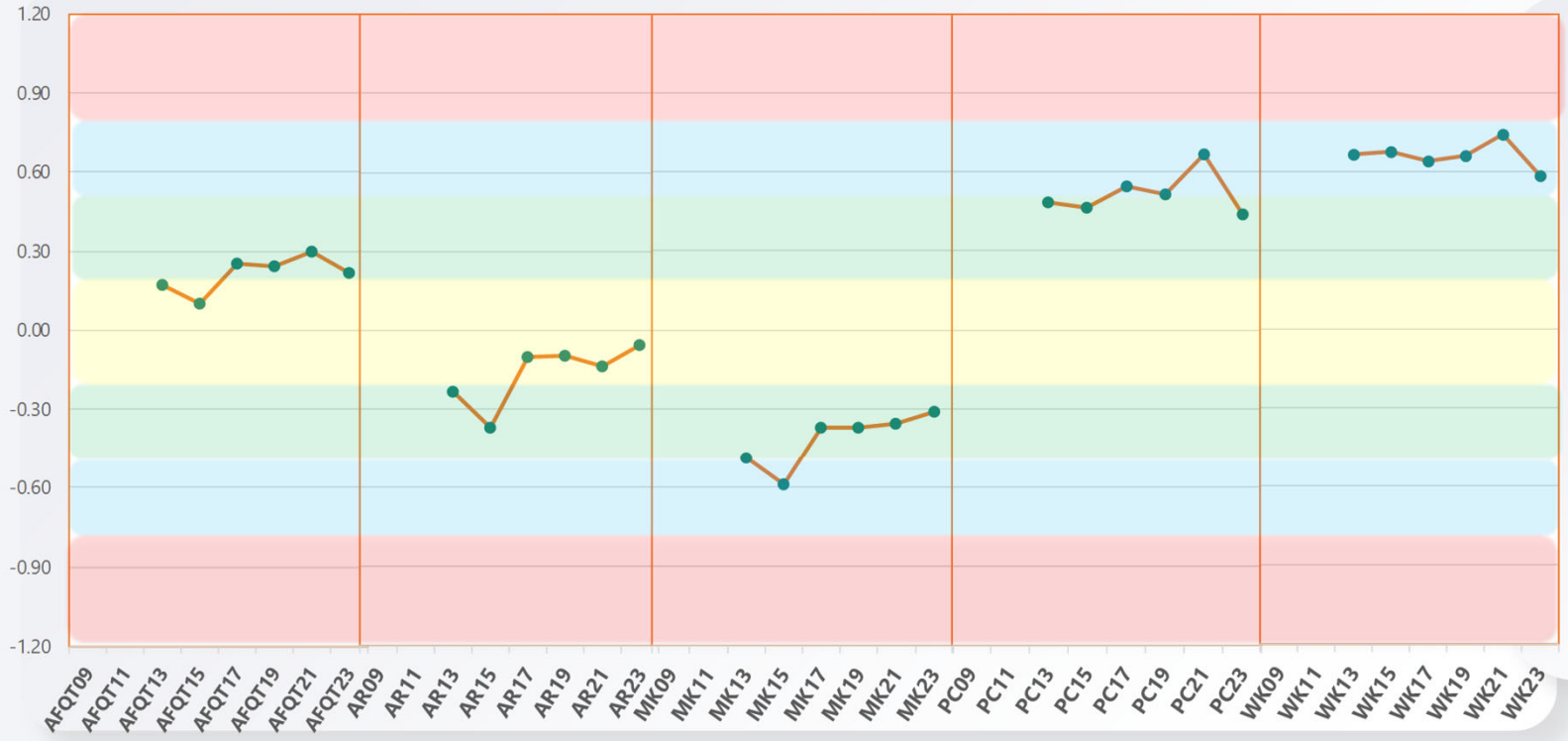
Female vs Male | HW vs NHW | NHA vs NHW | **NHB vs NHW**

Year	AO	AS	EI	GS	MC
FY09	0.68	1.19	1.01	0.97	1.10
FY11	0.65	1.16	1.01	0.99	1.10
FY13	0.62	1.18	1.00	0.97	1.10
FY15	0.63	1.18	1.01	0.98	1.11
FY17	0.69	1.20	1.07	1.04	1.17
FY19	0.64	1.20	1.02	1.02	1.13
FY21	0.64	1.23	1.06	1.08	1.19
FY23	0.66	1.19	1.00	1.00	1.13

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes for Odd-Numbered FY 09-23 (AFQT Tests/Scores)



Effect Sizes

Comparison

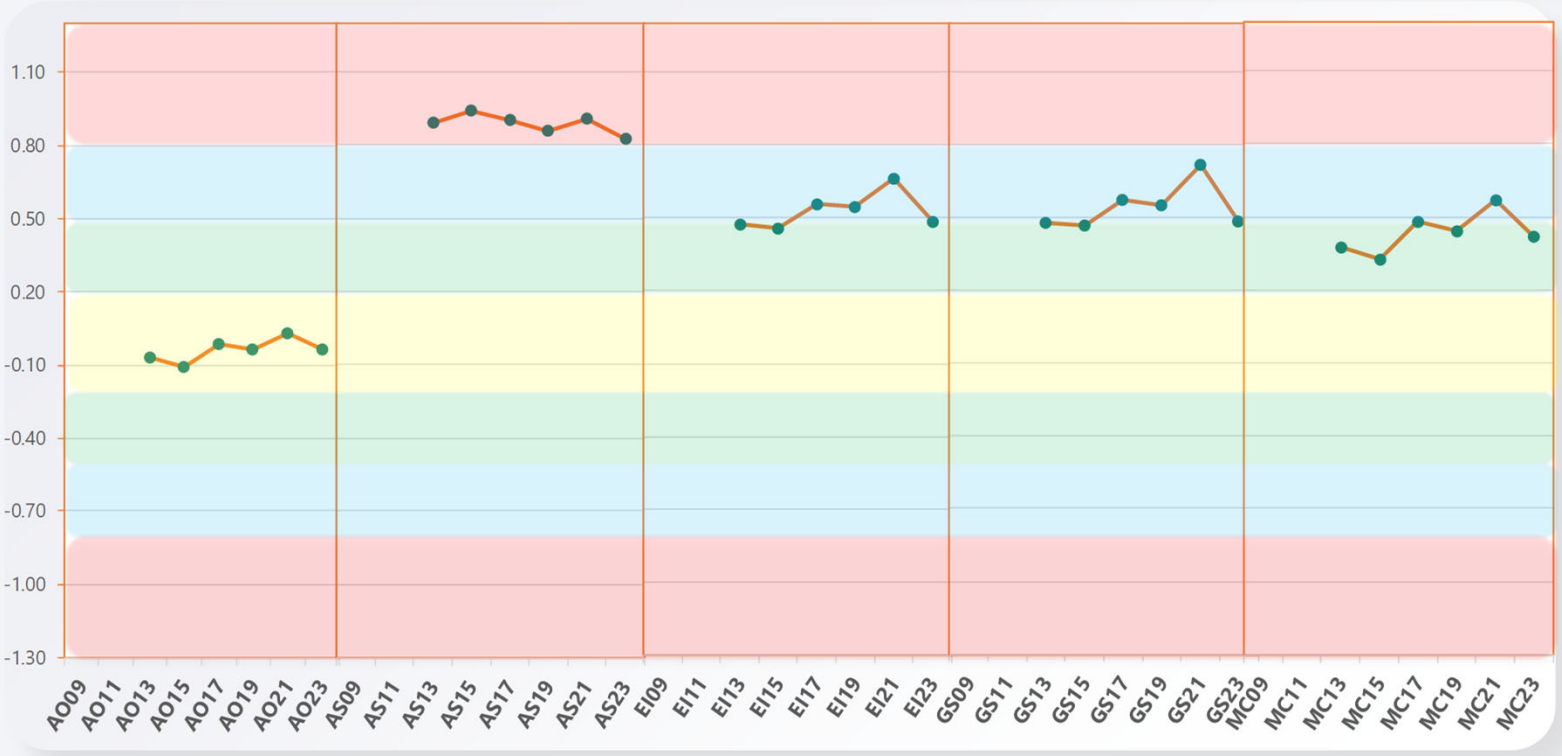
Female vs Male	HW vs NHW
NHA vs NHW	NHB vs NHW

Year	AFQT	AR	MK	PC	WK
FY09					
FY11					
FY13	0.17	-0.24	-0.49	0.49	0.67
FY15	0.10	-0.37	-0.59	0.46	0.68
FY17	0.25	-0.10	-0.37	0.55	0.64
FY19	0.24	-0.10	-0.37	0.51	0.66
FY21	0.30	-0.14	-0.36	0.67	0.74
FY23	0.21	-0.06	-0.31	0.44	0.58

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes for Odd-Numbered FY 09-23 (non-AFQT Tests)



Effect Sizes

Comparison

Female vs Male	HW vs NHW
NHA vs NHW	NHB vs NHW

Year	AO	AS	EI	GS	MC
FY09					
FY11					
FY13	-0.07	0.89	0.48	0.48	0.38
FY15	-0.11	0.94	0.46	0.47	0.33
FY17	-0.01	0.90	0.56	0.57	0.48
FY19	-0.04	0.86	0.55	0.55	0.44
FY21	0.03	0.91	0.66	0.72	0.57
FY23	-0.04	0.83	0.49	0.49	0.42

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison with Other Large-Scale Testing Programs

OPA
OFFICE OF PEOPLE ANALYTICS

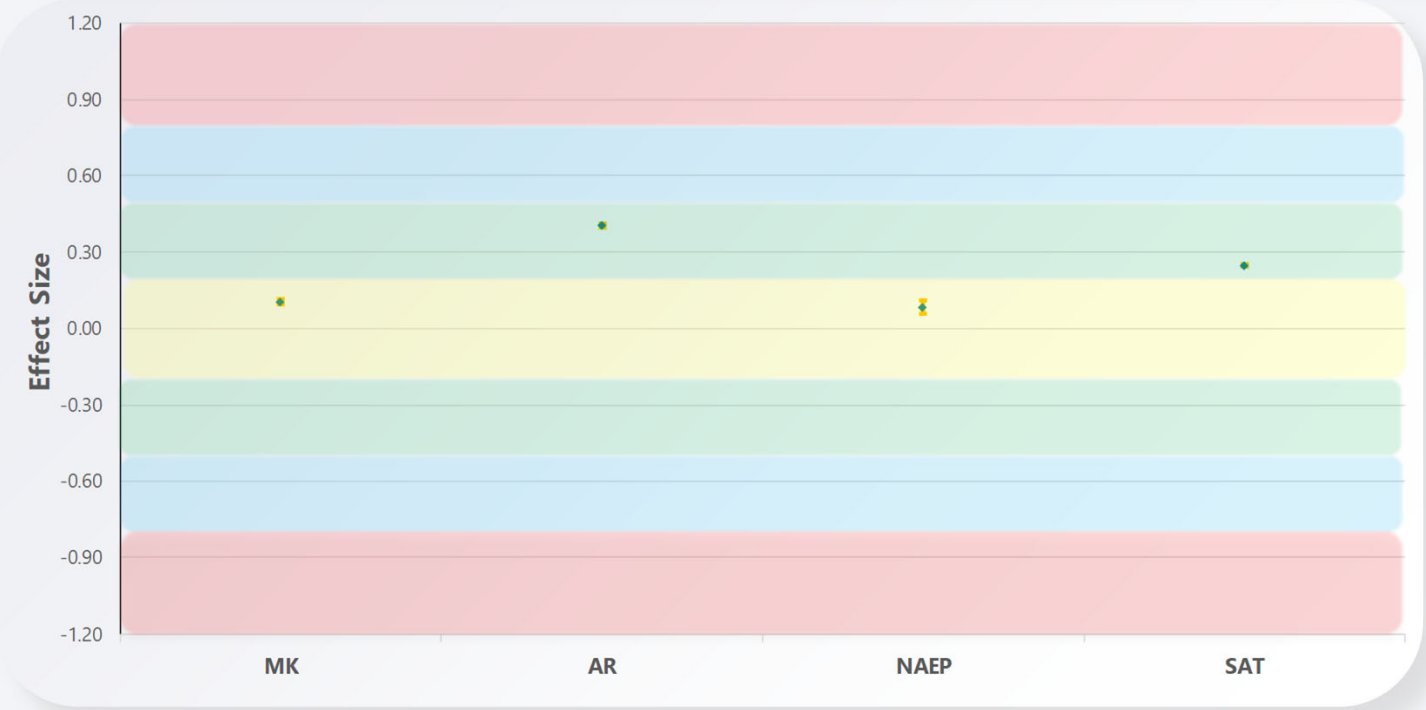
What Does It Mean?

- The magnitude of impact on the ASVAB has remained fairly consistent across fiscal years, but still varies in size from negligible to large across tests and groups.
- A comparison of impact across different testing programs gives some indication of whether the observed FY2023 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 — 2019 Grade 12 (Reading, Math, and Science)

*SAT stopped reporting SDs for demographic comparisons after 2016 in publicly available online content, limiting the ability to calculate effect sizes for more recent years without submitting data requests.

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs (Math)



Effect Sizes

Comparison

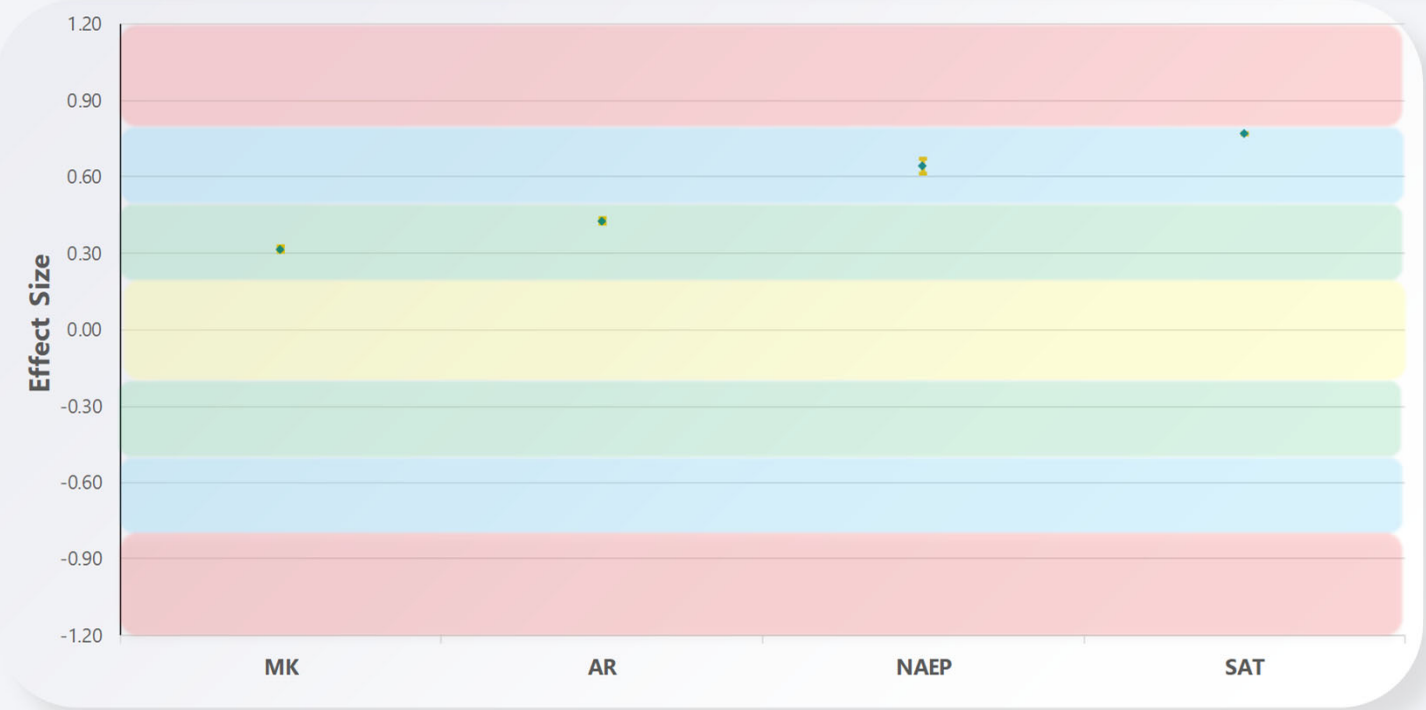
- **Female vs Male**
- H vs NHW
- NHA vs NHW
- NHB vs NHW

Test	Effect Size
MK	**0.11**
AR	**0.40**
NAEP	**0.08**
SAT	**0.25**

Trivial **Small** **Mod** **Large**

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs (Math)



Effect Sizes

Comparison

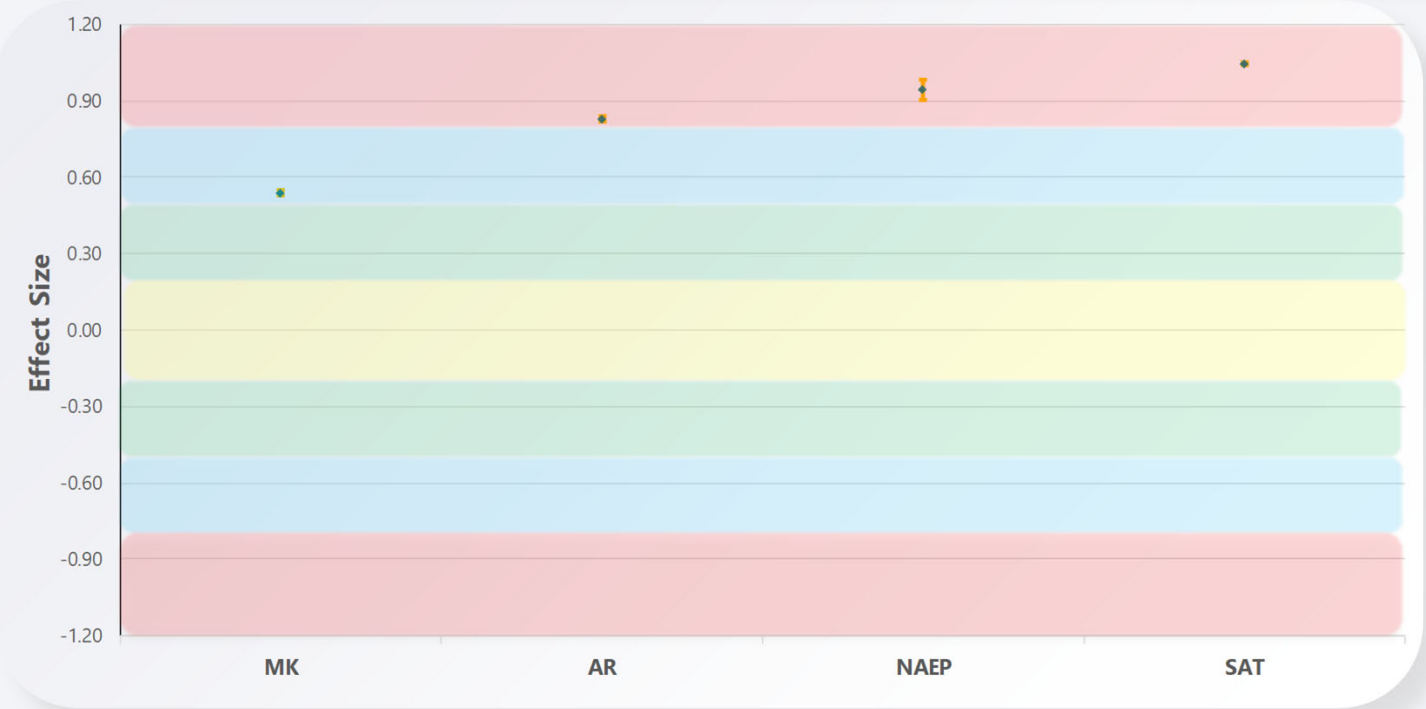
- Female vs Male
- **H vs NHW**
- NHA vs NHW
- NHB vs NHW

Test	Effect Size
MK	**0.32**
AR	**0.43**
NAEP	**0.64**
SAT	**0.77**

Trivial **Small** **Mod** **Large**

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs (Math)



