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

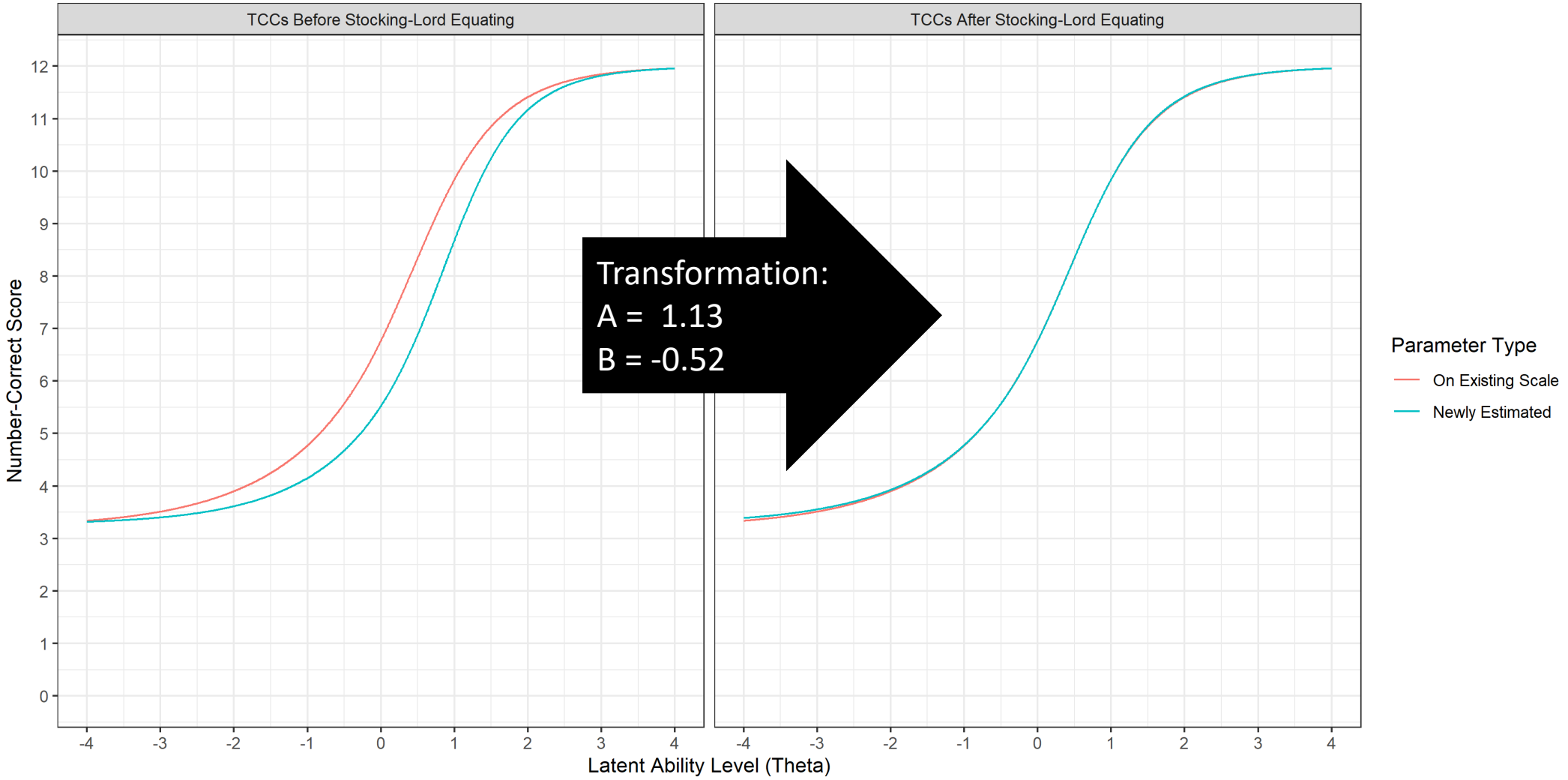
Background and Motivation (Continued)