Effect Sizes

Comparison

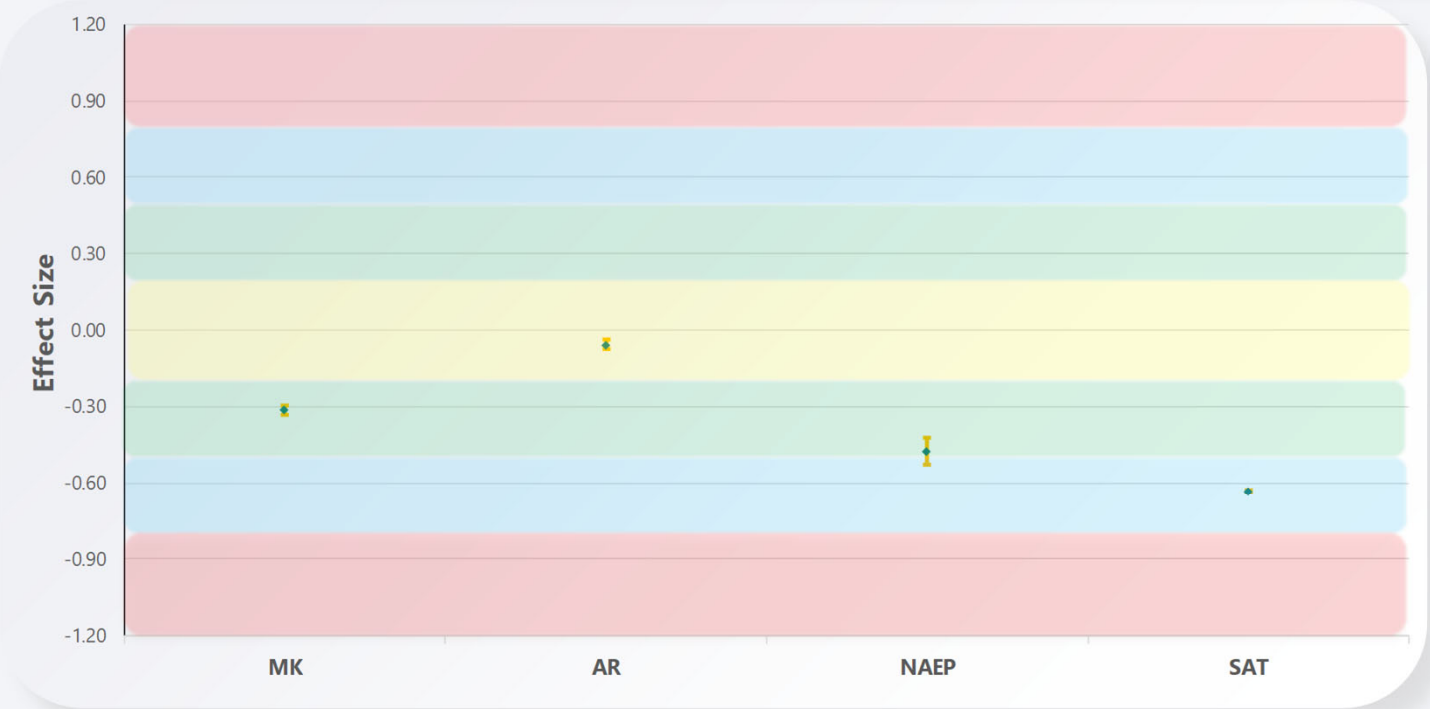
- Female vs Male
- H vs NHW
- NHA vs NHW
- **NHB vs NHW**

Test	Effect Size
MK	**0.54**
AR	**0.83**
NAEP	**0.95**
SAT	**1.04**

Trivial **Small** **Mod** **Large**

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs (Math)



Effect Sizes

Comparison

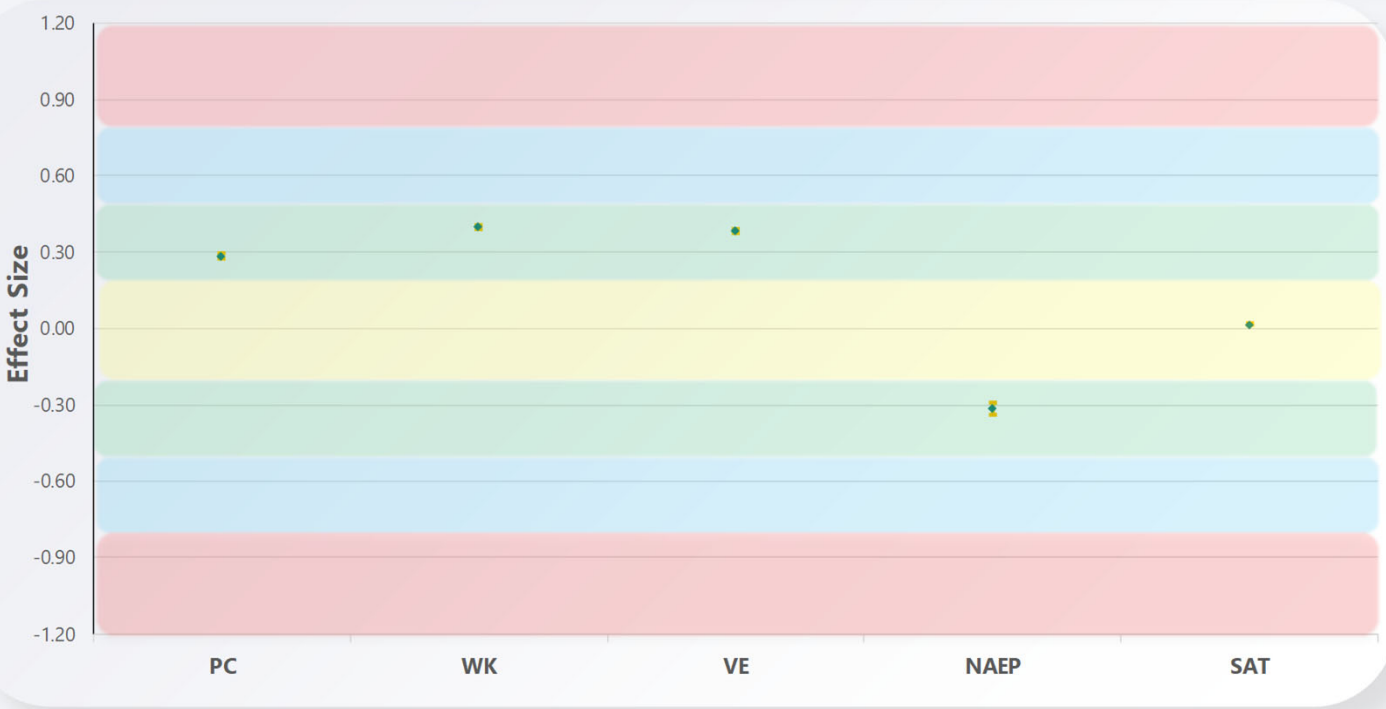
- Female vs Male
- H vs NHW
- **NHA vs NHW**
- NHB vs NHW

Test	Effect Size
MK	-0.31
AR	-0.06
NAEP	-0.48
SAT	-0.63

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs (Reading)



Effect Sizes

Comparison

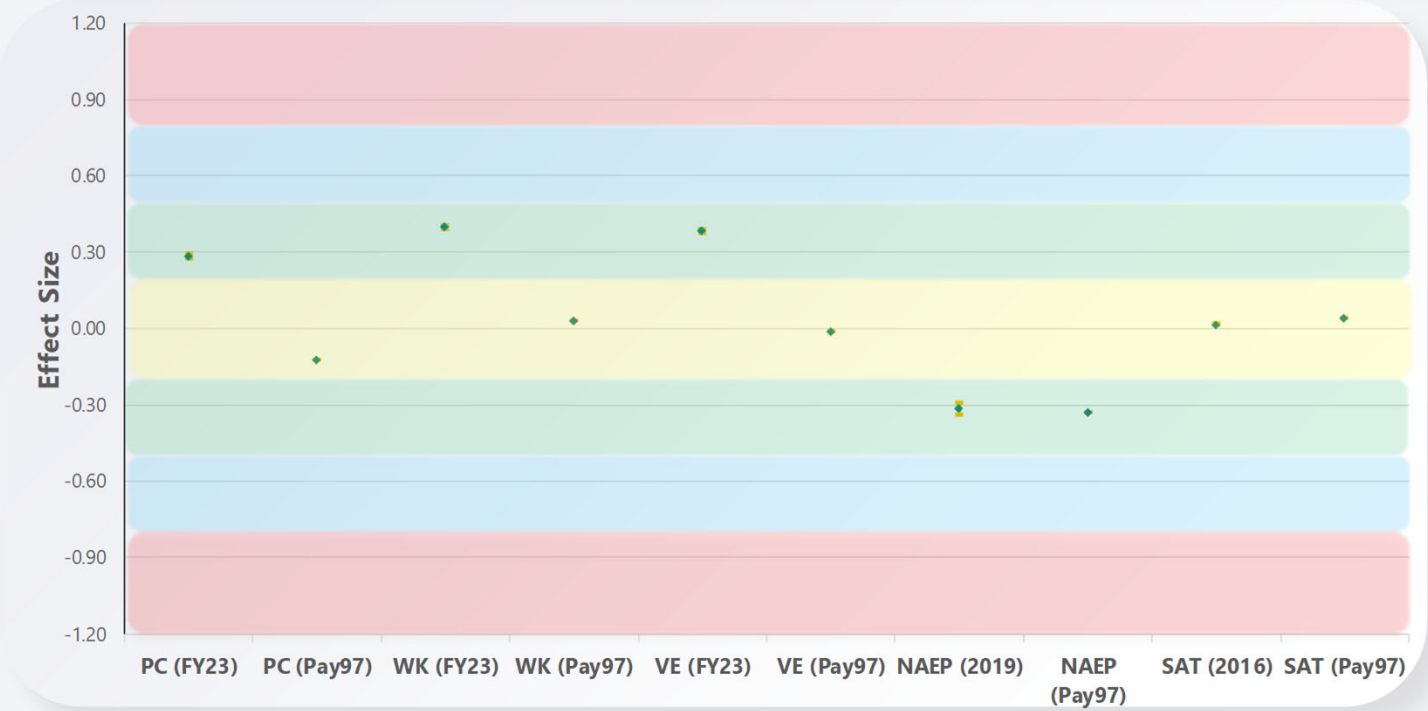
- **Female vs Male**
- H vs NHW
- NHA vs NHW
- NHB vs NHW

Test	Effect Size
PC	0.29
WK	0.40
VE	0.38
NAEP	-0.31
SAT	0.02

Trivial **Small** **Mod** **Large**

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs (Reading/Verbal) Female vs. Male



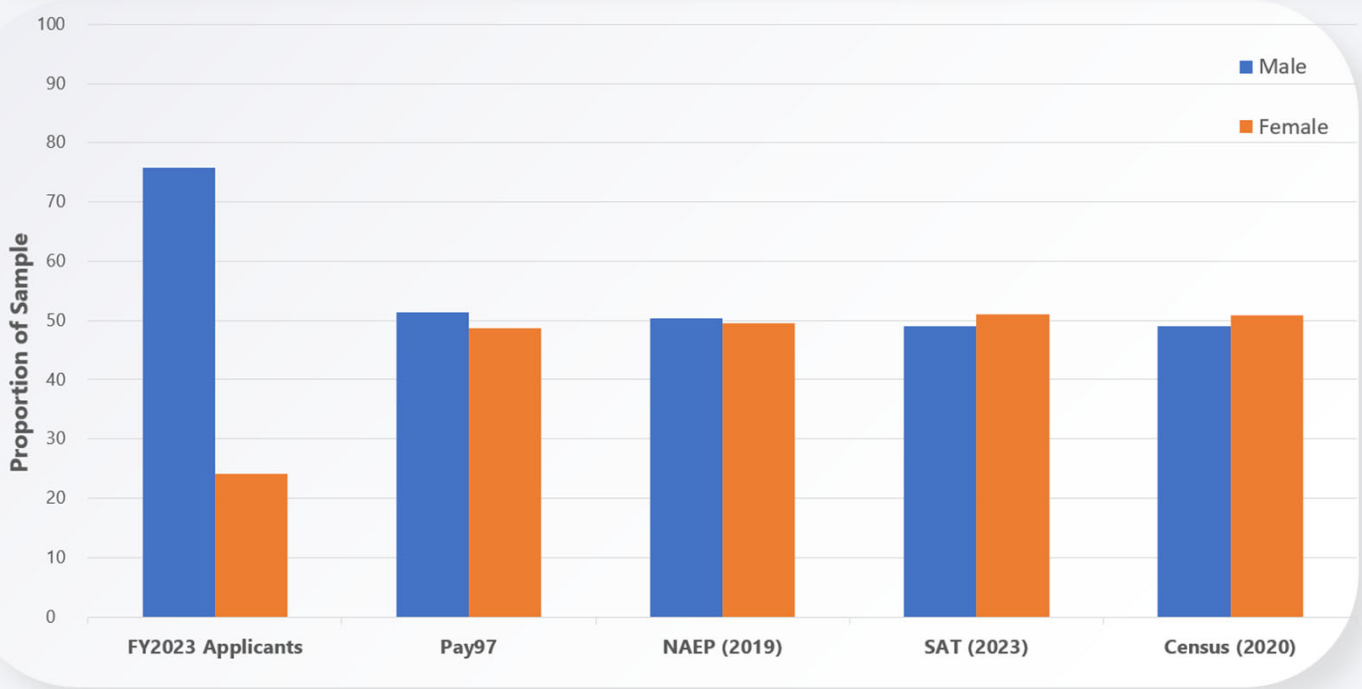
Effect Sizes

Test	Effect Size
PC (FY23)	0.29
PC (Pay97)	-0.12
WK (FY23)	0.40
WK (Pay97)	0.03
VE (FY23)	0.38
VE (Pay97)	-0.01
NAEP (2019)	-0.31
NAEP (Pay97)	-0.33
SAT (2016)	0.02
SAT (Pay97)	0.04

Trivial | Small | Mod | Large

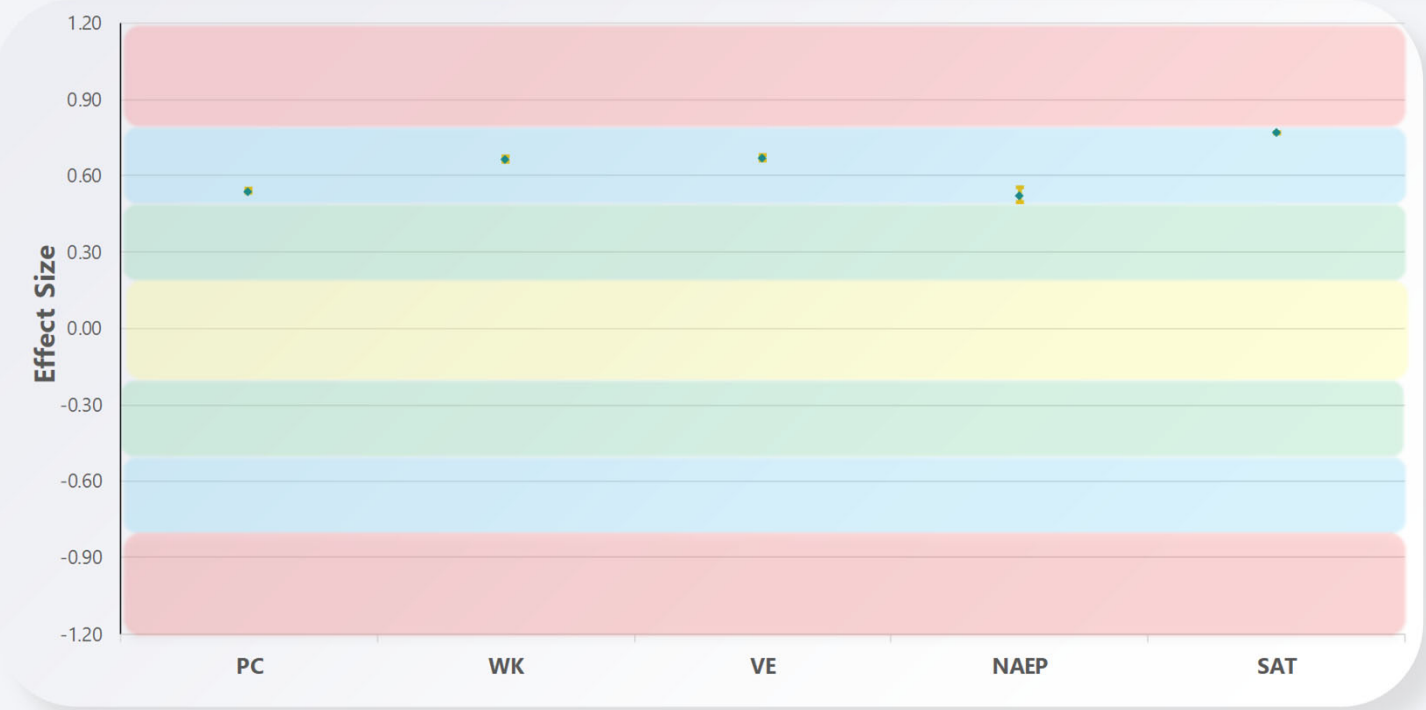
OPA
OFFICE OF PEOPLE ANALYTICS

Gender Representation Across Samples



OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs (Reading)



Effect Sizes

Comparison

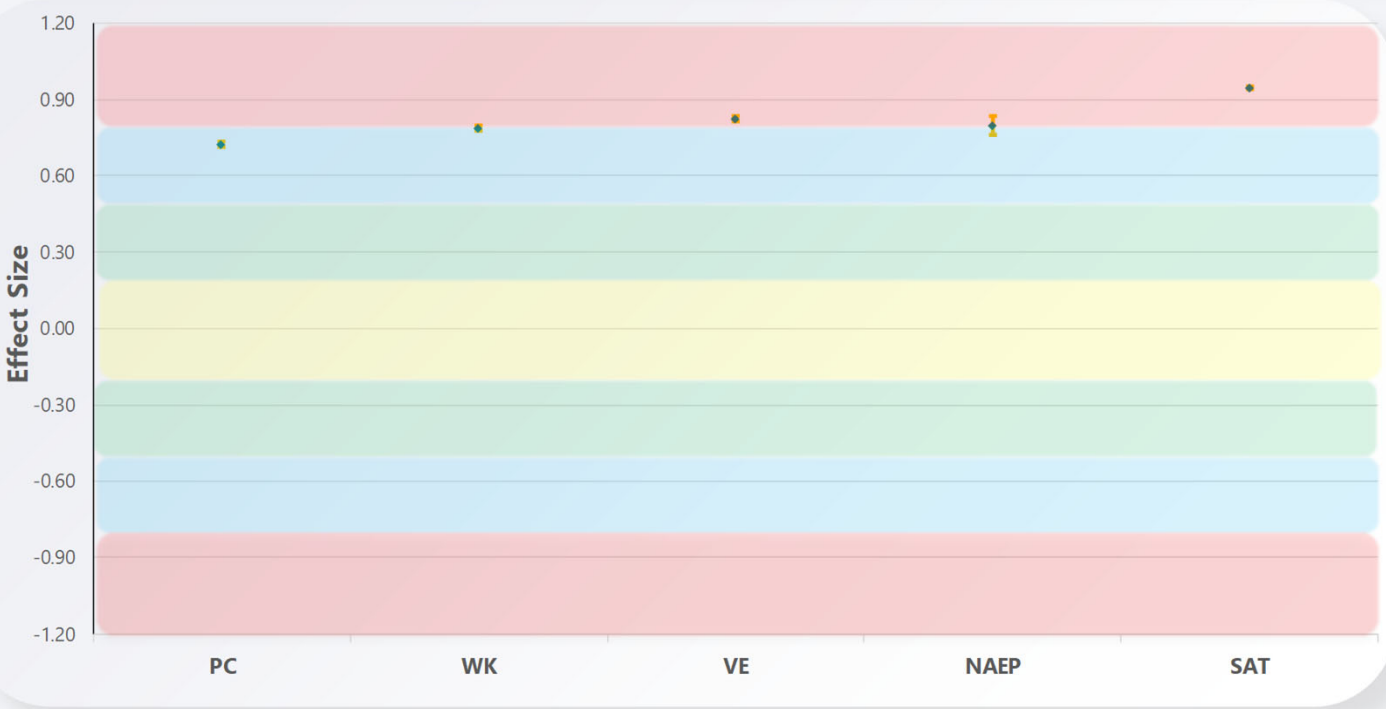
- Female vs Male
- **H vs NHW**
- NHA vs NHW
- NHB vs NHW

Test	Effect Size
PC	**0.54**
WK	**0.67**
VE	**0.67**
NAEP	**0.53**
SAT	**0.77**

Trivial **Small** **Mod** **Large**

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs (Reading)



Effect Sizes

Comparison

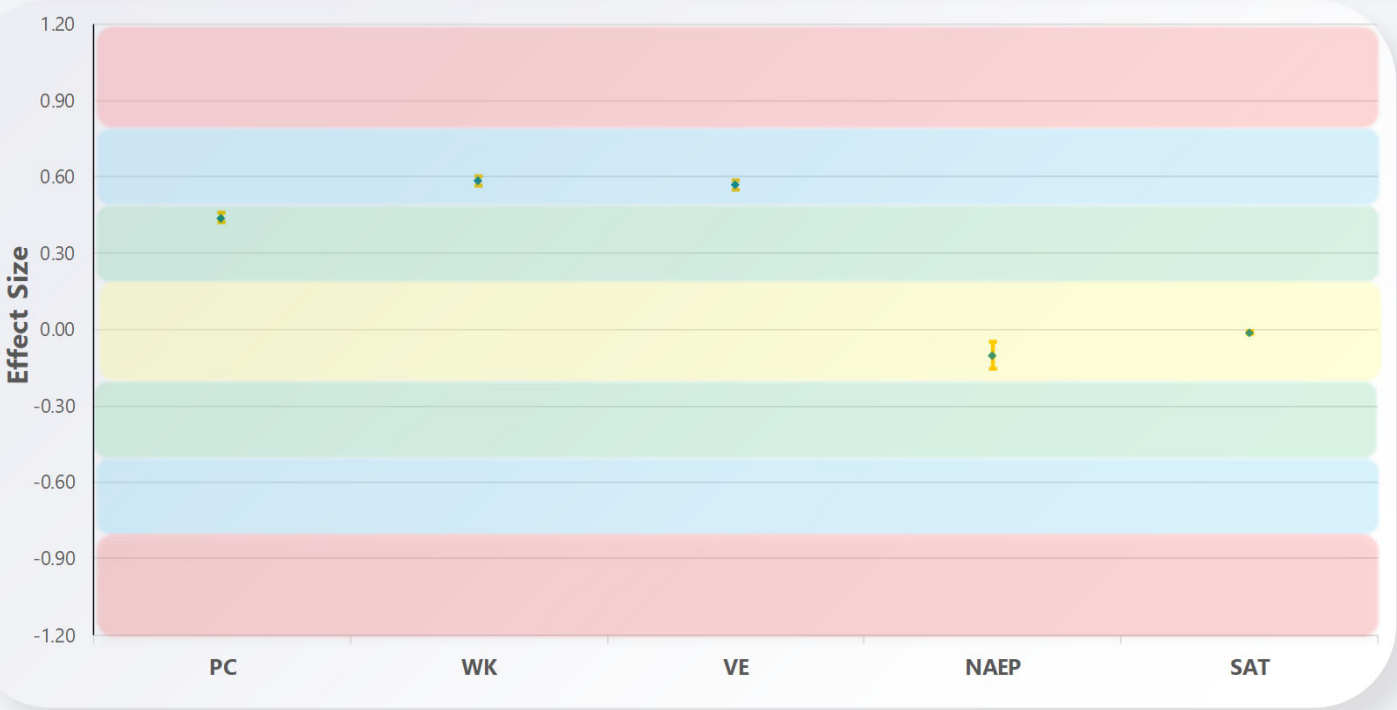
- Female vs Male
- H vs NHW
- NHA vs NHW
- **NHB vs NHW**

Test	Effect Size
PC	0.73
WK	0.79
VE	0.82
NAEP	0.80
SAT	0.95

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs (Reading)



Effect Sizes

Comparison

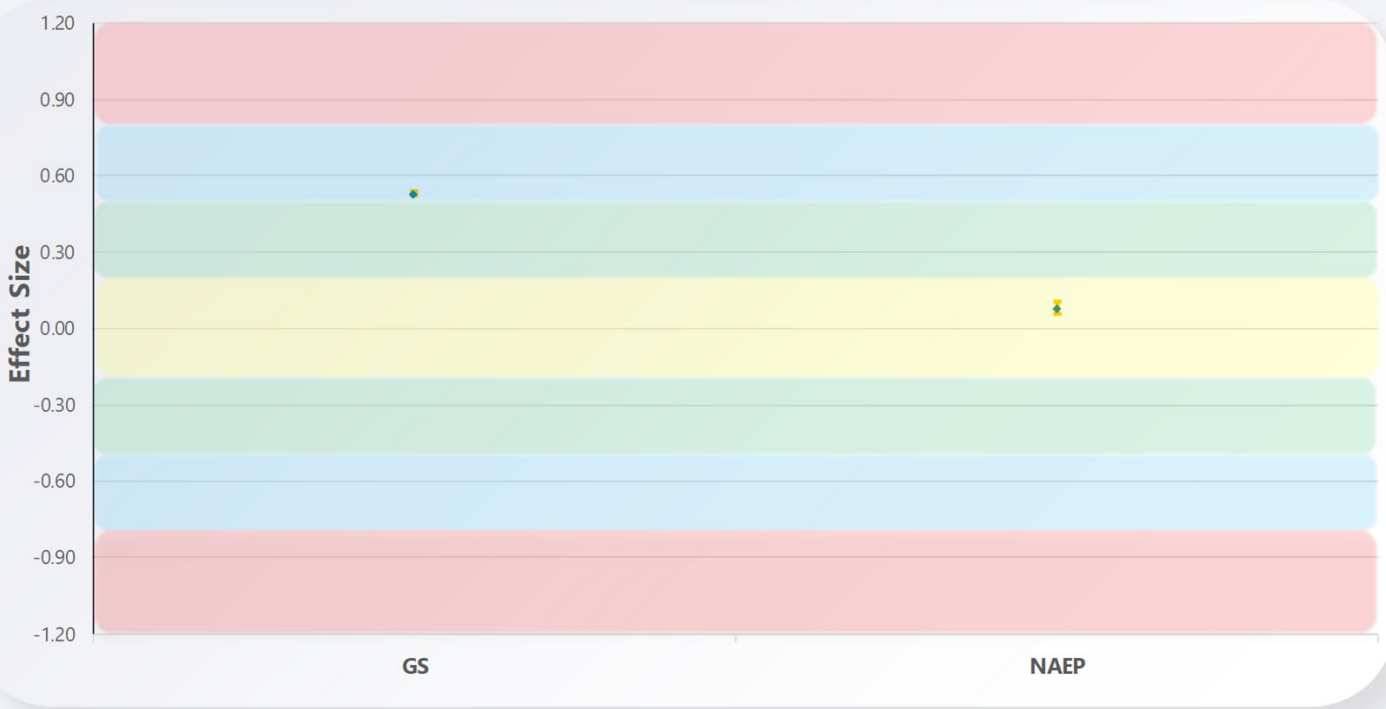
- Female vs Male
- H vs NHW
- **NHA vs NHW**
- NHB vs NHW

Test	Effect Size
PC	0.44
WK	0.58
VE	0.57
NAEP	-0.10
SAT	-0.01

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs (Science)



Effect Sizes

Comparison

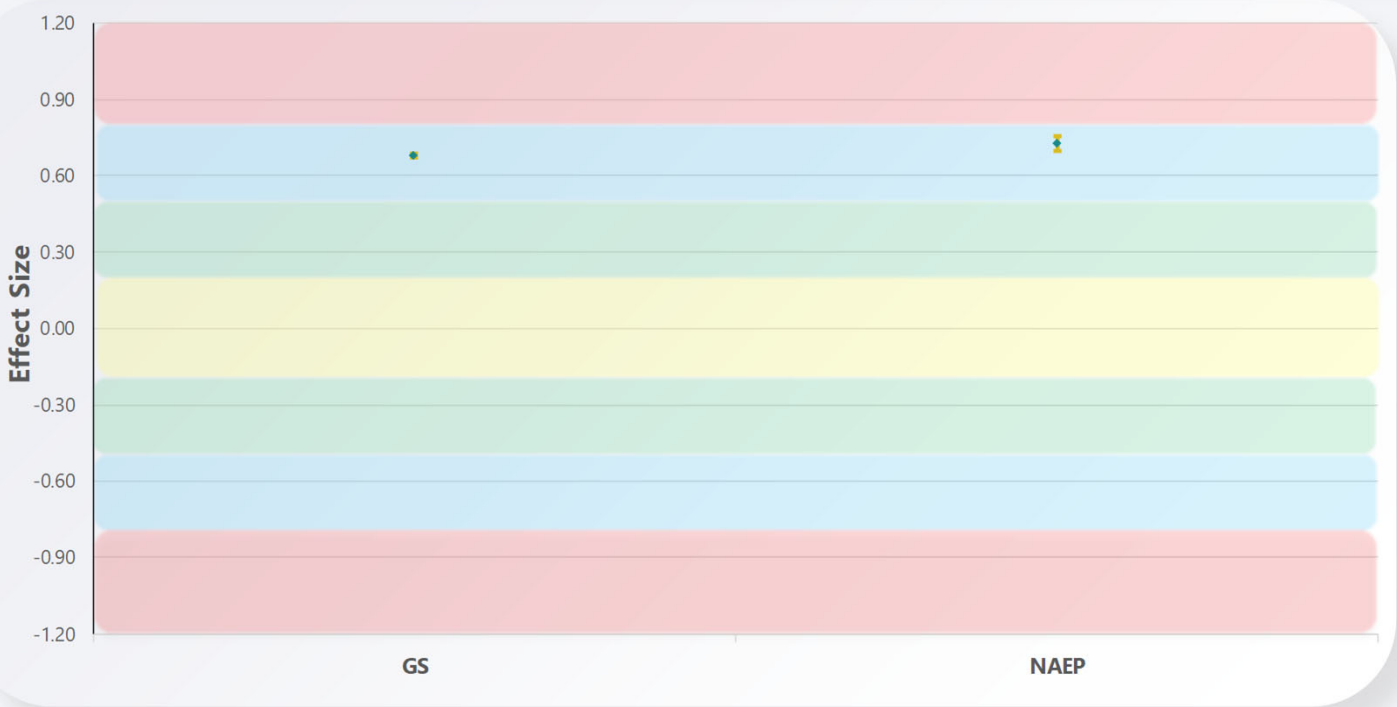
- **Female vs Male**
- H vs NHW
- NHA vs NHW
- NHB vs NHW

Test	Effect Size
GS	0.53
NAEP	0.08

Trivial **Small** **Mod** **Large**

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs (Science)



Effect Sizes

Comparison

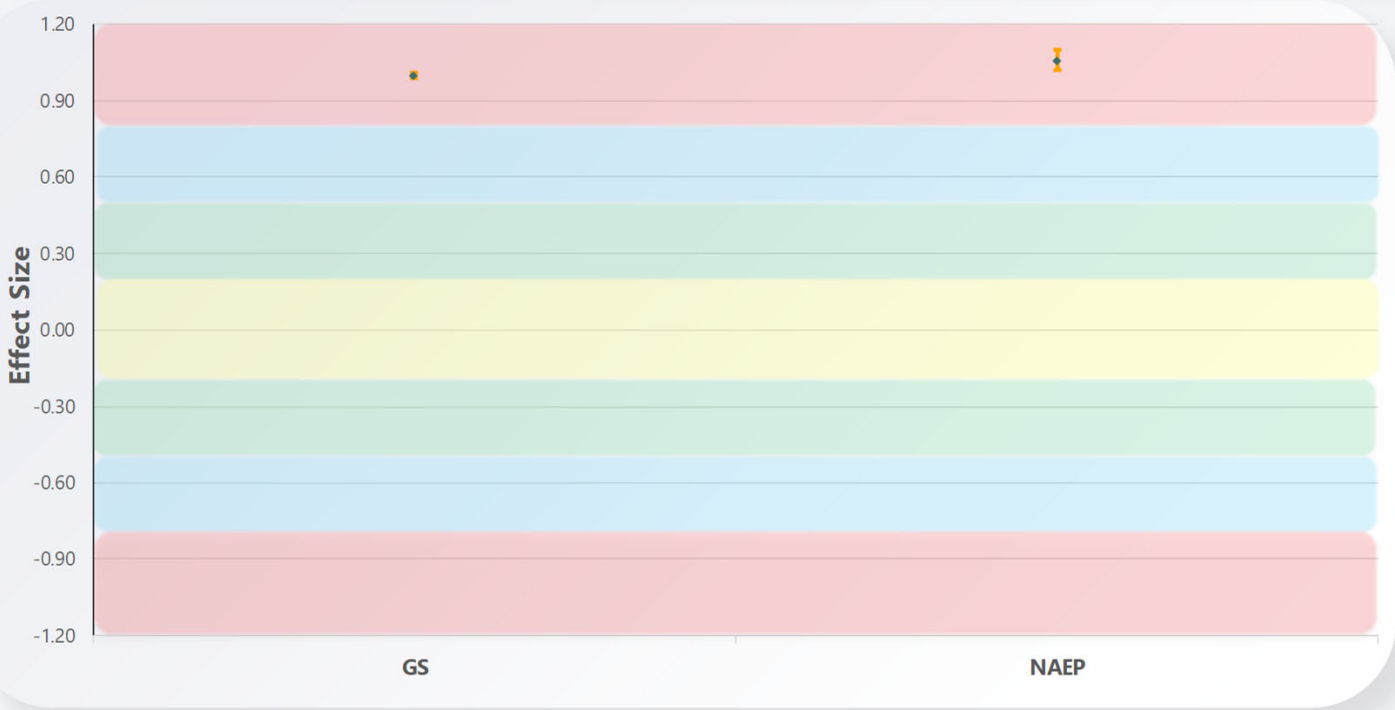
- Female vs Male
- **H vs NHW**
- NHA vs NHW
- NHB vs NHW

Test	Effect Size
GS	**0.68**
NAEP	**0.73**

Trivial **Small** **Mod** **Large**

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs (Science)



Effect Sizes

Comparison

- Female vs Male
- H vs NHW
- NHA vs NHW
- **NHB vs NHW**

Test	Effect Size
GS	**1.00**
NAEP	**1.06**

Trivial **Small** **Mod** **Large**

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of Effect Sizes Across Testing Programs (Science)

Effect Sizes

Comparison

- Female vs Male
- H vs NHW
- **NHA vs NHW**
- NHB vs NHW

Test	Effect Size
GS	**0.49**
NAEP	**-0.14**

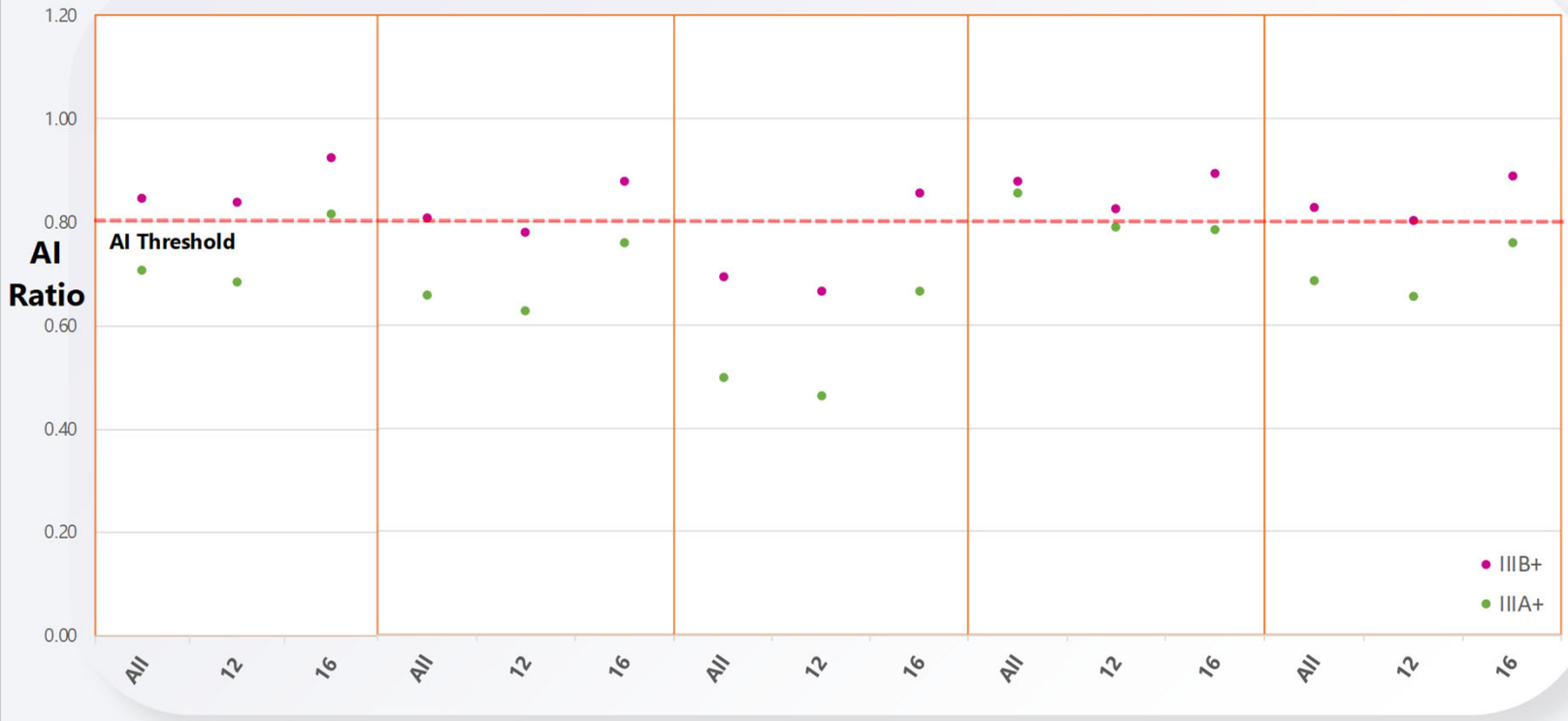
Trivial **Small** **Mod** **Large**

OPA
OFFICE OF PEOPLE ANALYTICS

Impact Ratio by Education Level

OPA
OFFICE OF PEOPLE ANALYTICS

Comparison of FY2023 Impact Ratios for Years of Education Group



Adverse Impact Ratios

Year	Education	IIIB+	IIIA+
Female vs Male	All	0.85	0.71
Female vs Male	12	0.84	0.68
Female vs Male	16	0.93	0.82
HW vs NHW	All	0.81	0.66
HW vs NHW	12	0.78	0.63
HW vs NHW	16	0.88	0.76
NHB vs NHW	All	0.69	0.50
NHB vs NHW	12	0.67	0.46
NHB vs NHW	16	0.86	0.67
NHA vs NHW	All	0.88	0.86
NHA vs NHW	12	0.83	0.79
NHA vs NHW	16	0.89	0.78
NWH vs NHW	All	0.83	0.69
NWH vs NHW	12	0.80	0.66
NWH vs NHW	16	0.89	0.76

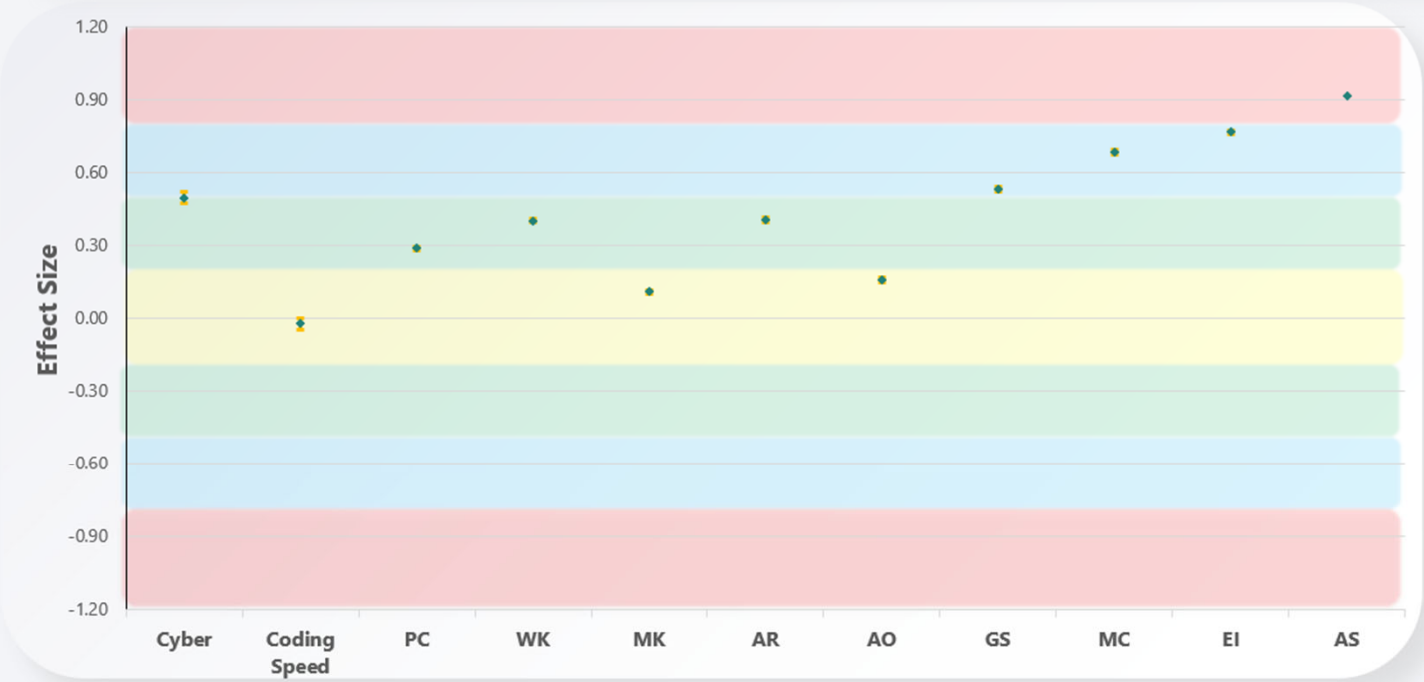
Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Effect Sizes for Special Tests

OPA
OFFICE OF PEOPLE ANALYTICS

Effect Sizes for ASVAB Scores FY2023 with Cyber Test



Effect Sizes

Comparison

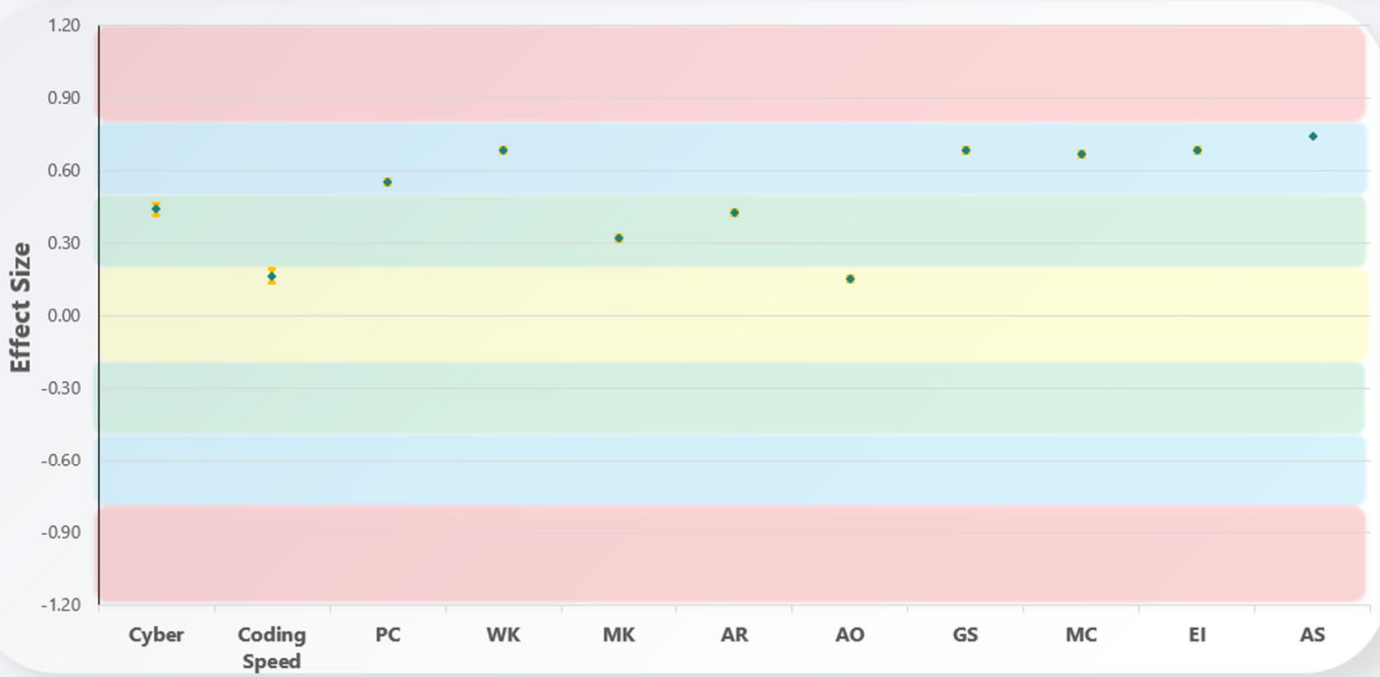
Female vs Male	HW vs NHW
NHA vs NHW	NHB vs NHW
NWH vs NHW	