- 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

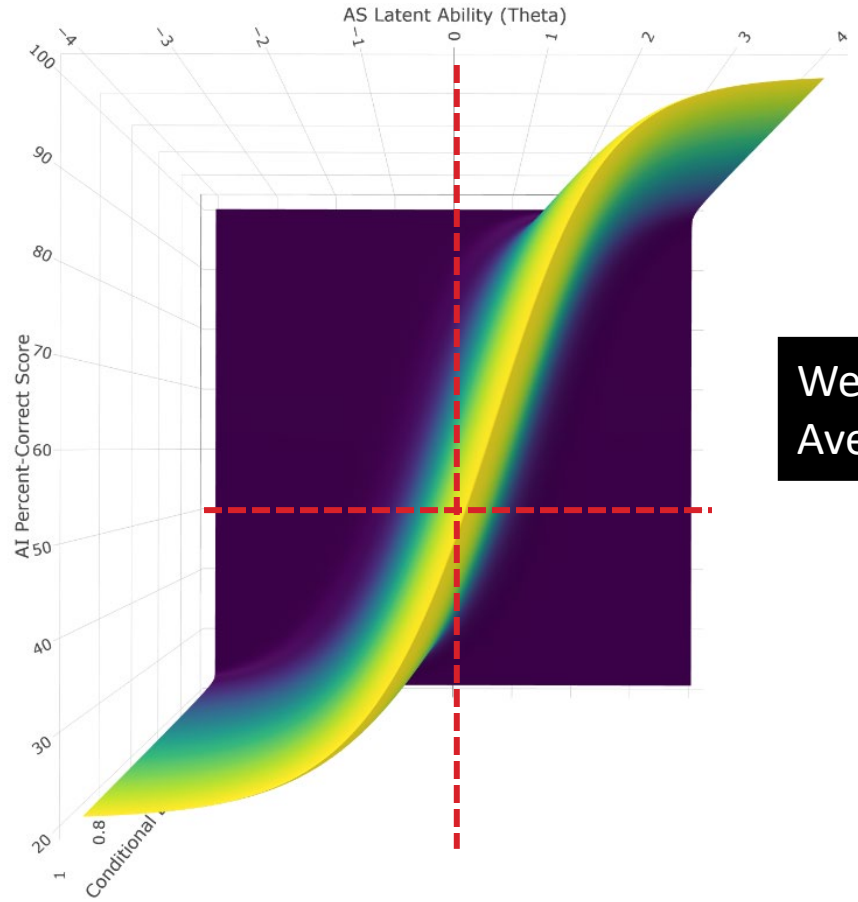
If anchor items can provide the scaling information needed to rescale item parameters, could we use alternative scale-anchoring information to achieve the same effect?

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

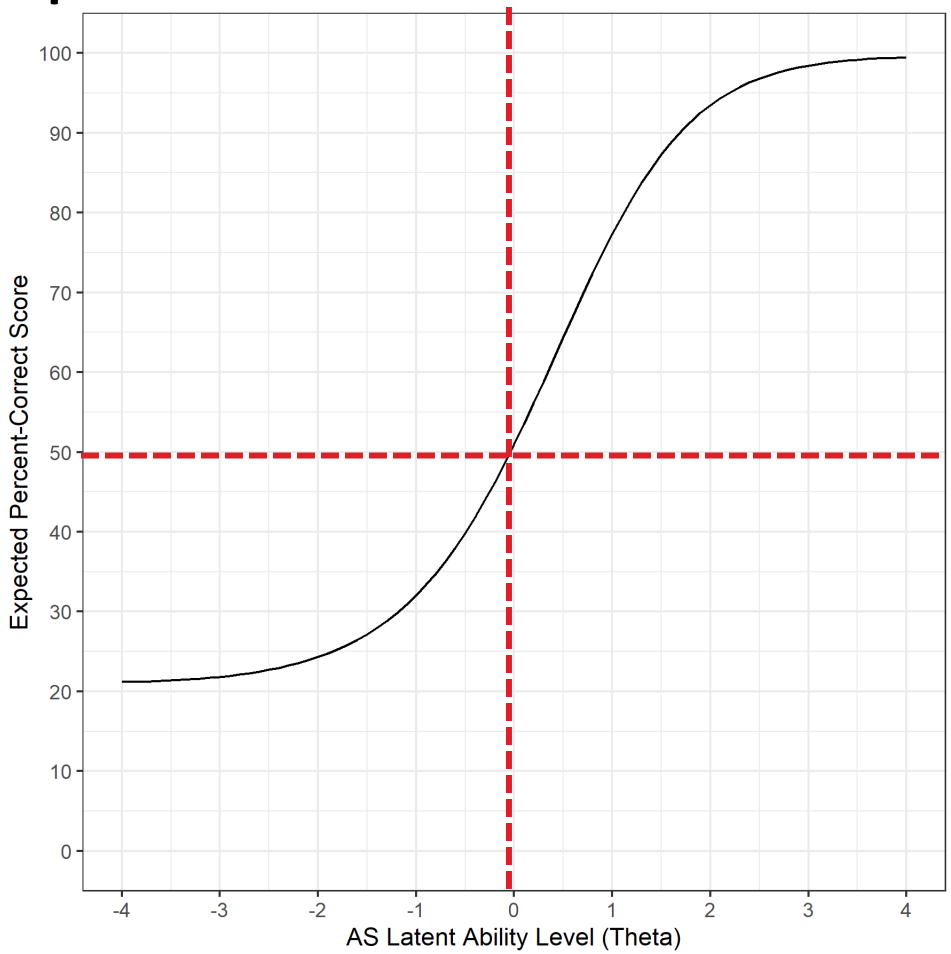
Deriving a Target TCC from Distributional Information About Abilities