Test	Effect Size
Cyber	0.50
Coding Speed	-0.02
PC	0.29
WK	0.40
MK	0.11
AR	0.40
AO	0.16
GS	0.53
MC	0.68
EI	0.76
AS	0.91

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Effect Sizes for ASVAB Scores FY2023 with Cyber Test



Effect Sizes

Comparison

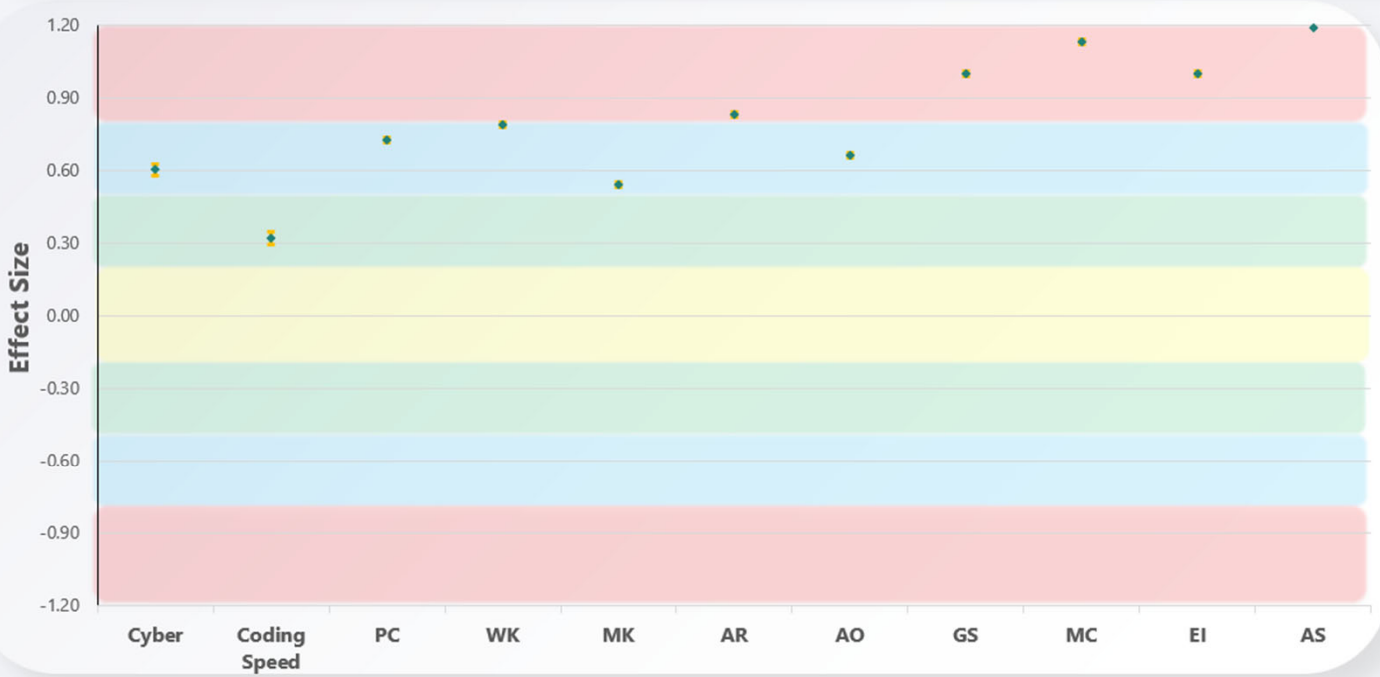
Female vs Male	**HW vs NHW**
NHA vs NHW	NHB vs NHW
NWH vs NHW	

Test	Effect Size
Cyber	0.44
Coding Speed	0.16
PC	0.55
WK	0.68
MK	0.32
AR	0.42
AO	0.15
GS	0.68
MC	0.67
EI	0.68
AS	0.74

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Effect Sizes for ASVAB Scores FY2023 with Cyber Test



Effect Sizes

Comparison

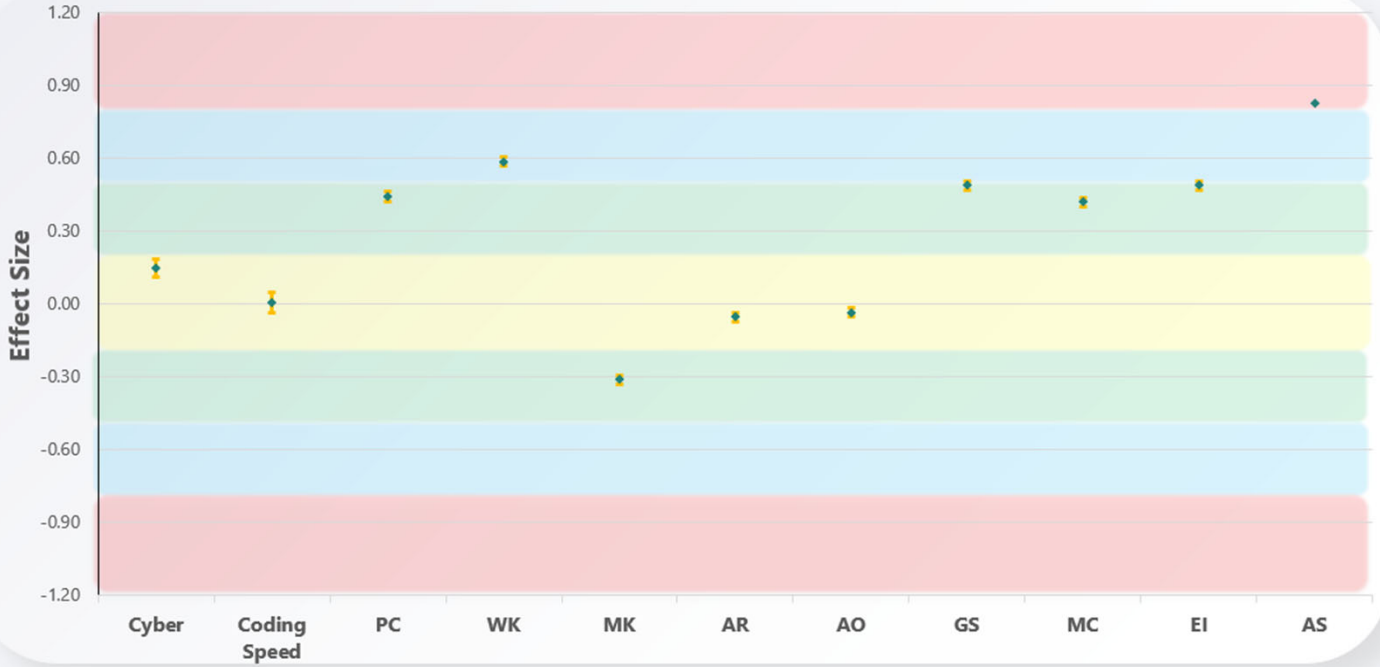
Female vs Male	HW vs NHW
NHA vs NHW	**NHB vs NHW**
NWH vs NHW	

Test	Effect Size
Cyber	0.60
Coding Speed	0.32
PC	0.73
WK	0.79
MK	0.54
AR	0.83
AO	0.66
GS	1.00
MC	1.13
EI	1.00
AS	1.19

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Effect Sizes for ASVAB Scores FY2023 with Cyber Test



Effect Sizes

Comparison

Female vs Male	HW vs NHW
NHA vs NHW	NHB vs NHW
NWH vs NHW	

Test	Effect Size
Cyber	0.15
Coding Speed	0.00
PC	0.44
WK	0.58
MK	-0.31
AR	-0.06
AO	-0.04
GS	0.49
MC	0.42
EI	0.49
AS	0.83

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Effect Sizes for ASVAB Scores FY2023 with Cyber Test

Effect Sizes

Comparison

Female vs Male | HW vs NHW | NHA vs NHW | NHB vs NHW | **NWH vs NHW**

Test	Effect Size
Cyber	0.40
Coding Speed	0.27
PC	0.49
WK	0.59
MK	0.32
AR	0.49
AO	0.23
GS	0.68
MC	0.73
EI	0.70
AS	0.84

Trivial | Small | Mod | Large

OPA
OFFICE OF PEOPLE ANALYTICS

Conclusions

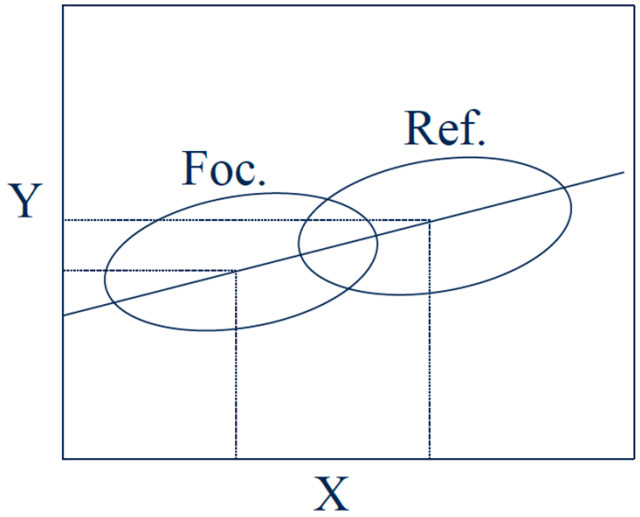
OPA
OFFICE OF PEOPLE ANALYTICS

Conclusions and Caveats

- For the AFQT tests and GS, the direction and magnitude of overall impact is generally consistent with comparable SAT and NAEP tests, which suggests that impact on ASVAB tests is reflective of differences in job or training performance
  - Comparisons across programs may be somewhat restricted due to differences in group definitions, testing populations, test content, etc.
    - NAEP is effectively an unrestricted sample
    - Those self-selecting into the Armed Services likely differ from SAT test-takers in terms of personality, motivation, and other characteristics

Conclusions and Caveats

- Adverse impact does **not** reflect test 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 test bias is indicated when the regression line relating the test score [X] and a criterion [Y] is the same for each group.
  - This was the approach taken by Putka et al. (2022).



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

Conclusions for Special Tests

- Cyber Test and Coding Speed generally exhibited small-to-moderate effects and were usually as low or lower than most ASVAB tests
  - Effects for CT and CS were also generally consistent with those found in FY21
    - Exception: CS NHW-NHB ES in FY21 was near 0, but was near .30 in FY23
- CS usually had very small effects (ranging from 0 to 0.30)

OPA
OFFICE OF PEOPLE ANALYTICS

Questions for the DAC

- Does the DAC have any general feedback or recommendations based on these results?
- For future analyses, are there any other results the DAC would be interested in seeing?

Acknowledgments

- Co-PD: Jessica Johnston-Fisher
- Analyst Team: Vanessa Nguyen, Eryn Nielsen
- Other HumRRO Contributors: Matt Reeder, Insu Paek, Emily Borawski, Cathedia Rose, Jeff Dahlke
- DTAC: Matt Trippe, Greg Manley, Ping Yin, Mary Pommerich, Liz Waterbury, Tom Waterbury

OPA
OFFICE OF PEOPLE ANALYTICS

Thank You!

For more information please contact:

Nicholas Howald

nhowald@humrro.org

703.881.6044

OPA
OFFICE OF PEOPLE ANALYTICS

Tab Q

OPA

OFFICE OF PEOPLE ANALYTICS

High School Curriculum Study

Rod McCloy

Human Resources Research Organization

Briefing presented to the DACMPT
January 23, 2025

Background

- Goals
  - Design a research study to:
    - Determine how ASVAB subtests align with content taught in high schools
    - Explore how ASVAB content is taught
    - Map ASVAB content to other relevant sources
  - Design should include:
    - Review of previous high school curriculum and high school assessment alignment studies with ASVAB content
    - Review of previous mappings between ASVAB and other tests
    - Review of any available National Assessment of Educational Progress (NAEP) transcript studies
    - Method for assessing if there are differences between course-taking patterns of military applicants and the general high school population

Trends in Teaching Practices

OPA
OFFICE OF PEOPLE ANALYTICS

Trends in Teaching Practices

- Most significant (relatively) recent development was the introduction of the Common Core State Standards (CCSS) in 2009 and the Next Generation Science Standards (NGSS) in 2011
- CCSS recommends (a) regular practice with complex texts and writing assignments involving the use of evidence and (b) practices that support gaining a conceptual understanding of mathematical principles
- NGSS recommends emphasis on in-depth development of core explanatory ideas, using ideas to generate and apply models to various phenomena, and treating science as a coherent progression over the course of K–12 education with knowledge built over time and across disciplines
- In both cases, research has produced mixed results regarding impact (Kane et al., 2016; Loveless, 2014, 2015; Song et al., 2019; Gao et al., 2018, 2022)

Trends in Teaching Practices (cont.)

- Integrated Instruction
  - Blending content within or across disciplines
  - Research has shown mixed results with more positive results at lower grades (Becker et al., 2011; Winarno et al., 2020)
- *Learning progressions* is a research-based method for developing instruction
  - Identify ultimate objective of instructional unit/sequence and work back to identify all prerequisites
- *Microlearning* involves breaking material into small chunks and including assessments to gauge incremental understanding
- *Flipped instruction* moves the presentation of content to outside the classroom so class time can be devoted to more in-depth discussion

Trends in Teaching Practices (cont.)

- *Project-based instruction* assigns students real-world issues to work on individually or in groups
- *Use of technology in instruction*
  - Gray et al. (2021) found that 47% of schools reported employing self-contained instructional practices to a moderate or great extent
  - National Center for Education Statistics (NCES) study found that 84% of schools reported using technology for activities normally done in the classroom, and 54% indicated use for activities that would not be possible without technology

Implications for the ASVAB

- Given the largely decentralized status of public education, attempting to adapt to various trends would be difficult
  - Some states adopted, then replaced, CCSS
  - New York moved to integrated math curricula, then returned to traditional format
- Larger implication may be in the way student knowledge is assessed
  - Recent comparison of ASVAB and Smarter Balanced Assessment Consortium (SBAC) math items found the latter required students to demonstrate skills in a more diverse and language-intense context
  - Review of SBAC items found them to often involve fairly lengthy reading passages with multiple questions related to each passage
    - Identify an inference that can be drawn from the passage, then select the portion of the text that supports your answer
  - Also often involve open-ended questions that require students to think critically and cite evidence in their response

Implications for the ASVAB (cont.)

- Could suggest more complex item types, e.g.,
  - Include a passage that presents a particular point of view on a topic
  - Examinee is told that the passage must be shortened by selecting the most relevant points and arranging them in a cohesive order
- Implementation would involve challenges
  - Need valid and reliable automated scoring options for open-ended items given the volume of testing
  - Likely increase in item development costs
  - Significant programming efforts to implement
  - Could result in increased testing times
- Such changes might run contrary to the desire to incorporate more language-free content into the ASVAB to accommodate non-native English speakers

Prior ASVAB Alignment Studies

OPA
OFFICE OF PEOPLE ANALYTICS

Prior ASVAB Alignment Studies

- Oppler et al. (1997) focused on technical tests and General Science (GS)
  - Examined 1990 High School Transcript Study (HSTS) data
  - Conducted an Exposure to Content survey of recruits
    - Both indicated high levels of exposure to GS content; less so for technical tests
    - Recruit sample was “technically better prepared than the HSTS sample”; likely a selection effect
  - Results from a survey of military SMEs indicated that ASVAB content is relevant to military training/jobs
- Waugh et al. (2015) examined content blueprints of ASVAB subtests in relation to educational/assessment programs that address similar subject areas (e.g., NAEP, SAT, ACT)
  - Developed alternate subtest taxonomies
    - Found a good deal of overlap between ASVAB and sources reviewed
    - Revised taxonomies provided more detailed breakouts of content domains that could increase the breadth of the subject matter covered

Prior ASVAB Alignment Studies (cont.)

- Summary
  - Results from Oppler et al. (1997) and more recent work (Adams et al., 2022) indicate that ASVAB science and technical tests are relevant to military jobs
  - Waugh et al. (2015) found a good deal of overlap between ASVAB test blueprints and other relevant sources (e.g., SAT, ACT, NAEP), particularly those tests that address content regularly taught in schools (i.e., Word Knowledge [WK], Paragraph Comprehension [PC], Arithmetic Reasoning [AR], Math Knowledge [MK], and General Science [GS])
    - Technical tests more questionable
    - Relevant comparison sources found for Auto Information (AI) and Shop Information (SI), but not Mechanical Comprehension (MC) and Electronics Information (EI)

High School Course Taking

OPA
OFFICE OF PEOPLE ANALYTICS

High School Course Taking

- Review of literature identified four broad categories of research
  - Course-taking and changes in course-taking over time
  - Impact of course-taking on future outcomes
  - Changes in and impact of Career and Technical Education (CTE) course taking
  - Methodological studies

- Much of the research based on NCES-sponsored studies
  - High School Longitudinal Study (HSLS:2005, 2009)
  - High School Transcript Study (HSTS: 1990, 1994, 1998, 2000, 2005, 2009, 2019)

High School Course Taking (cont.)

- Overall results indicate that students earned more credits and pursued more challenging curricula in 2009 compared to 1990, especially in math and science (NCES, 2011)
  - However, there are findings that suggest course titles may not reflect actual level of course content
  - 2019 data suggest only 12% of students followed a rigorous curricula and 23% were below standard (NCES n.d.)
- Results of several studies suggest students who do well in middle school math and science classes are more likely to take advanced classes in high school
- Students who take Algebra 1 before 9th grade are more likely to be proficient on standardized tests and more likely to go on to postsecondary institutions (NCES 2019)

OPA
OFFICE OF PEOPLE ANALYTICS

CTE Course Taking

- Results from a variety of studies yield the following general conclusions
  - Most students earn Career and Technical Education (CTE) credits while in high school
  - The percentage doing so has declined somewhat from 1990 to 2015
  - Course-taking patterns have shifted over time (e.g., less focus on areas such as agriculture, architecture/construction, and business/marketing, and greater focus on engineering/technology, health care, hospitality/tourism, and human services)
  - Consistent differences between males and females in areas of focus
    - A higher percentage of males earn credits in architecture and construction, engineering and technology, manufacturing, and transportation and logistics, while a higher concentration of females in health care and human services
  - Overall test scores and graduation rates for students taking CTE courses have risen over time
  - Limited data suggest no relationship between CTE course-taking and postsecondary pursuits

Methodological Studies

- Rosen et al. (2017) examined data from HSLS: 2009, comparing student reports of math courses taken to their actual transcripts
  - Overall self-reports were accurate regarding courses taken, with less accuracy about year taken and grade received
  - Greater accuracy in reporting grade received among higher-performing students
- NCES (2020) compared courses students reported taking as part of the NAEP studies conducted in 2000, 2005, and 2009 to their high school transcripts
  - For all math courses except pre-calculus and unified/integrated math, a higher percentage of students reported taking the class than was indicated by their transcript
    - In all standard math classes (Algebra 1, Geometry, Algebra 2), higher percentages of students reported taking the class than was indicated by their transcripts, with differences ranging from 2% to 7%

Current Study

OPA
OFFICE OF PEOPLE ANALYTICS

Current Study

1. Review relevant sources (e.g., NAEP, ACT) to determine if they have been updated/revised in a way that makes them more or less aligned with ASVAB
2. Conduct "pseudo-alignment" study in which SMEs review high school course catalogs with ASVAB test blueprints and make judgments regarding whether content is addressed in schools
3. Work with Joint Advertising, Market Research, and Studies (JAMRS) to include questions on course-taking/extracurricular activities in their Ad Tracking Survey, which examines awareness of and reactions to military advertising campaigns
    - Survey conducted quarterly with a stratified random sample of U.S. youth 16–24 years old who previously responded to the *Futures Survey*, which obtains information on attitudes toward the military and propensity to enlist
4. Explore HSTS: 19 data to identify relevant results that have not been reported in the literature (in process)

Review of Comparable Taxonomies

- PC—ACT Curriculum Study
  - High school ELA teachers indicated topic areas most frequently taught
  - Highest rated were composing skills and strategies, vocabulary, comprehension strategies, analysis and evaluation of texts, and inferential comprehension of texts
  - HumRRO PC editors reviewed findings and agreed that vocabulary is covered (in WK), inferential comprehension is addressed, and analysis and evaluation of texts is partially covered (no evaluation)
  - Composing skills and strategies are not addressed

OPA
OFFICE OF PEOPLE ANALYTICS

Review of Comparable Taxonomies (cont.)

- PC—ACT Reading and Readiness Standards
  - Standards set for various reading score ranges (i.e., 13–15, 16–19, 20–23)
  - Comparisons with ASVAB are not clear-cut due to inclusion in the standards of the phrases “somewhat challenging” and “challenging” passages
  - ASVAB PC passages limited to 100–180 words to eliminate scrolling;
    ACT averages ~800 words
  - HumRRO PC editors agreed that most standards are addressed
  - Exceptions include determining cause-effect relationships and making comparisons between passages

Review of Comparable Taxonomies (cont.)

- PC—NAEP Reading Assessment and Achievement Level Definitions
  - Again, comparisons not straightforward
  - Reading assessment includes items that require comparisons between two or more texts, and passage length can range from 500–1,500 words
  - Seven item types, only one of which is used in PC (i.e., single-selection multiple choice)
  - PC editors agreed that Basic Achievement Level Standards are addressed in the ASVAB
  - Those at higher levels (i.e., proficient, advanced) not covered or only partially covered
  - Common characteristics of standards not covered include
    - Diagrams and charts
    - Comparison between texts
    - Requiring analysis, evaluation, synthesis, and critique of texts

Review of Comparable Taxonomies (cont.)

- MK/AR—ACT Curriculum Study
  - Math teachers rate most important skills to be developed
  - Four skills not included in MK/AR blueprint are higher level (e.g., Math 3, Algebra 2)
- MK/AR—ACT Math Readiness Standards
  - Standards set for various ACT Math Score Ranges (i.e., 13–15, 16–19, 20–23)
  - Six of the 12 skills at the 13–15 level are addressed, and remainder could be covered in ASVAB, assuming they could be assessed through multiple-choice questions (e.g., locate positive rational numbers on number line, estimate length of line segment based on other lengths in geometric figure)
  - All skills at the 16–19 level are or could be addressed in ASVAB except one involving probability, which is not in the existing blueprints
  - Several skills at the 20–23 score level were judged to be outside the AR/MK blueprint (e.g., add two matrices that have whole number entries); others were judged to be included or candidates for inclusion in AR/MK

Review of Comparable Taxonomies (cont.)

- MK/AR—2022 and 2024 NAEP Mathematics Assessment Framework
  - Includes objectives deemed appropriate for assessment by subtopic and grade
  - All objectives in Numbers Properties and Operations are covered, partially covered, or could be covered in the ASVAB, although addressing some would require expanding item types (e.g., identify situations where estimation is appropriate)
  - Most objectives in Measurement are covered, partially covered, or could be covered, except measurement in triangles (e.g., solve problems using the fact that trigonometric ratios stay constant in similar triangles)
  - Most objectives in Geometry, Algebra, and Data Analysis/Statistics/Probability were judged outside of the MK/AR blueprint
    - Most would require more expansive item types (e.g., describe, analyze, explain)

Review of Comparable Taxonomies (cont.)

- GS—Next Generation Science Standards
  - Cover three broad areas—Physical Sciences, Life Sciences, Earth/Space Sciences, which are also addressed in the ASVAB
  - Subareas within each define skills high school students should be able to demonstrate
  - Emphasis is on application of knowledge rather than retention
  - As a result, most would require alternate means of assessment (e.g., conduct a project, write a paper) or more expansive item types (e.g., develop a model, communicate scientific information)

Review of Comparable Taxonomies (cont.)

- GS—ACT Science Test Topic Areas
  - Cover three broad areas—Life Science/Biology, Physical Science/Chemistry/Physics, Earth/Space Science
  - HumRRO GS editor judged all to be covered in GS
- GS—ACT Science College and Career Readiness Standards
  - Describe what students at various score levels should be able to do (13–15, 16–19, 20–23)
  - Three broad areas—Interpretation of Data, Scientific Investigation, Evaluation of Models/Inferences/Experimental Results
  - HumRRO GS editor indicated that the descriptors do not represent the way in which content is covered by ASVAB (e.g., Compare, Determine) although certain topics are addressed (e.g., understand basic scientific terminology)

Review of Comparable Taxonomies (cont.)

- GS—National Academy of Sciences, National Research Council Framework for K–12 Science Education
  - Covers four broad areas—Physical Sciences, Life Sciences, Earth/Space Science, Engineering/Technology/Application of Science
  - HumRRO GS editor judged nearly all are covered in GS except Engineering, Technology, and Applications of Science
- GS—2028 NAEP Science Framework
  - Addresses the first three of the four topic areas above
  - HumRRO GS editor identified all topic areas as addressed in GS except Evidence of Common Ancestry and Diversity

Review of Comparable Taxonomies (cont.)

- Conclusions
  - ASVAB addresses the preponderance of content covered in other reviewed sources
  - Possible additions to test blueprints identified
  - Many skills not assessed by ASVAB would be difficult to address through a test or would require more complex/varied item types
  - Differences in underlying purpose of the ASVAB (selection/classification) and other tests (diagnostic/developmental) may obviate the need to assess knowledge/skills in similar ways

Preliminary Results—Alignment Study

- School Sampling Approach
  - Randomly selected one state from each of the 9 Census Regions: RI, PA, MI, MN, VA, TN, AR, MT, and CA
  - Created an extract of Common Core of Data public school directory for each state
  - Sorted schools by level and eliminated Pre-K, elementary, and middle schools
  - Sorted schools by type and eliminated special education, unknown, and alternative schools
  - Generated random numbers to select 5 schools from each state

Preliminary Results—Alignment Study (cont.)

- Compared distribution of jurisdiction sizes to national data
  - Significant underrepresentation of City/Large (national = 15.08% of schools; sample = 8.41% of schools)
  - Three states in sample have no City/Large jurisdictions (AR, MT, and RI)
  - MN had two City/Large schools randomly chosen
  - PA, MI had none—randomly chose one City/Large school from each
- Added TX and FL to represent high-recruitment states

OPA
OFFICE OF PEOPLE ANALYTICS

Preliminary Results—Alignment Study (cont.)

- Logged on to school websites and sought course catalogs
  - Found detailed course descriptions for 40 of 57 schools
  - Schools lacking course catalogs tended to have small student populations (e.g., < 250)
- Drew additional sample within state/size jurisdiction groups, when necessary, until catalogs located
  - Implication: Smaller schools may be underrepresented

Preliminary Results—Alignment Study (cont.)

- Identified SMEs (item writers/editors) for MK/AR, GS, EI, AI, SI, MC, Cyber
- Created ratings spreadsheet
- Conducted meetings to provide overview of task
  - Purpose
  - How schools were selected
  - Use of rating sheet
- As of the time these slides were generated, ratings still in progress
- Results reflect data obtained to date

Preliminary Results—Alignment Study (cont.)

- AR/MK—All ASVAB content covered either in prerequisite courses (to those in the catalogs) or by basic courses in the catalogs
  - Possible exception: ASVAB time/temperature, with SMEs identifying few explicit mentions in catalogs
- GS—Almost all topics covered in a mixture of basic and advanced courses
  - Exception: ASVAB Life Science/Botany, which was not addressed in ~60% of high school course catalogs
- AI—Of the 56 catalogs reviewed thus far, 34 were identified as having no automotive technology/repair classes

OPA
OFFICE OF PEOPLE ANALYTICS

Preliminary Results—Alignment Study (cont.)

- SI—Content available in approximately two-thirds of the catalogs reviewed thus far
- MC—All six blueprint elements covered in the catalogs reviewed thus far
- Cyber—10 schools offered no related classes, and 8 provided courses only in use of IT and software
  - All test components covered in 14 schools
  - Topics most likely to be omitted were Network Configuration, Offensive Methods, and PC Configuration and Maintenance

OPA
OFFICE OF PEOPLE ANALYTICS

Ad Tracking Survey Results—Course Taking

- Small number of propensed respondents (89 of 880)
- Significantly higher proportions of respondents *not* considering military service reported taking biology, chemistry, physics, calculus, and statistics/probability
- Significantly higher percentage of those in the “definitely not enlist” category compared to those in the “probably not enlist” category took chemistry and statistics/probability
- Significantly higher proportion of those in the “definitely not enlist” category took business/marketing compared to those in the propensed group

Ad Tracking Survey Results—Course Taking

Courses	Total (n = 880)	Probably/Definitely Enlist (n = 89)	Probably Not Enlist (n = 357)	Definitely Not Enlist (*n* = 433)
Algebra	79%	69%	82%	80%
Biology	74%	56%	74%↑	76% ↑
Geometry	67%	52%	71%↑	67%
Chemistry	63%	35%	60%↑	71%↑↑
Health Sciences	45%	31%	45%	47%
Physics	39%	19%	39%↑	42%↑
Calculus	31%	5%	29%↑	36%↑
Computer Science	27%	18%	25%	30%
Statistics/Probability	26%	5%	23%↑	32%↑↑
Business/Marketing	21%	14%	17%	24%↑
Agriculture/Food Science	17%	19%	17%	15%
Engineering	13%	17%	13%	12%
Woodworking	12%	10%	15%	11%
Electronics/Electrical Systems	9%	15%	8%	8%
Manufacturing/Welding	7%	14%	8%	6%
Architecture/Construction	5%	11%	4%	5%
Transportation/Auto Repair	3%	10%	4%	2%
None of the Above	4%	3%	6%	4%
Refused	10%	19%	10%	8%

↑ = higher than propensed youth

↑ = higher than "Probably Not" propensed youth

Ad Tracking Survey Results—Extracurricular Activities

- Participation in extracurricular activities below 10% in most cases
- Highest participation levels in social service/volunteer efforts, sports and cheerleading, computer-related pursuits
- Few significant differences across groups
- Most notable difference was higher percentages of those in the medium- and high-propensity groups taking part in automobile and construction activities

OPA
OFFICE OF PEOPLE ANALYTICS

Ad Tracking Survey Results—Extracurricular Activities

Activity	Total (n = 1,150)	Probably/ Definitely Enlist (n = 134)	Probably Not Enlist (n = 443)	Definitely Not Enlist (n = 573)
Sports/Cheerleading/Drill Team	17%	20%	20%	14%
Academic Clubs				
Mathematics	4%	8%	6%↑	2%
Biology	3%	2%	4%	3%
Chemistry	2%	1%	1%	2%
English/Creative Writing	4%	6%	3%	5%
Debate	2%	5%	2%	2%
History	2%	4%	2%	1%
Foreign Language	7%	4%	8%	7%
Special Interest Clubs				
Cooking	3%	5%	4%	1%
Film	2%	1%	3%	2%
Photography	3%	4%	3%	3%
Chess	4%	5%	6%	2%
Art (Painting, Pottery)	8%	9%	7%	9%
Music (Band, Orchestra, Choir)	15%	14%	16%	13%
Social Service (Animal Welfare, Food Bank)	23%	21%	26%	21%
Computers/Electronics (Assembly, Repair, Programming)	11%	17%	13%	8%
Automobiles (Repair, Restoration)	7%	14%↑	10%↑	3%
Construction (Buildings, Furniture)	8%	17%↑	11%↑	3%
Boy/Girl Scouts	2%	2%	3%	2%
Agriculture (4-H, Future Framers of America)	4%	9%	4%	3%
Other	10%	7%	12%	8%
None of the Above	29%	22%	24%	35%↑↑
Refused	9%	14%	8%	9%

↑ = higher than propensed youth

↑ = higher than "Probably Not" propensed youth

↑ = higher than "Definitely Not" propensed youth

Conclusions

- ASVAB content largely addressed in relevant frameworks (e.g., ACT, NAEP)
  - Some suggestions for additions to blueprints
  - Addressing some skills would require expansion of item types
- ASVAB academic content areas (e.g., GS, AR/MK) typically addressed in high school courses
- Technical content coverage is spottier
- Some indication of course-taking differences between propensed and non-propensed youth, with the latter taking higher-level courses
- Some indication that propensed youth more likely to take part in extracurricular activities relevant to ASVAB (e.g., automotive, construction)

Questions for the DAC

- Does the DAC have recommendations on how this work can improve the composition of the ASVAB for selection and classification purposes?
- ASVAB currently assesses both knowledge learned in school and knowledge and skills needed in the military that may not be addressed in formal education.
  - Do you have any thoughts on how Next Generation ASVAB can continue to bridge that gap?

OPA
OFFICE OF PEOPLE ANALYTICS

Thank you!

For more information
please contact:

Peter Ramsberger
pramsberger@humrro.org
703.706.5686

OPA
OFFICE OF PEOPLE ANALYTICS

References

Adams, K. A., Oppler, S. H., Yee Prendez, J., & Robertson, S. A. (2022). *Training relevance survey for the Armed Services Vocational Aptitude Battery (ASVAB).* Human Resources Research Organization.

Becker, K. & Park, K. (2011). Effects of integrative approaches among science, technology, engineering, and mathematics (STEM) subjects on students' learning: A preliminary meta-analysis. *Journal of STEM Education (12)*5/6, 23-370.

Gao, N., Adan, S., Lopes, L., & Lee, G. (2018). *Implementing the next generation science standards: Early evidence from California.* Public Policy Institute of California.

Gao, N., DiRanna, K., & Chang Fay, M. (2022). *The impact of COVID-19 on science education: Early evidence from California.* Public Policy Institute of California.

Gray, L., Lewis, L., & Chapman, C. (2021). *Use of educational technology for instruction in public schools: 2019–2020* (NCESS 2021-017). National Center for Education Statistics. https://nces.ed.gov/pubs2021/2021017Summary.pdf

Kane, T. J., Owens, A. M., Marinell, W. H., Thal, D. R., & Staiger, D. O. (2016). *Teaching higher: Educators' perspectives on common core implementation.* Harvard University, Center for Education Policy Research. https://cepr.harvard.edu/teaching-higher

Loveless, T. (2014). *The 2014 Brown Center report on American education: How well are American students learning?* Brown Center on Education Policy at Brookings. https://www.brookings.edu/wp-content/uploads/2016/06/2014-Brown-Center-Report_FINAL-4.pdf

Loveless, T. (2015). *The 2015 Brown Center report on American education: How well are American students learning?* https://www.brookings.edu/wp-content/uploads/2016/06/2015-Brown-Center-Report_FINAL-3.pdf

National Center for Education Statistics (n.d.) *2019 NAEP high school transcript study (HSTS results).* https://www.nationsreportcard.gov/hstsreport/#home

National Center for Education Statistics (2011). *NAEP 2009 year in review* (NCES 2011-471).

References

National Center for Education Statistics (2019). *Algebra I coursetaking and postsecondary enrollment* (NCES 2019-154).

National Center for Education Statistics (2020). *From algebra to zoology: How well do students report mathematics and science coursetaking?* NCES 2020-037.

Oppler, S. H., Russell, T. L., Rosse, R. L., Keil, C. T., Meiman, E. P., & Welsh, J. R. (1997). *Item evaluation for the Armed Services Vocational Aptitude Battery (ASVAB) science and technical test specification: Final report* (DMDC Technical Report 97-024). Defense Manpower Data Center.

Rosen, J. A., Porter, S. R., & Rogers, J. (2017). Understanding student self-reports of academic performance and course-taking behavior. *AERA Open, 3*(2), 1–14.

Song, M., Yang, R., Garet, M. (2019). *Effects of adoption of college- and career-readiness standards on student achievement.* American Institutes for Research. https://www.c-sail.org/sites/default/files/Effects%20of%20CCR%20standards%20on%20stu%20achievement_4-2019_AERA.pdf

Waugh, G., Knapp, D., Ramsberger, P., & Caramagno, J. (2015). *Refining ASVAB item and test development procedures* (2014 No. 082). Human Resources Research Organization.

Winarno, N., Rusdiana, D., Riandi, R., Susilowati, E., & Afifah, R. M. (2020). Implementation of integrated science curriculum: A critical review of the literature. *Journal of the Education of Gifted Young Scientists, 8*(2). https://files.eric.ed.gov/fulltext/ED606270.pdf

OPA
OFFICE OF PEOPLE ANALYTICS

Tab R

SLIDES ONLY
NO SCRIPT PROVIDED

OFFICE OF PEOPLE ANALYTICS

ASVAB CEP: Updates to Program and Non-Cognitive Measures

Dr. Irina Rader, *Defense Testing & Assessment Center*
Dr. Rod McCloy, *Human Resources Research Organization*
Dr. Maura Burke, *Human Resources Research Organization*

Briefing presented to the DACMPT
January 23, 2025

25-P-0267

Discussion Topics

- ASVAB Career Exploration Program (CEP) Update
- New Form for the Find Your Interests (FYI) Inventory
- Initial Analysis of Responses from the Work Values Situational Judgment Activity (WV SJA)
- Summary
- Questions and Discussion

ASVAB CEP
Current State

OPA
OFFICE OF PEOPLE ANALYTICS

Achieving ASVAB CEP Vision—A Strategic Approach

MISSION

The ASVAB CEP is a program sponsored by the Department of Defense (DoD) with a two-part mission: to provide a career exploration service to American youth and provide qualified leads to military recruiters.

VISION

The ASVAB CEP assesses academic ability and vocational interests, which together help inform career decisions. Personalized career exploration, awareness of career-field entry requirements, and future-oriented planning tools help students work with parents and educators to develop post-secondary plans. Eligible participants can use their scores to explore enlistment and have no obligation to military service.

ASVAB CEP Usage Metrics Year to Date (YTD)

OPA
OFFICE OF PEOPLE ANALYTICS

Nationwide Participation

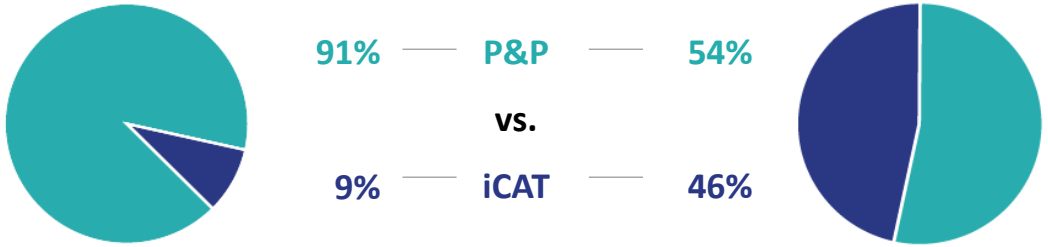
13,105 SCHOOLS PARTICIPATED

619,926 STUDENTS TESTED

LEADS PROVIDED TO MILITARY SERVICES

339,463

Paper & Pencil (P&P) vs. iCAT Participants



There has been a significant shift from Paper & Pencil assessments to CEP iCAT assessments over the past five years.

OPA
OFFICE OF PEOPLE ANALYTICS

Areas of Focus by Business Strategies

School Year 2024/2025

OPA
OFFICE OF PEOPLE ANALYTICS

2024 ASVAB CEP JAMBOREE

The CEP Jamboree is a three-day strategic planning session with stakeholders from the Defense Testing and Assessment Center (DTAC), Accession Policy (AP), Personnel and Readiness (P&R)/Manpower and Reserve Affairs (M&RA)/Military Personnel Policy (MPP), Defense Personnel Analytic Center (DPAC), and U.S. Military Entrance Processing Command (USMEPCOM). The event focuses on reviewing the past year's performance and achievements and brainstorming the direction of ASVAB CEP for school year (SY) 24–25 and beyond.

ASVAB CEP team members from AP and DTAC led the collaborative meeting with members of USMEPCOM HumRRO, Lattice Form, and Written LLC.

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB Career Exploration Program (CEP) Ecosystem of Integrated Business Strategies

School Year 2024–2025

1. TECHNOLOGY

Optimize user experience by enhancing features and addressing bugs. Migrate CEP websites into Defense Personnel Assessment Center System (DPACS) boundary to enhance security. Consolidate backend systems for operational efficiency. Expand data analytics to inform decision-making.

2. NEW RESEARCH & INNOVATION NEW

Studies to evaluate and improve CEP measures/processes: (a) students' readiness to benefit from CEP, (b) use of AI to improve occupational crosswalks, (c) evaluation of non-cognitive measures, (d) expansion of post-test interpretation (PTI) delivery, and (e) use of external data to inform program impact.

3. OCCUPATIONAL WEBSITE DATA & CONTENT

One of the primary benefits to users of the ASVAB CEP is the data contained on the program's websites. This initiative focuses on the activities undertaken to collect, analyze, store, and share occupational data.

4. PROMOTION & ENGAGEMENT

Advertising, social media, content marketing, national events, and stakeholder engagement provide opportunities for knowledge sharing and interaction with various customer segments of ASVAB CEP's target audiences.

5. WORKFORCE MULTIPLIER

The personnel responsible for delivering the ASVAB CEP require awareness and training. This initiative seeks to expand the numbers and the knowledge of those who can speak to the benefits of the program.

6. LEGISLATIVE ACTIVITIES

Monitoring ASVAB CEP legislative activities: (a) weekly monitoring and tracking of state and federal legislative activities, state education websites, and news sites, (b) systematize Department of Education connections, and (c) follow up on and maintain connections made at conferences.

7. UNDERSERVED POPULATIONS

The ASVAB CEP benefits young adults. This initiative seeks to expand access to ASVAB CEP among eligible populations including post-secondary institutions, homeschool students, and students enrolled at schools that don't offer ASVAB CEP.