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)

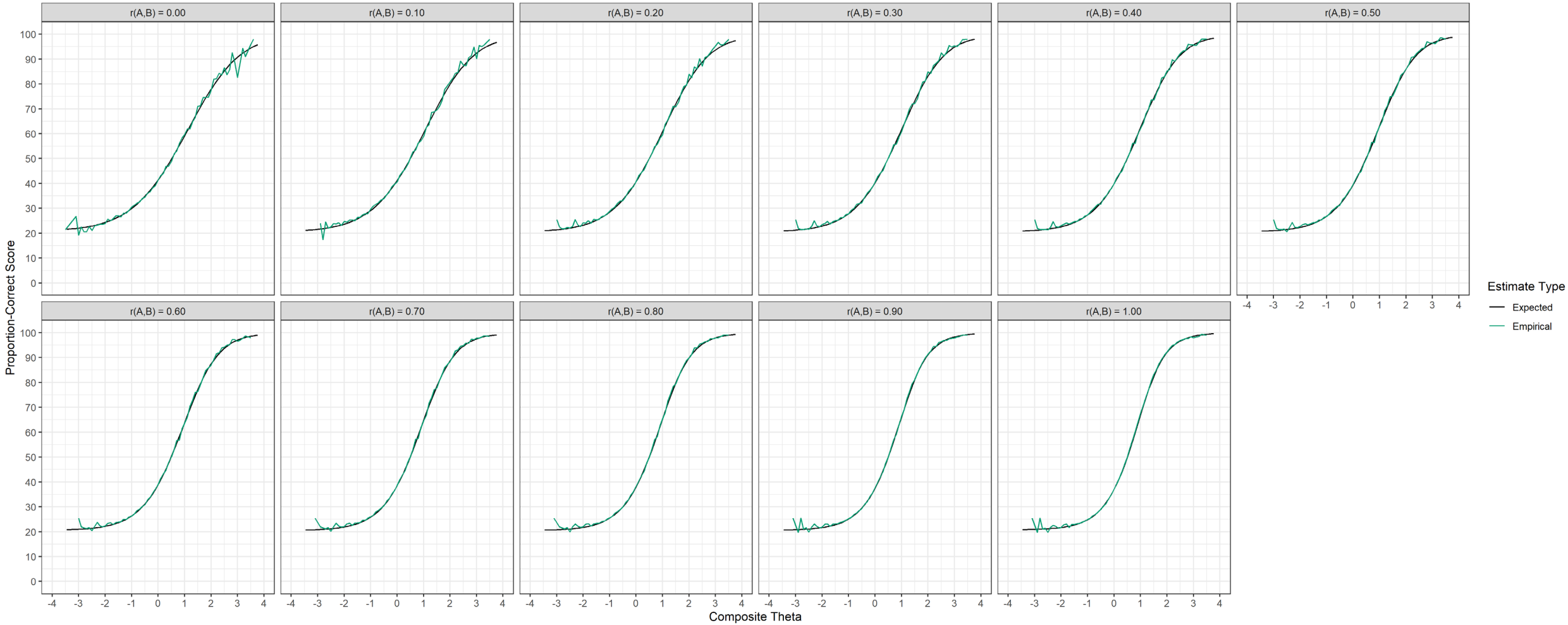
- Simulated item responses using person and item parameters
  - For each simulee-item combination, the simulee's true theta and the item's true IRT parameters were used to estimate the probability of a correct response
  - To introduce measurement error, simulee's probabilities of correct responses were 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