ASVAB CEP Business Strategy SY24/25 Goals

- **Technology—**Migrate ASVAB Program and Careers in the Military (CITM) websites into the DPACS Boundary NLT August 2025
- **Research & Innovation—**Leverage research and innovation to enhance the ASVAB CEP program, improve occupational crosswalks, and address stakeholder needs and concerns
- **Occupational Data and Content—**Define Occupational Crosswalk Process and explore utilization of AI to further enhance collection and analysis

ASVAB CEP Business Strategy SY24/25 Goals

- **Promotion & Engagement** Execute SY24/25 Social Media Strategic Plan, increase program awareness, and grow social media presence
  - Continue to support States with ASVAB CEP Month Proclamations (Alabama, Oklahoma, and Louisiana)

ASVAB CEP and members of all Service branches gather to witness Governor Kay Ivey sign proclamation at Alabama State Capitol declaring October ASVAB Career Exploration Month.

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB CEP Business Strategy SY24/25 Goals

- **Workforce Multiplier—**Continue to expand the PTI training program, including updates to the training content and tracking; work strategic partnerships with U.S. Army Recruiting and Retention College leaders, JROTC, and MEPS Battalion Commanders
- **State Legislative Activities—**Continue tracking state and federal legislation and development of interactive mapping and visualization tooling
- **Underserved Populations—**Create pilot program with the goal to increase private and homeschool testing as well as post-secondary institution participation

New Form for the Find Your Interests (FYI) Inventory

OPA
OFFICE OF PEOPLE ANALYTICS

The FYI Inventory

- Original form developed c. 2005
  - Replaced the Interest Finder (Wall & Baker, 1997; Wall et al., 1996)
- 90-item RIASEC measure
  - Dislike/Indifferent/Like response options
- Scores are reported using norms
  - Total group
  - Sex-specific

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB CEP Expert Panel

- DTAC convened an ASVAB CEP Expert Panel in 2017 to comment on updated ASVAB CEP
  - Reviewed all components of the revamped program
  - Gave particular emphasis to the FYI Inventory
    - Lauded the measure
    - Suggested updating it to ensure (a) currency/relevance of items and (b) construct coverage per basic interests (Su et al., 2019)
- ASVAB CEP Expert Panel suggestions
  - Update dated/obsolete/biased items
    - “Study the effect of acid rain on plants”
    - “Add up store receipts”
  - Link FYI to basic interests
    - Original charge: “Develop basic interests scales”

Basic Interests (from Su et al., 2019)

R	I	A	S	E	C
Agriculture	Life Science	Applied Arts and Design	Healthcare Service	Business Initiatives	Accounting
Animal Service	Mathematics/ Statistics	Creative Writing	Human Resources	Law	Finance
Athletics	Medical Science	Culinary Arts	Humanities and Foreign Language	Management/ Administration	Information Technology
Construction/ Woodwork	Physical Science	Media	Personal Service	Marketing/Advertising	Office Work
Engineering		Music	Religious Activities	Politics	
Mechanics/ Electronics		Performing Arts	Social Science	Professional Advising	
Outdoors		Visual Arts	Social Service	Public Speaking	
Physical/Manual Labor			Teaching/Education	Sales	
Protective Service					
Transportation/ Machine Operations					

R = Realistic, I = Investigative, A = Artistic, S = Social, E = Enterprising, C = Conventional

OPA
OFFICE OF PEOPLE ANALYTICS

FYI Form Development and Analysis

- HumRRO drafted 450 new FYI items for field testing beginning in 2019; effort driven by expert panel guidance
  - Focus on content validity, emphasizing construct coverage
  - Identify contemporary content related to emerging economic changes
  - Build on existing items with an enhanced item pool rated by a panel of experts
  - Identify Basic Interest Indicators, using Su et al. (2019) and the Strong Interests Inventory as frameworks for potential detailed basic interest markers
- "HumRRO employed Natural Language Processing (NLP) procedures to detect newly developed items, which might be considered 'enemies' or close clones of previously developed items" (Burke et al., p. 4).

OPA
OFFICE OF PEOPLE ANALYTICS

FYI Form Development and Analysis (cont.)

- Using field test data (230 of the 450 items were field tested), DTAC developed/proposed a new FYI form
  - Followed original FYI development process (Baker et al., 2010; Pommerich, 2004)
  - For each RIASEC scale, DTAC retained 7–10 items from the current form, adding 5–8 new items
    - Eight new items for Artistic; six new items for Enterprising
- Initial attempt
  - Used field test data (230 of the 450 items were field tested)
  - Followed original FYI development process (Baker et al., 2010; Pommerich, 2004)
    - Items selected based on item statistics and IRT item parameters
  - For each RIASEC scale, 7–10 items from the current form were retained, adding 5–8 new items
    - Eight new items for Artistic, six new items for Enterprising

FYI Form Development and Analysis (cont.)

- Following construction, we reviewed the form for content
  - Emphasis given to the basic interests taxonomy per guidance from the ASVAB CEP Expert Panel
- Only partial coverage (61%) of the 41 basic interests in Su et al.'s (2019) taxonomy
  - **3 of 10 for Realistic**
  - **3 of 4 for Investigative**
  - 6 of 7 for Artistic
  - **3 of 8 for Social**
  - 7 of 8 for Enterprising
  - 3 of 4 for Conventional
- O*NET has included Su et al.'s basic interests taxonomy in their recent update
- Given that CEP links to O*NET occupational information, HumRRO proposed two other options for the new form that would increase coverage of the basic interests

OPA
OFFICE OF PEOPLE ANALYTICS

FYI Form Development and Analysis (cont.)

- Three forms considered
  - Form Version 1
    - Assembled with focus on item statistics and IRT parameters
    - Retains majority of original FYI Items
  - Form Version 2
    - More focus on basic interests, but . . .
    - Retains mix of original and field test FYI items
  - Form Version 3
    - Primary focus on basic interests
    - “From scratch”
      - Items selected to ensure coverage of all basic interests; no requirement to retain any previous items
      - Retains 19 (21.1%) items from the current form

Original (Current) FYI Form—Basic Interest Coverage

Form (% BI)	RIASEC	Total BI	# BI Covered	% BI Covered	Missing BI
Original (61%)	Realistic	10	5	50.0	Animal Service, Athletics, Engineering, Outdoors, Protective Service
	Investigative	4	2	50.0	Mathematics/Statistics, Medical Science
	Artistic	7	6	85.7	Culinary Arts
	Social	8	3	37.5	Human Resources, Humanities/Foreign Language, Personal Service, Religious Activities, Social Science*
	Enterprising	8	6	75.0	Law, Professional Advising
	Conventional	4	3	75.0	Information Technology

*Social Science appears under Social in Su et al.'s (2019) taxonomy. We have chosen to include it under Investigative given our items' content and our choice to focus more on actions than context. Putka et al. (2023) include Social Science under both Social and Investigative interests.

OPA
OFFICE OF PEOPLE ANALYTICS

Proposed Form Version—Basic Interest Coverage

Form (% BI)	RIASEC	Total BI	# BI Covered	% BI Covered	Missing BI
Version 1 (61%)	Realistic	10	3	30.0	Agriculture, Animal Service, Athletics, Engineering, Outdoors, Protective Service, Transportation/Machine Operations
	Investigative	4	3	75.0	Mathematics/Statistics, Medical Science
	Artistic	7	6	85.7	Culinary Arts
	Social	8	3	37.5	Human Resources, Humanities/Foreign Language, Personal Service, Religious Activities, Social Science*
	Enterprising	8	7	87.5	Law
	Conventional	4	3	75.0	Information Technology
Version 2 (87.8%)	Realistic	10	6	60.0	Animal Service, Athletics, Outdoors, Transportation/Machine Operations
	Investigative	4	4	100.0	
	Artistic	7	7	100.0	
	Social	8	8	100.0	
	Enterprising	8	7	87.5	Sales
	Conventional	4	4	100.0	
Version 3 (100.0%)	Realistic	10	10	100.0	
	Investigative	**5**	**5**	100.0	
	Artistic	7	7	100.0	
	Social	**7**	**7**	100.0	
	Enterprising	8	8	100.0	
	Conventional	4	4	100.0	

*Social Science appears under Social in Su et al.'s (2019) taxonomy. We have chosen to include it under Investigative given the items' content and our choice to focus more on actions than context. These changes are reflected in the Total BI counts for Version 3 **(red text)**.

Form Version Reliability
(Internal Consistency—Cronbach's α)

Occupational Theme	Current Form	Form Version 1	Form Version 2	Form Version 3
Realistic	.95	.96	.95	.85
Investigative	.94	.94	.93	.89
Artistic	.91	.93	.92	.88
Social	.92	.93	.91	.82
Enterprising	.92	.92	.91	.87
Conventional	.94	.93	.90	.86

Lower internal consistency reliability estimates for the HumRRO forms, but this is ***desirable*** given the heterogeneity of each RIASEC dimension.

OPA
OFFICE OF PEOPLE ANALYTICS

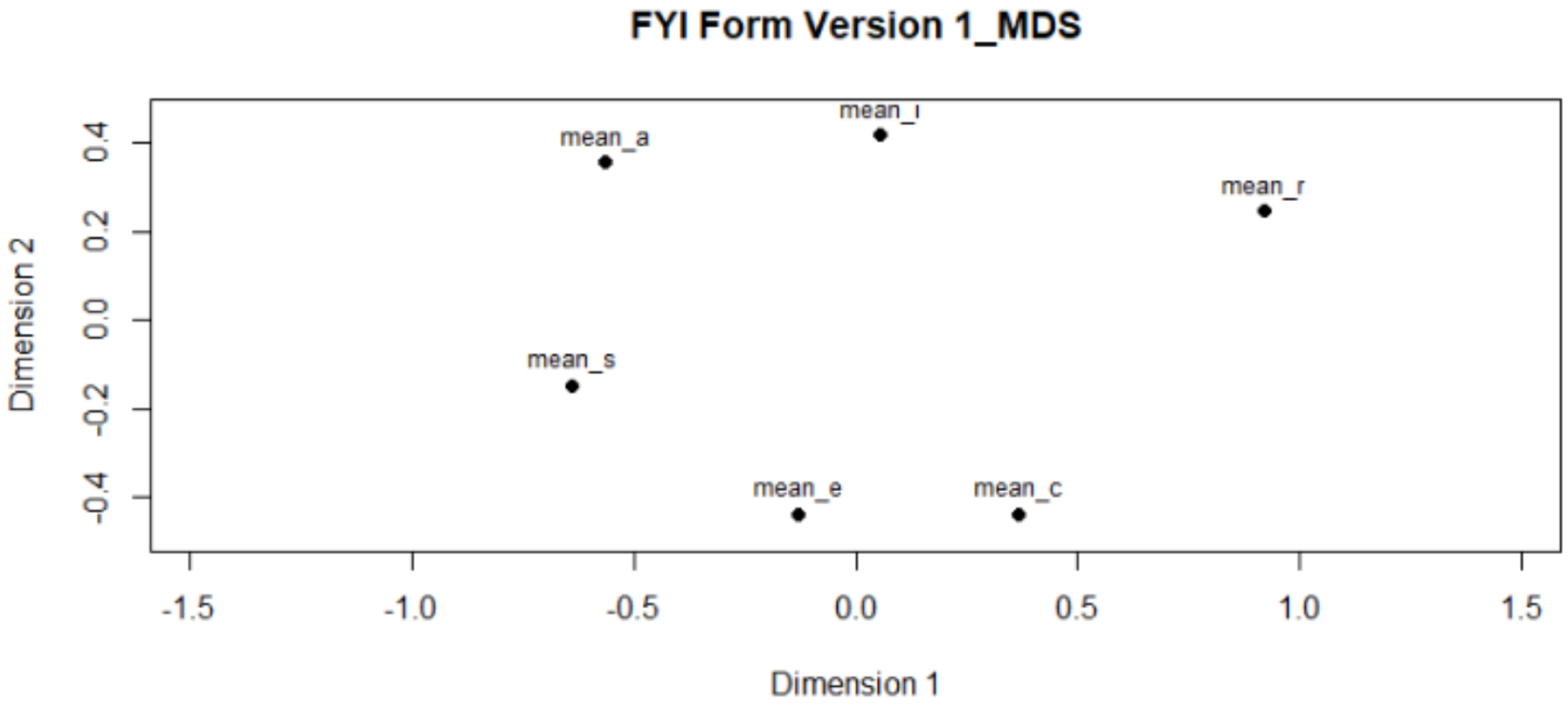
Form Version Sex Differences
(Cohen's d, Male–Female)

Occupational Theme	Current Form	Form Version 1	Form Version 2	Form Version 3
Realistic	0.88	0.91	0.90	0.72
Investigative	0.23	0.16	0.17	0.14
Artistic	−0.37	−0.32	−0.33	−0.30
Social	−0.72	−0.72	−0.70	−0.55
Enterprising	0.18	0.08	0.09	−0.02
Conventional	0.07	0.18	0.20	0.20

Smallest subgroup differences for our proposed Form Version 3. Higher Conventional *d* values due to inclusion of Information Technology.

OPA
OFFICE OF PEOPLE ANALYTICS

Multidimensional Scaling Results: Proposed Form Version 1



Model: Symmetric SMACOF

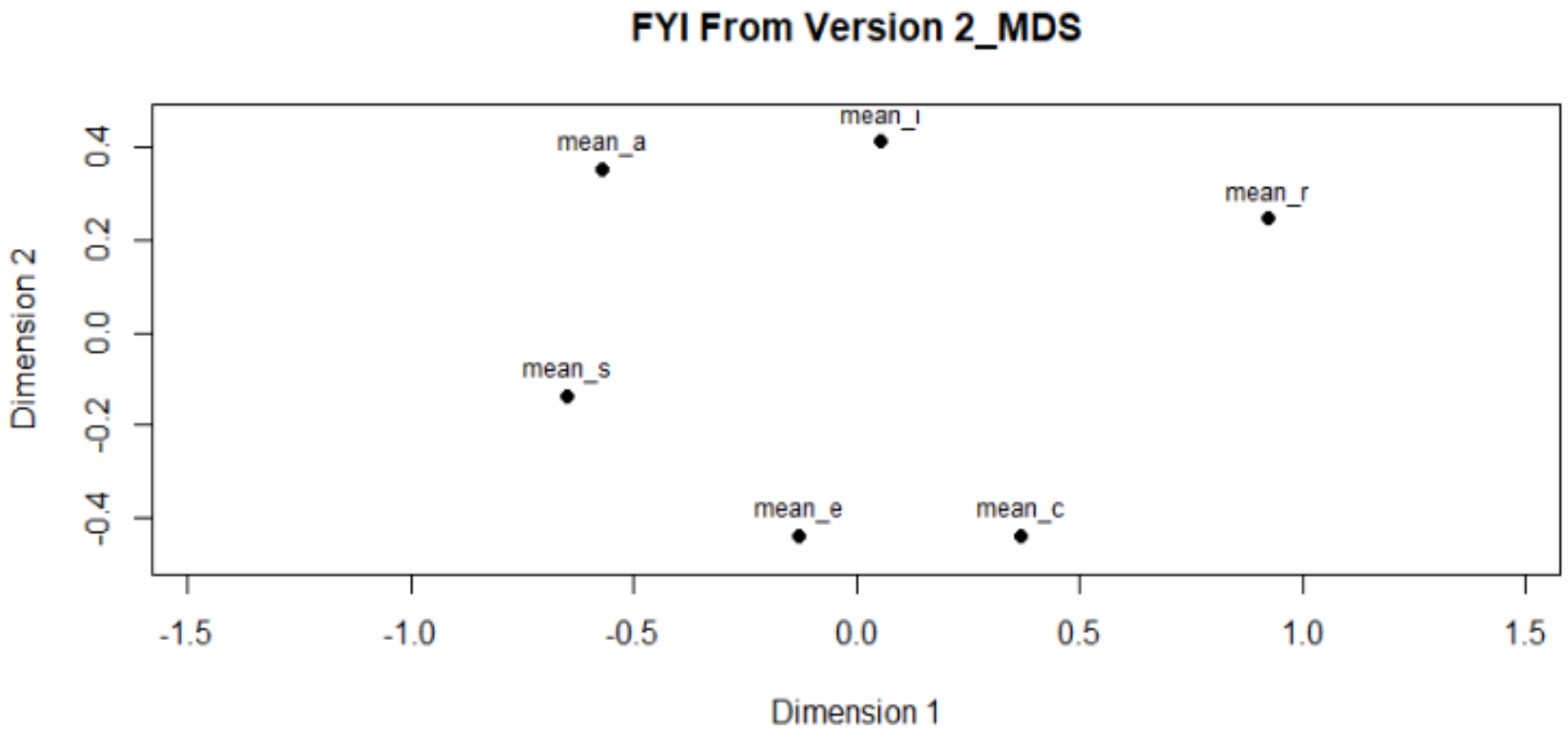
Number of objects: 6

Stress-1 value: 0.01

Number of iterations: 47

OPA
OFFICE OF PEOPLE ANALYTICS

Multidimensional Scaling Results: Proposed Form Version 2



Model: Symmetric SMACOF

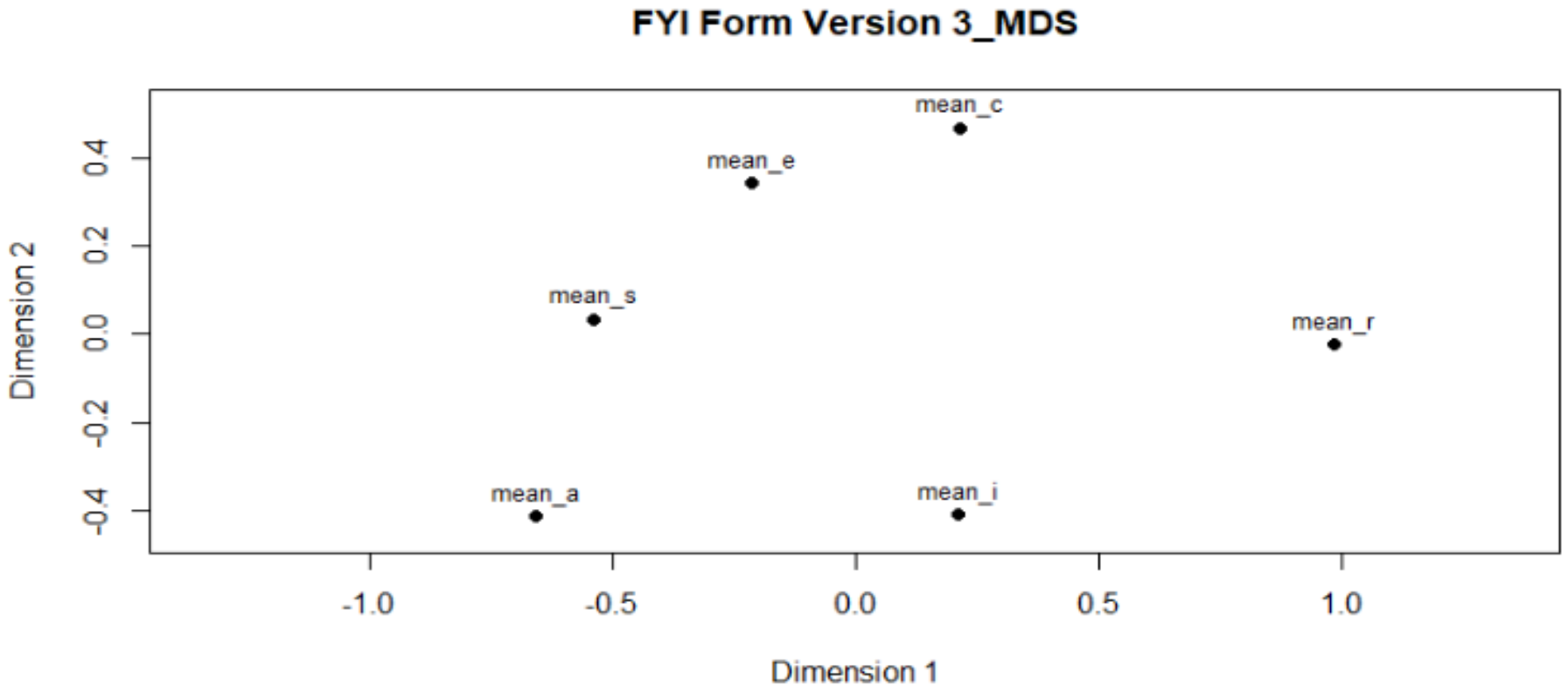
Number of objects: 6

Stress-1 value: 0.01

Number of iterations: 59

OPA
OFFICE OF PEOPLE ANALYTICS

Multidimensional Scaling Results: Proposed Form Version 3



Model: Symmetric SMACOF

Number of objects: 6

Stress-1 value: 0.004

Number of iterations: 91

OPA
OFFICE OF PEOPLE ANALYTICS

Summary: FYI

OPA
OFFICE OF PEOPLE ANALYTICS

Summary: FYI

- New FYI form
  - Recommend Form Version 3 (“from scratch”)
  - Provides strong psychometric characteristics
    - More reasonable internal consistency reliability estimates (still high, just not too high)
    - Smallest subgroup differences despite not purposefully selecting items with this criterion in mind
    - Complete coverage of the basic interests
  - Next steps
    - Finalize dimensionality analyses (item-level EFA; CFA models [standard, circumplex])
    - Field test and analyze new form
    - Establish norms for new form

Questions for the DAC: FYI

OPA
OFFICE OF PEOPLE ANALYTICS

Questions for the DAC: FYI

- What are your reactions to the new FYI form? Any concerns?
- Is there additional analysis/information you would like to see before field testing the proposed new form?
- What suggestions might you have for designing the field test of the new FYI form?
- What recommendations might you have for establishing norms (sex-based, total-group) for the new form?