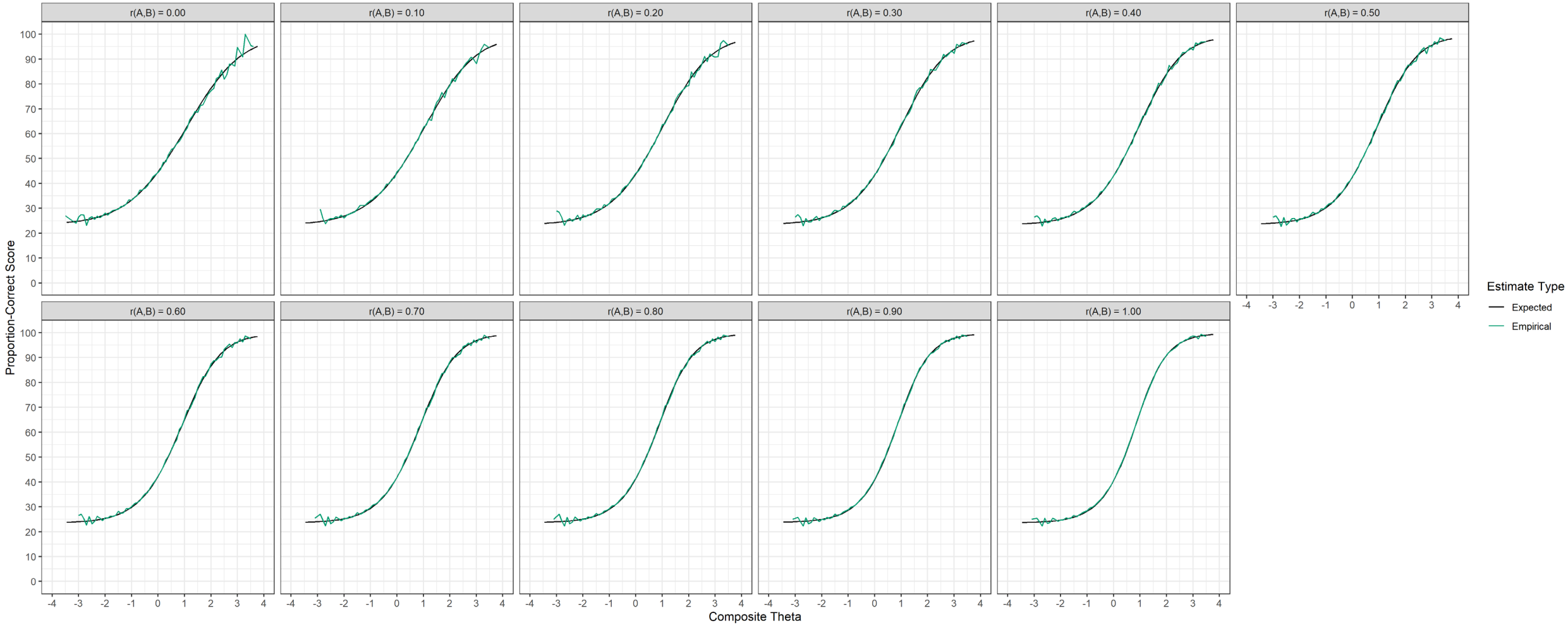
OPA
OFFICE OF PEOPLE ANALYTICS

Expected and Empirical TCC Alignment: Test A



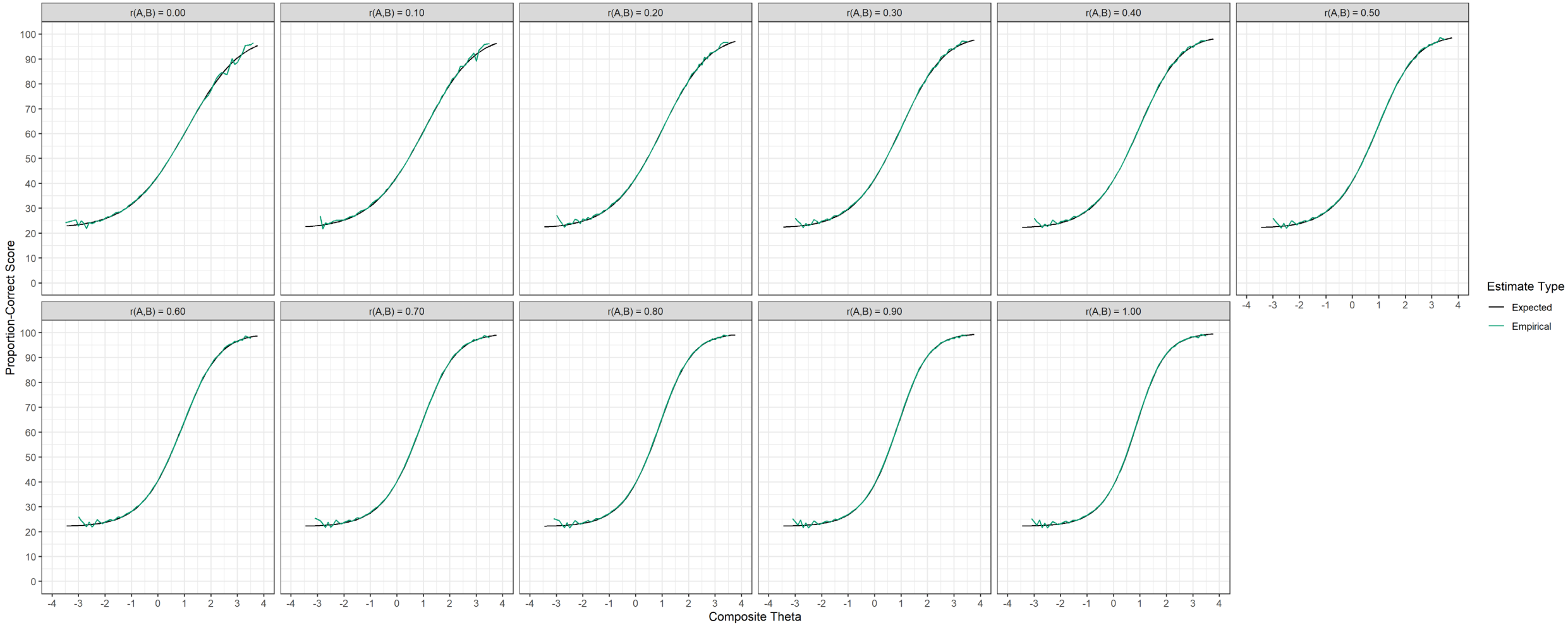
OPA
OFFICE OF PEOPLE ANALYTICS

Expected and Empirical TCC Alignment: Test B



OPA
OFFICE OF PEOPLE ANALYTICS

Expected and Empirical TCC Alignment: Combined Test

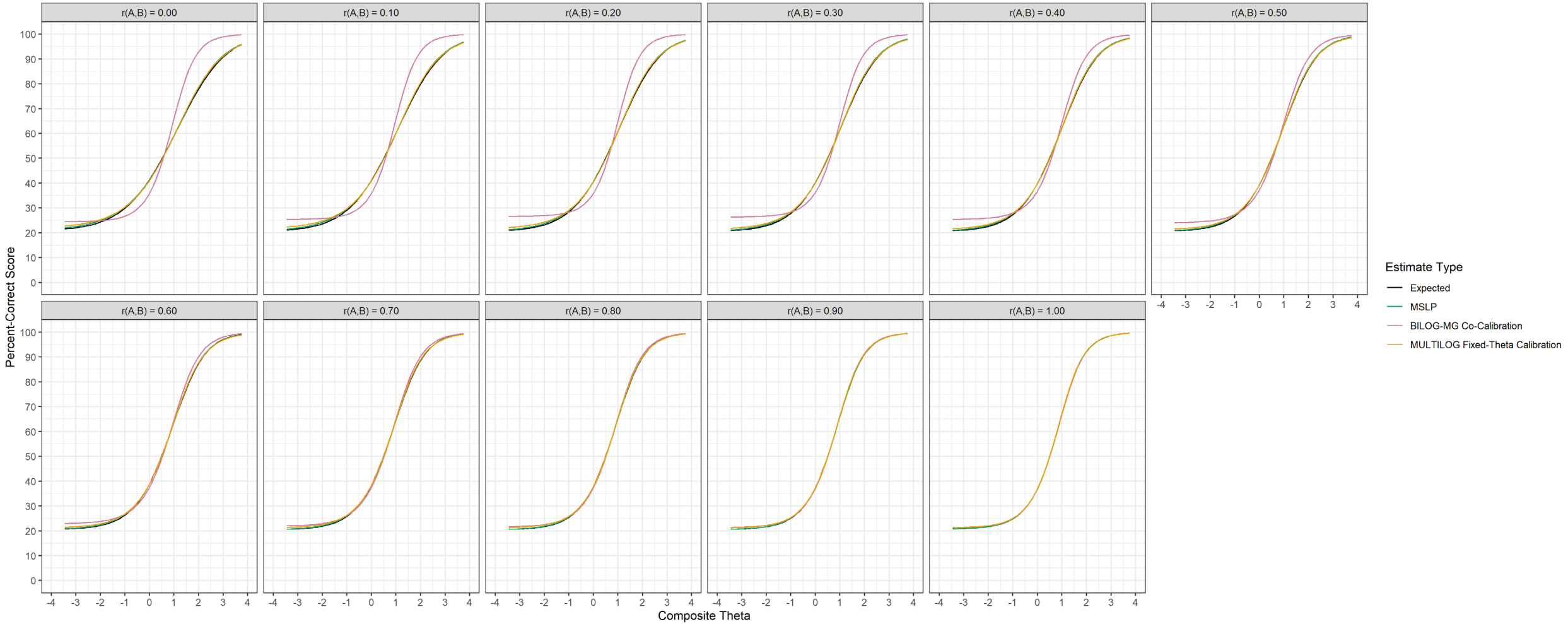