Initial Analysis of Responses from the Work Values Situational Judgment Activity (WV SJA)

OPA
OFFICE OF PEOPLE ANALYTICS

A New Non-Cognitive Assessment for ASVAB CEP

- Goal: Explore the possibility of creating a work values assessment to add to the ASVAB CEP
- Work values tend to have greater meaning and utility for experienced workers
- The original idea was to introduce CEP participants to the concept of work values
  - Example: To facilitate discussions between students and counselors or teachers

OPA
OFFICE OF PEOPLE ANALYTICS

ASVAB CEP Work Values Situational Judgment Activity (WV SJA)

- Conducted a systematic review of pinnacle research publications
- Proposed various work values inventory formats
  - Ipsative, IRT-based scoring model pairing work values statements against one another
  - Situational policy-capture approach to measuring work values using regression-based methods for scoring
  - Multiple-choice item with basic mathematics for scoring
- Chose the third option in light of DTAC's valuing (a) administration time and (b) accessibility with paper-and-pencil administration

OPA
OFFICE OF PEOPLE ANALYTICS

Products of Our Development Work

- WV SJA
- Other proposed activities (versions were created for DTAC's review)
  - Realistic Job Preview
  - Personal Values and Work Values
  - The Intersection of Work Values and Work Interests
  - How Has the Pandemic Made You Think About What You Value?
  - Structured Interview

WV SJA

- Situational Judgment Test (SJT) assessing the **six work values** from the Theory of Work Adjustment (Dawis et al., 1964, 1968; Dawis & Lofquist, 1976, 1978)
- Introduces students to work values
- Linked to occupations (as are the ASVAB and FYI) to permit career exploration in terms of work values

Work Values and Their Definitions

Achievement—Workers who score high on Achievement are results-oriented. These workers often pursue jobs where employees are able to apply their strengths and abilities, which gives them a sense of accomplishment.

Independence—Workers who score high on Independence value the ability to approach work activities with creativity. These workers want to make their own decisions and plan their work with little supervision from a manager.

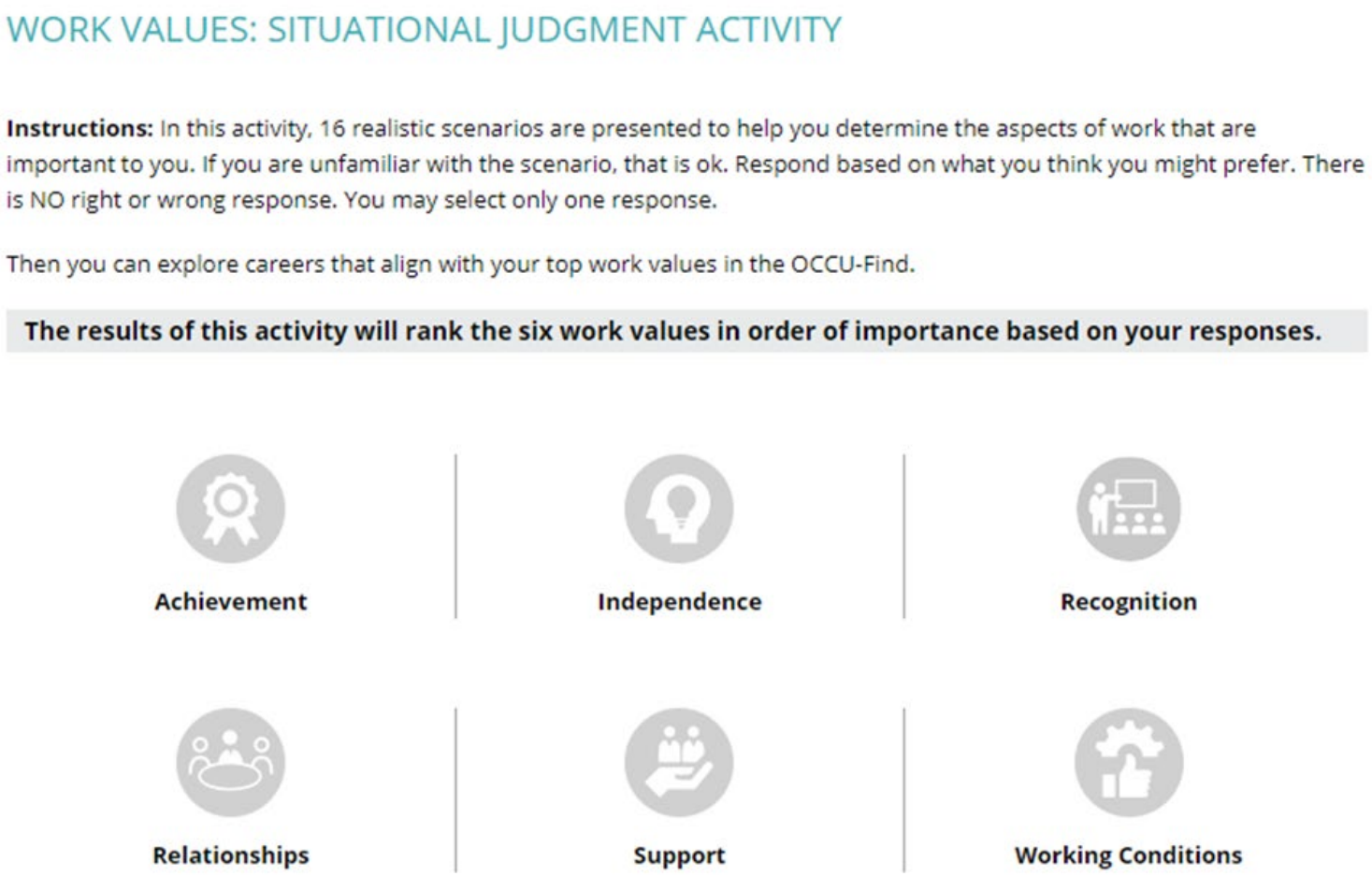
Recognition—Workers who score high on Recognition pursue jobs with opportunities for advancement and leadership responsibilities that allow them to give direction and instruction to others. These workers are often considered prestigious by their peers and others in their organization and receive recognition for the work they contribute.

Relationships—Workers who score high on Relationships prefer jobs that provide services to others and working with co-workers in a friendly, non-competitive environment. Workers in these jobs value getting along well with others and do not like to be pressured to do things that go against their morals or sense of what is right and wrong.

Support—Workers who score high on Support appreciate when their company's leadership stands behind and supports their employees. People in these types of jobs like to feel they are being treated fairly by the company and have supervisors who spend time and effort training their workers to perform well.

Working Conditions—Workers who score high on Working Conditions value job security and pleasant working conditions. These workers enjoy being busy and want to be paid well for the work they do. They enjoy developing ways of doing things with little or no supervision and depend on themselves to get the work done. They pursue steady employment that offers something different to do on a daily basis.

WV SJA: Introductory Screen



OPA
OFFICE OF PEOPLE ANALYTICS

Situational Judgment Test Format

Achievement

School Context

Your school requires you to fulfill a certain amount of internship hours in order to graduate. Which internship opportunity do you prefer most?

- An internship where you feel a sense of accomplishment from the work you do.
- An internship that will give you the opportunity to advance in your next job.
- An internship where you have a supportive supervisor.
- An internship that will keep you busy.
- An internship where you get along well with your coworkers.
- An internship where you are able to try out your ideas.

Work Context

Your supervisor has set up a meeting to discuss your performance over the last 6 months at work. Which type of feedback would you value most?

- A supervisor telling you that you are supported by management.
- A supervisor telling you that you worked on a variety of projects.
- A supervisor telling you that you are good at making independent decisions.
- A supervisor acknowledging your goal achievement and setting new goals for the next six months.
- A supervisor telling you that you are highly regarded by your peers.
- A supervisor telling you that you get along well with your co-workers.

Relationships

OPA
OFFICE OF PEOPLE ANALYTICS

WV SJA Results Page

YOUR WORK VALUE ORDER

1 Support

Workers who score high on Support appreciate when their company's leadership stands behind and supports their employees. People in these types of jobs like to feel like they are being treated fairly by the company and have supervisors who spend time and effort training their workers to perform well.

2 Relationships

Workers who score high on Relationships prefer jobs that provide services to others and working with co-workers in a friendly, non-competitive environment. Workers in these jobs value getting along well with others and do not like to be pressured to do things that go against their morals or sense of what is right and wrong.

3 Recognition

Workers who score high on Recognition pursue jobs with opportunities for advancement and leadership responsibilities that allow them to give direction and instruction to others. These workers are often considered prestigious by their peers and others in their organization and receive recognition for the work they contribute.

START EXPLORING

Retake Assessment

OPA
OFFICE OF PEOPLE ANALYTICS

WV SJA Ties Report

YOUR WORK VALUES ARE TIED IN SOME AREAS

Two or more of your work values ranked the same.

Which sounds most like you? Select one of the four tied values.

○ **Independence**		Workers who score high on Independence value the ability to approach work activities with creativity. These workers want to make their own decisions and plan their work with little supervision from a manager.
○ **Relationships**		Workers who score high on Relationships prefer jobs that provide services to others and working with co-workers in a friendly, non-competitive environment. Workers in these jobs value getting along well with others and do not like to be pressured to do things that go against their morals or sense of what is right and wrong.
○ **Support**		Workers who score high on Support appreciate when their company's leadership stands behind and supports their employees. People in these types of jobs like to feel like they are being treated fairly by the company and have supervisors who spend time and effort training their workers to perform well.
○ **Working Conditions**		Workers who score high on Working Conditions value job security and pleasant working conditions. These workers enjoy being busy and want to be paid well for the work they do. They enjoy developing ways of doing things with little or no supervision and depend on themselves to get the work done. These workers pursue steady employment that offers something different to do on a daily basis.

OPA
OFFICE OF PEOPLE ANALYTICS

Preliminary Results

OPA
OFFICE OF PEOPLE ANALYTICS

Analysis of WV SJA Response Data

- Currently have > 42k responses (uncleaned data)
- Initial results
  - Modal response profiles
  - Differences by sex, context (i.e., school, work)

OPA
OFFICE OF PEOPLE ANALYTICS

WV SJA Analysis Sample: Demographics

Demographic	Demographic Detail	Total Number of Students
Sex	Male	20,110
	Female	20,301
	NA	2,130
Education	10	9,444
	11	21,803
	12	8,916
	13	99
	14	22
	15	139
	NA	2,118
Race/Ethnicity	American Indian	1,961
	Asian	1,612
	African	2,593
	Native Hawaiian	469
	White	24,011
	Hispanic	6,888
	Not Hispanic	21,636
Total	**Total**	**42,541**

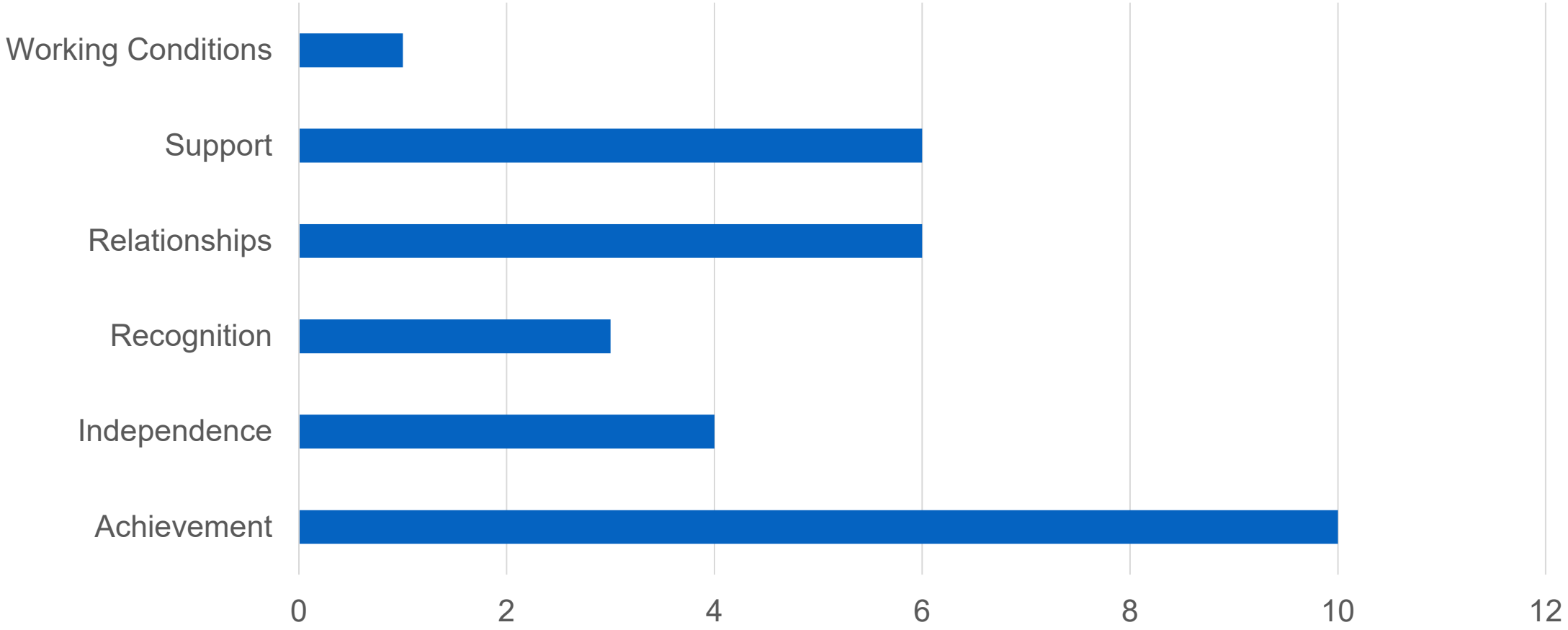
OPA
OFFICE OF PEOPLE ANALYTICS

Top Work Values Profiles

Rank Order	Work Value (1st Position)	Work Value (2nd Position)	Work Value (3rd Position)	Number of Students
1	Relationships	Support	Achievement	948
2	Achievement	Relationships	Support	799
3	Achievement	Independence	Recognition	729
4	Relationships	Achievement	Support	708
5	Achievement	Support	Relationships	692
6	Achievement	Recognition	Independence	666
7	Support	Relationships	Achievement	624
8	Achievement	Working Conditions	Independence	595
9	Recognition	Achievement	Independence	536
10	Support	Achievement	Relationships	535

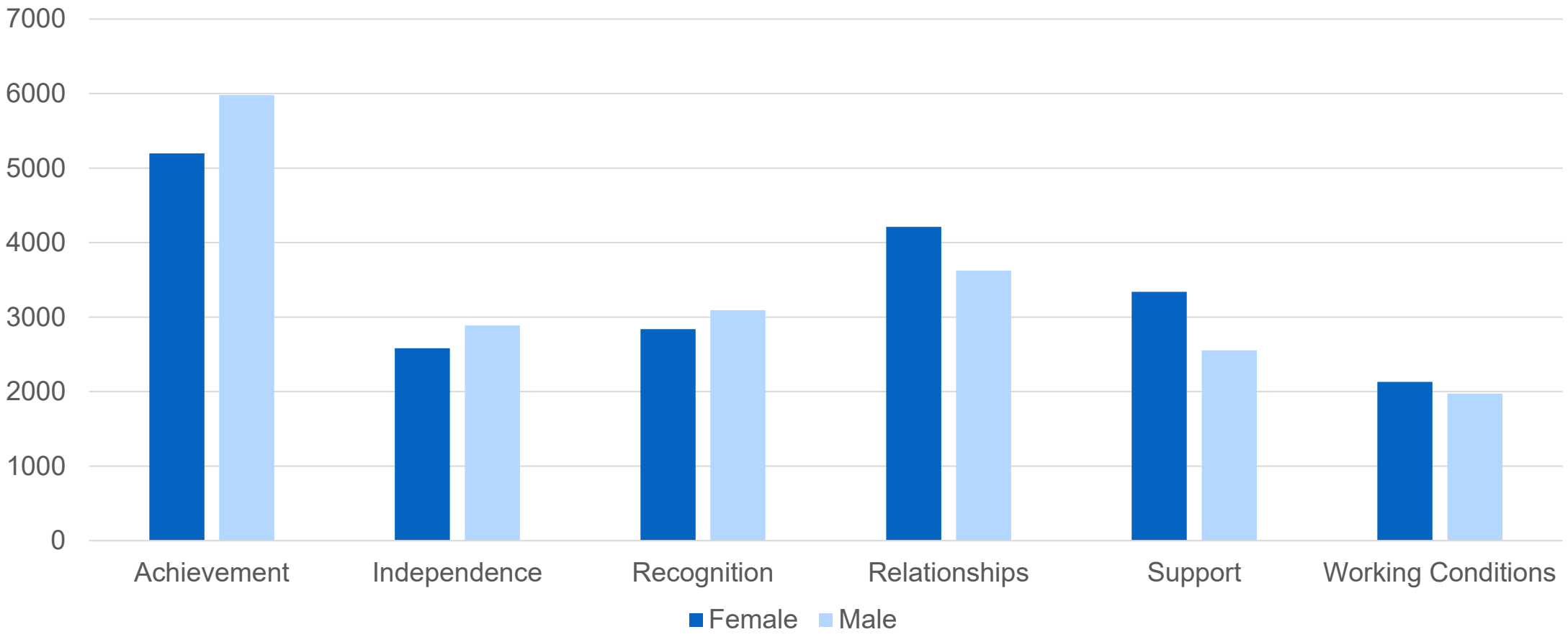
OPA
OFFICE OF PEOPLE ANALYTICS

Work Values Occurrence in Top Ten Profiles



OPA
OFFICE OF PEOPLE ANALYTICS

Top Work Values by Sex



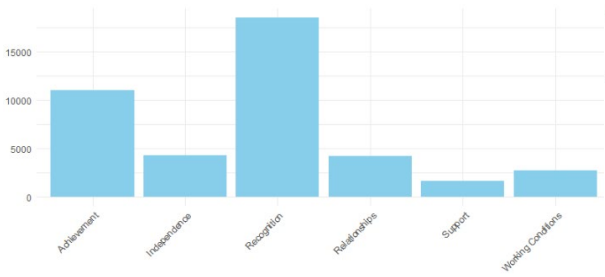
OPA
OFFICE OF PEOPLE ANALYTICS

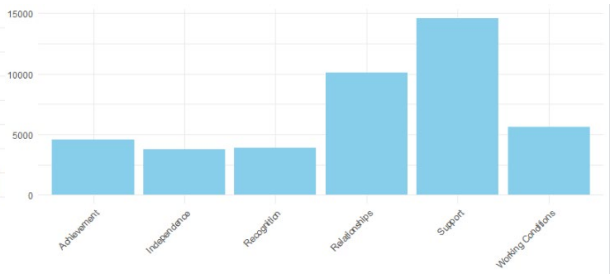
Top Work Values Profile by Sex

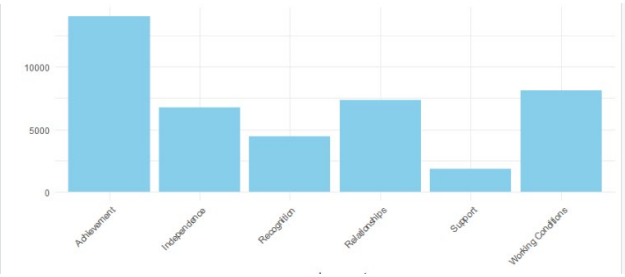
Female	Male
1st: Relationships, Support, Achievement	1st: Achievement, Independence, Recognition
2nd: Achievement, Relationships, Support	2nd: Achievement, Recognition, Independence
3rd: Relationships, Achievement, Support	3rd: Relationships, Support, Achievement