OPA
OFFICE OF PEOPLE ANALYTICS

TCC Comparisons for Rescaling/Calibration Methods

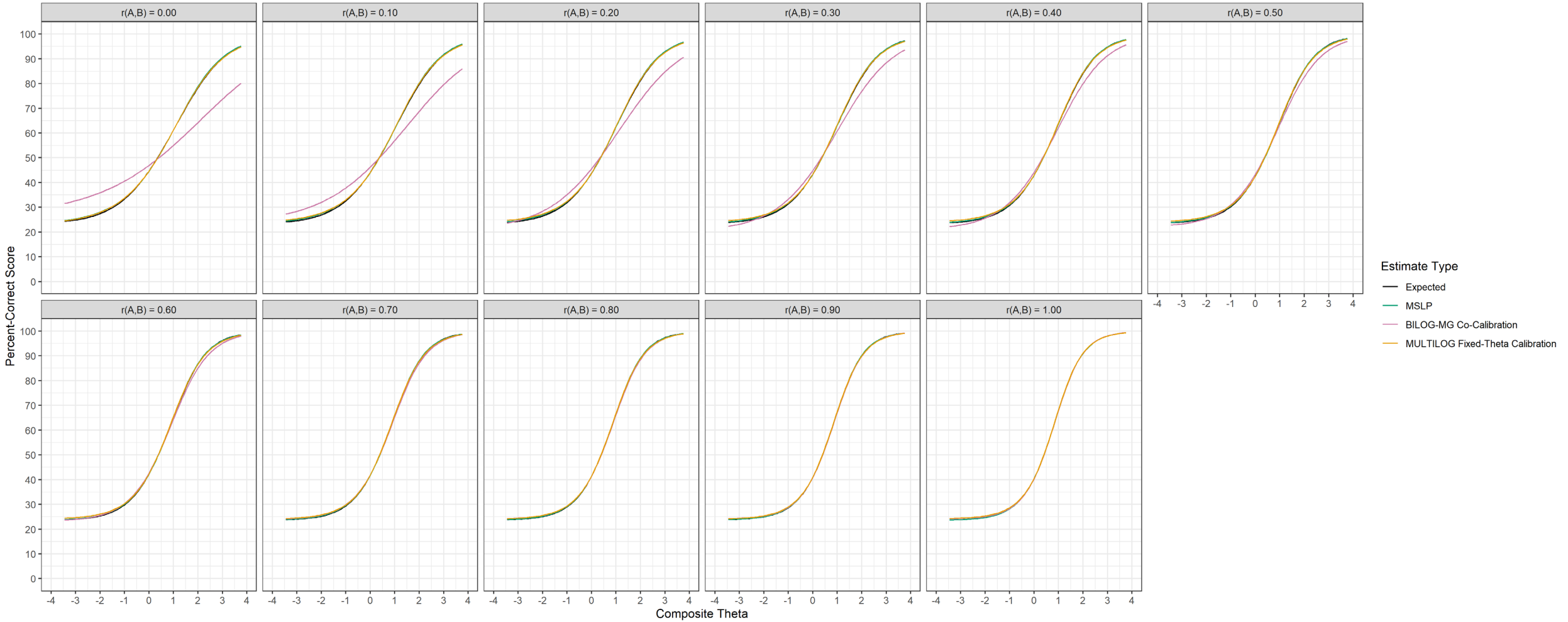
OPA
OFFICE OF PEOPLE ANALYTICS