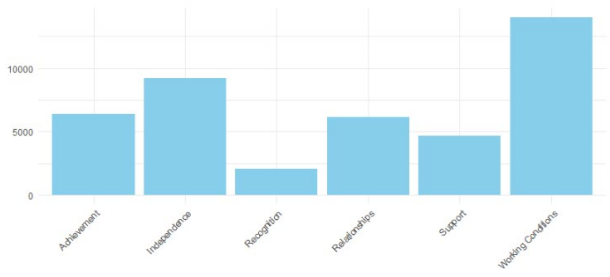
OPA
OFFICE OF PEOPLE ANALYTICS

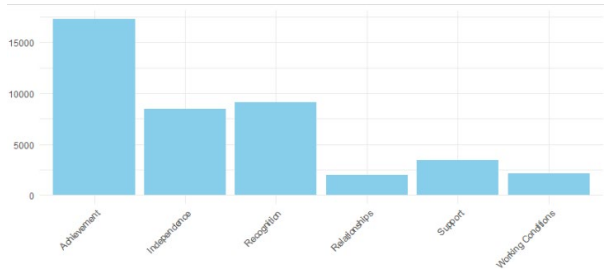
WV SJA Item Endorsement: School Context

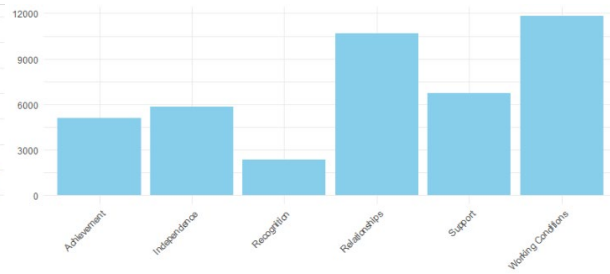


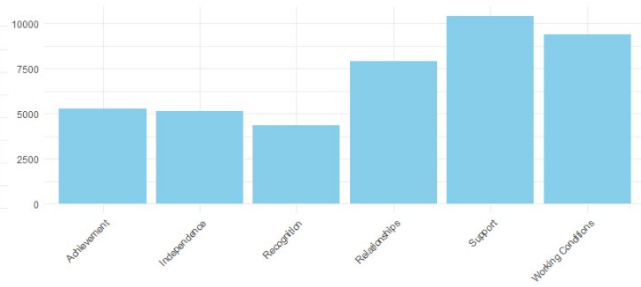


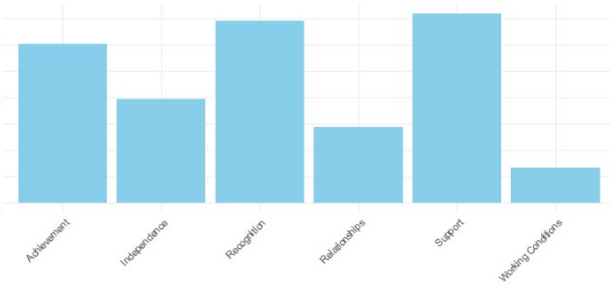






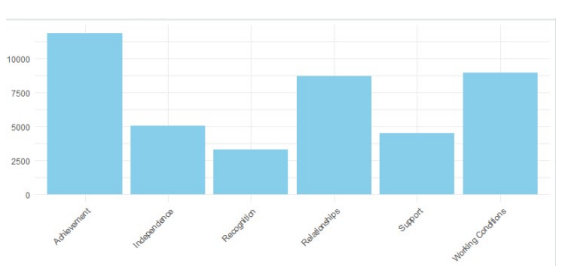


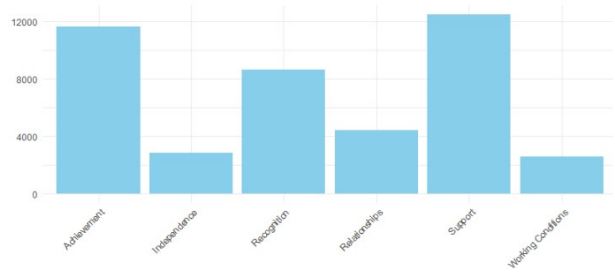


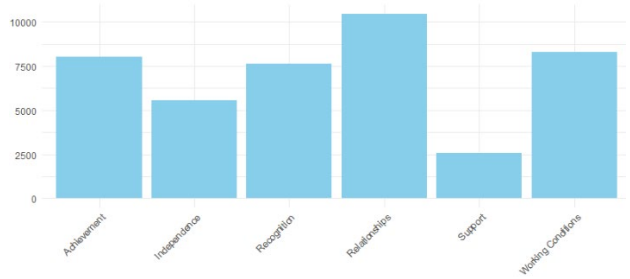


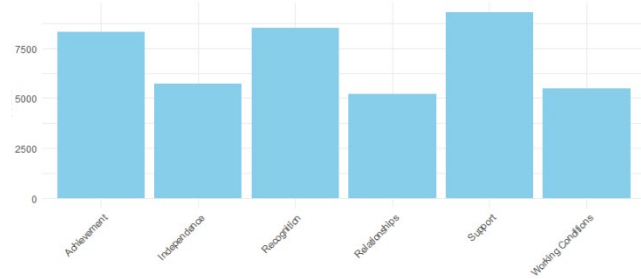
OPA
OFFICE OF PEOPLE ANALYTICS

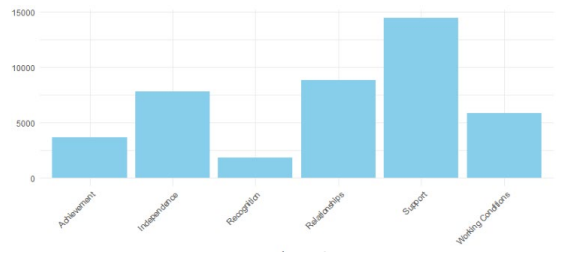
WV SJA Item Endorsement: Work Context

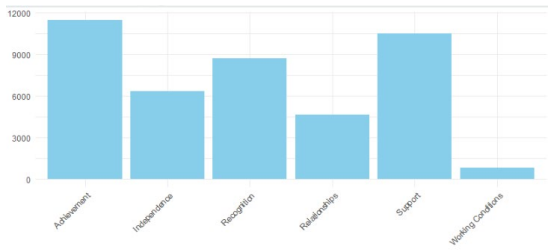


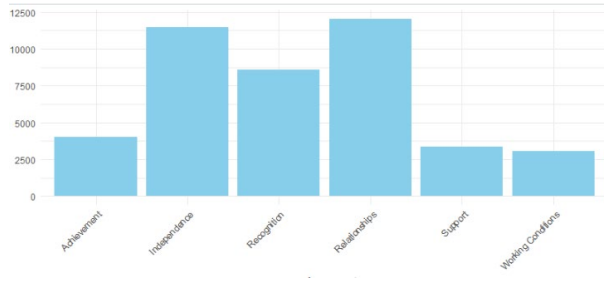


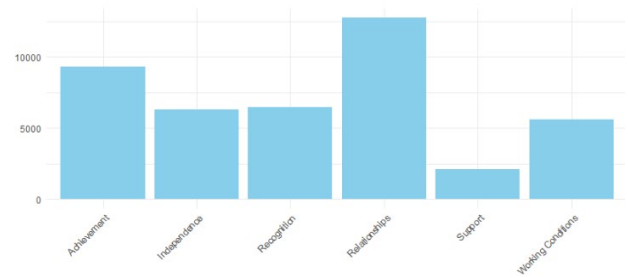












OPA
OFFICE OF PEOPLE ANALYTICS

WV SJA Average Endorsement Between Contexts

Work Value	School Context Mean	Work Context Mean	Significance
Achievement	1.514	1.801	Significant
Independence	1.318	1.048	Significant
Recognition	1.280	1.278	Not significant
Relationships	1.556	1.263	Significant
Support	1.252	1.417	Significant
Working Conditions	1.078	1.193	Significant

OPA
OFFICE OF PEOPLE ANALYTICS

Questions for the DAC: WV SJA

OPA
OFFICE OF PEOPLE ANALYTICS

Questions for the DAC: WV SJA

- Given the respondent population, should the WV SJA focus on a single context (i.e., work vs. school)?
- Do you have concerns with using the WV SJA to identify occupational matches?

ASVAB CAREER EXPLORATION PROGRAM

asvabprogram.com

Use access code **CEP4ME** to create an account.

CAREERS IN THE MILITARY
Powered by ASVAB CEP

careersinthemilitary.com

States considering incorporating the ASVAB CEP into their ESSA plans are encouraged to visit:
https://www.asvabprogram.com/legislation

Or contact

Irina Rader, EdD

irina.v.rader.civ@mail.mil or dodhra.asvab-cep@mail.mil

OPA
OFFICE OF PEOPLE ANALYTICS

Appendix

OPA
OFFICE OF PEOPLE ANALYTICS

Current FYI Form—Items and BIs Assessed

Items with yellow highlighting were retained and considered for use with the new FYI form.

Original FYI Form		
Item # (from form) and Item Stem	Basic Interest	BI Coverage
04. Adjust bicycle gears	Mechanics/Electronics	
10. Repair a leaky faucet	Construction/Woodwork	
16. Install kitchen cupboards	Construction/Woodwork	
22. Operate a farm	Agriculture	
28. Apply wood stains and varnishes to furniture	Construction/Woodwork	
34. Repair household appliances	Construction/Woodwork	
40. Build a deck for a house	Construction/Woodwork	
46. Tile a kitchen floor	Construction/Woodwork	
52. Use carpentry tools	Construction/Woodwork	
58. Build a stone wall	Construction/Woodwork	
64. Operate a riding mower	Transportation/Machine Operations	
70. Refinish the floors in a house	Construction/Woodwork	
76. Detail a car	Physical/Manual Labor	
82. Assemble playground equipment	Construction/Woodwork	
88. Frame a house	Construction/Woodwork	50.0%
03. Investigate stars and black holes	Physical Science	
09. Discover a new strain of virus	Life Science	
15. Test DNA samples	Life Science	
21. Explore ancient ruins	Physical Science (iffy for Investigative)	
27. Study an active volcano	Physical Science	
33. Identify an unknown chemical substance	Physical Science	
39. Conduct lab experiments	All	
45. Study environmental science	Physical Science	
51. Predict earthquakes	Physical Science	
57. Analyze ocean currents	Physical Science	
63. Study the effects of acid rain on plants	Life Science	
69. Observe and classify a new species	Life Science	
75. Study planetary storms	Physical Science	
81. Observe and record animal life cycles	Life Science	
87. Study changes in Earth's atmosphere	Physical Science	50.0%
01. Attend an art class	Visual Arts (iffy)	
07. Act on stage	Performing Arts	
13. Write a movie script	Creative Writing	
19. Compose music	Music	
25. Illustrate a book	Visual Arts	
31. Design a set for a play	Applied Arts and Design	
37. Play a role in a musical	Performing Arts	
43. Attend a poetry reading	(this is interest in poetry, not doing it)	
49. Design a museum exhibit	Applied Arts and Design	
55. Create sculptures	Visual Arts	
61. Direct a musical	Performing Arts	
67. Paint portraits	Visual Arts	
73. Write a short story	Creative Writing	
79. Film a documentary	Media	
85. Play in a jazz band	Music	85.7%
02. Help children with after-school homework	Teaching/Education	
08. Serve as a playground activity leader	Social Service	
14. Help people cope with loss	Social Service	
20. Volunteer for a local community service	Social Service	
26. Assist a teacher in the classroom	Teaching/Education	
32. Organize activities at a community center	Social Service	
38. Teach people how to cope with stress	Social Service	
44. Counsel others about substance abuse	Social Serivce	
50. Help people resolve personal problems	Social Serivce	
56. Take care of a disabled person	Healthcare Serivce	
62. Teach parenting skills	Teaching/Education	
68. Serve as a dormitory counselor	Social Serivce	
74. Lead a group therapy session	Social Serivce	
80. Mentor a troubled child	Teaching/Education	
86. Reassure a nervous patient	Healthcare Serivce	37.5%
06. Chair a committee meeting	Management/Administration	
12. Persuade committee members on an issue	Politics	
18. Campaign for a political office	Politics	
24. Manage a department in a company	Management/Administration	
30. Conduct a business seminar	Public Speaking	
36. Market new products to retail businesses	Marketing/Advertising	
42. Give a sales presentation	Sales	
48. Invest in new companies	Business Initiatives	
54. Recruit new customers for a business	Marketing/Advertising	
60. Give a press conference	Public Speaking	
66. Persuade someone to finance a business	Business initiatives	
72. Sell residential and business properties	Sales	
78. Publicize an event	Marketing/Advertising	
84. Plan meetings and conferences	Management/Administration	
90. Serve as a company's spokesperson	Public Speaking	75.0%
05. Count and balance a cash drawer	Accounting	
11. Enter data in an accounting ledger	Accounting	
17. Count the inventory of a small business	Accounting	
23. Do accounting for a business	Accounting	
29. Process company payrolls	Accounting	
35. Prepare bank deposits	Accounting	
41. Add up store receipts	Accounting	
47. Type legal papers and documents	Office Work	
53. Organize and maintain personnel files	Office Work	
59. Compute fees and charges	Accounting	
65. Review financial records	Finance	
71.Enter data in a database	Office Work	
77. Prepare bills and invoices	Accounting	
83. Maintain paper and electronic data files	Office Work	
89. Record business transactions	Accounting	75.0%

OPA
OFFICE OF PEOPLE ANALYTICS

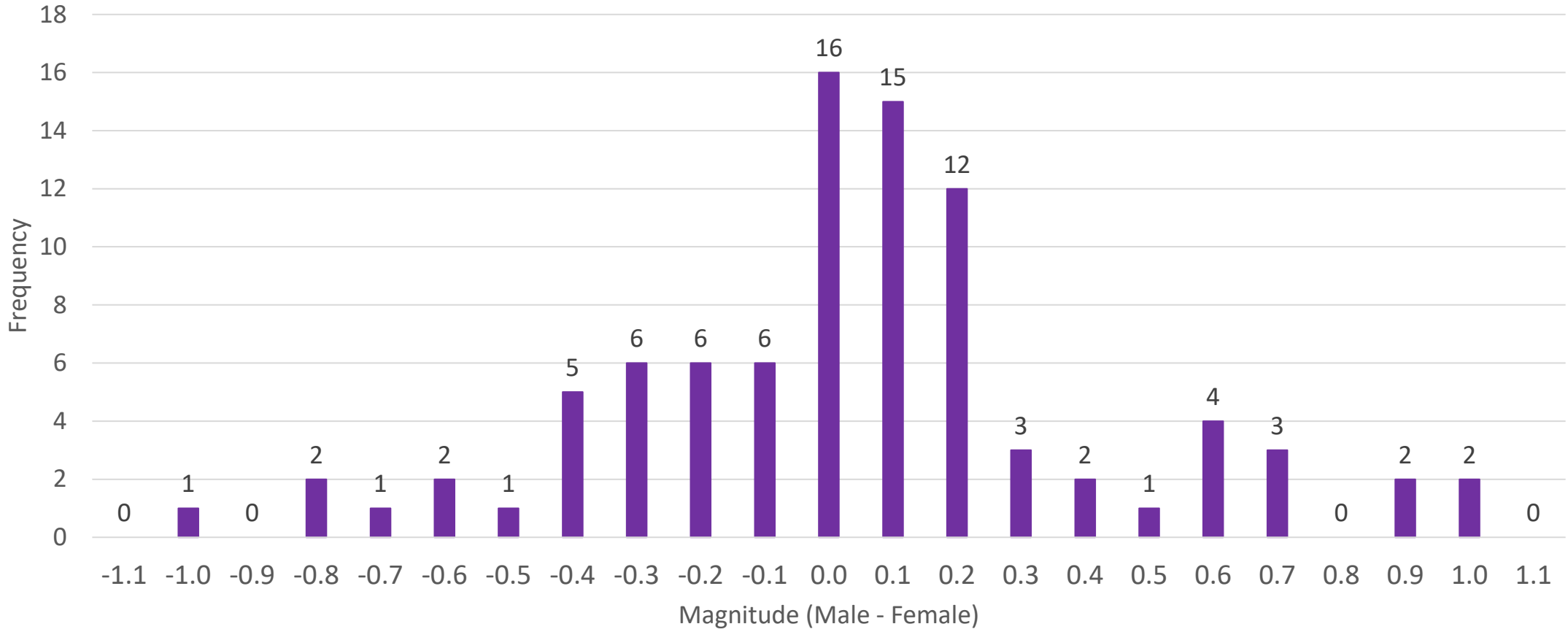
HumRRO's Proposed FYI Form—Items, BIs Assessed, and *d* Values

Items with magenta highlighting are on the current FYI form (19/90 = 21.1%)

Item ID	Item Stem	Basic Interest	Cohen's d (M-F)
UID5	Apply wood stains and varnishes to furniture	Construction/Woodwork	0.41
UID15	Frame a house	Construction/Woodwork	0.78
UID94	Train athletes	Athletics	0.28
UID98	Fight forest fires	Protective Service	0.60
UID99	Maintain hiking trails	Outdoors	0.30
UID101	Move furniture	Physical/Manual Labor	0.57
UID105	Groom an animal	Animal Service	-0.42
UID108	Plant crops	Agriculture	0.23
UID111	Treat an animal's wound	Animal Service	-0.22
UID115	Inspect equipment for defects	Engineering	0.78
UID117	Install electrical wiring	Mechanics/Electronics	1.00
UID119	Repair a computer	Engineering	0.98
UID122	Track and monitor wildlife	Outdoors	0.21
UID124	Pilot an aircraft	Transportation/Machine Operations	0.66
)131	Service a car	Mechanics/Electronics	0.98
UID17	Discover a new strain of virus	Medical Science	-0.04
UID21	Identify an unknown chemical substance	Physical Science	0.17
UID22	Conduct lab experiments	All	0.17
UID27	Observe and classify a new species	Life Science	0.13
UID30	Study changes in Earth's atmosphere	Physical Science	0.20
UID134	Forecast the amount of products a store should order	Mathematics/Statistics	0.20
UID136	Discover patterns in data	Mathematics/Statistics	0.28
UID139	Study how species evolve	Life Science	0.12
UID152	Study how medicine affects the body	Medical Science	-0.33
UID153	Conduct a scientific study	All	0.18
UID157	Diagnose a mental illness	Medical Science	-0.67
UID159	Create graphs to explain survey results	Mathematics/Statistics	0.10
UID160	Develop an internet survey	Social Science	0.09
UID162	Calculate the economic impact of a new law	Social Science	0.20
)143	Research environmental problems	Physical Science	-0.01
UID32	Act on stage	Performing Arts	-0.25
UID33	Write a movie script	Creative Writing	-0.09
UID35	Illustrate a book	Visual Arts	-0.39
UID44	Film a documentary	Media	-0.01
UID176	Create a new dessert	Culinary Arts	-0.31
UID178	Create a podcast	Media	0.09
UID182	Perform a song	Music	-0.09
UID183	Play a musical instrument	Music	-0.01
UID194	Create a music video	Media	-0.03
UID195	Design an office space	Applied Arts and Design	-0.31
UID199	Participate in a dance class	Performing Arts	-0.84
UID200	Decorate a cake	Culinary Arts	-0.80
UID203	Apply makeup to someone	Applied Arts and Design	-1.00
UID205	Draft a cartoon strip	Visual Arts	-0.01
)207	Write a fictional story	Creative Writing	-0.25
UID222	Bring food to those in need	Social Service	-0.44
UID51	Organize activities at a community center	Social Service	-0.52
UID247	Teach at an elementary school	Teaching/Education	-0.71
UID54	Help people resolve personal problems	Social Service	-0.49
UID166	Use artifacts to understand an ancient civilization	Humanities and Foreign Language	0.28
UID218	Monitor the health of a patient	Healthcare Service	-0.63
UID217	Assist a patient with mobility	Healthcare Service	-0.37
UID224	Help others improve their work	Human Resources	-0.16
UID231	Educate employees about a new policy	Human Resources	-0.02
UID239	Serve as an interpreter	Humanities and Foreign Language	-0.16
UID241	Provide spiritual guidance	Religious Activities	-0.30
UID243	Provide personal training at a gym	Personal Service	0.37
UID244	Lead a prayer service	Religious Activities	0.01
UID248	Tutor a student	Teaching/Education	-0.47
)250	Serve others beverages	Personal Service	-0.29
UID64	Manage a department in a company	Management/Administration	0.16
UID73	Publicize an event	Marketing/Advertising	-0.14
UID74	Plan meetings and conferences	Management/Administration	-0.22
UID252	Start a business	Business Initiatives	0.19
UID259	Promote a new policy	Politics	0.11
UID261	Resolve a customer complaint	Sales	-0.04
UID263	Communicate a company's strategy	Public Speaking	0.21
UID264	Persuade a jury	Law	-0.13
UID268	Interpret the law	Law	0.07
UID273	Coach a sports team	Professional Advising	0.22
UID274	Promote a product	Marketing/Advertising	0.10
UID278	Persuade the public to support an issue	Politics	-0.11
UID280	Lead a workshop on professional achievement	Professional Advising	0.33
UID282	Convince others to try a product	Sales	0.11
)283	Act as a spokesperson for a group	Public Speaking	-0.06
UID80	Process company payrolls	Accounting	0.23
UID84	Organize and maintain personnel files	Office Work	-0.15
UID88	Prepare bills and invoices	Accounting	0.16
UID289	Review financial transactions	Finance	0.28
UID290	Prepare financial reports for a business	Finance	0.14
UID291	Prepare someone's taxes	Accounting	0.12
UID293	Prepare a budget	Accounting	0.15
UID294	Create computer code	Information Technology	0.60
UID295	Build computers	Information Technology	1.00
UID296	Estimate the cost of a product	Finance	0.47
UID303	Document steps taken during a study	Office Work	-0.09
UID306	Manage someone else's schedule	Office Work	-0.43
UID316	Monitor security technology	Information Technology	0.77
UID319	Record court proceedings	Office Work	-0.29
)321	Program computer updates	Information Technology	0.68

OPA
OFFICE OF PEOPLE ANALYTICS

Frequency Chart of HumRRO FYI Form Item Effect Sizes (n_i = 90)



Mean effect size = 0.05
61.1% of effect sizes fall between −0.3 < d < 0.3

d > 0 = 48 (Males score higher)
d < 0 = 42 (Females score higher)

OPA
OFFICE OF PEOPLE ANALYTICS

Tab S

OPA

OFFICE OF PEOPLE ANALYTICS

Future Topics

Mary Pommerich

Defense Testing & Assessment Center

Briefing presented to the DACMPT

January 23, 2025

Possible Future Topics

- ASVAB evaluations
  - Assembling Object dimensionality
  - Differential prediction analyses for composite scores
  - Impact of COVID/score trends for Enlisted Testing Program (ETP) & Career Exploration Program (CEP)
- CAT-ASVAB/Form development methodology
  - Evaluation of Differential Item Functioning methodology
- Unproctored testing
  - Pending Internet Computerized Adaptive Test (PiCAT)/Verification test (VTest) updates
  - AFQT Predictor Test (APT)
- Super-scoring
- Adding new non-cognitive measures
  - Military compatibility assessment for the officer population
  - Criterion domain/performance metrics for military compatibility
- Calculator effort
  - Impact study results
  - Simulation study results
  - Needs assessment results
- ASVAB validity
  - Updates to ASVAB validity framework
  - Updates to TAPAS validity framework
- Explore AI/GAI/technology advancements
  - Status report on ASVAB effort
  - Status report on non-cognitive effort
- Next generation testing
  - Roadmap update
  - High school curriculum study update
- Adding new cognitive tests/composites
  - Complex Reasoning update
  - Computational Thinking validation study