TCC Comparisons for Rescaling/Calibration Methods: Test A



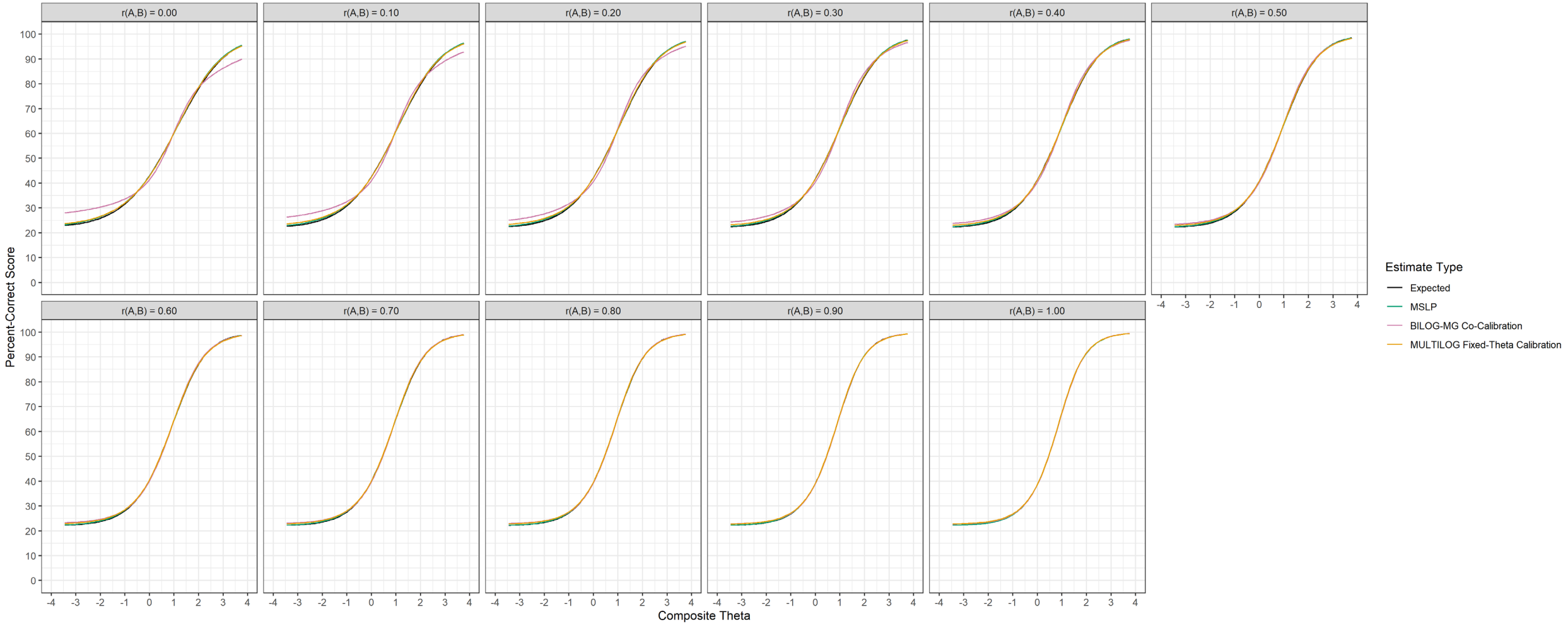
OPA
OFFICE OF PEOPLE ANALYTICS

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

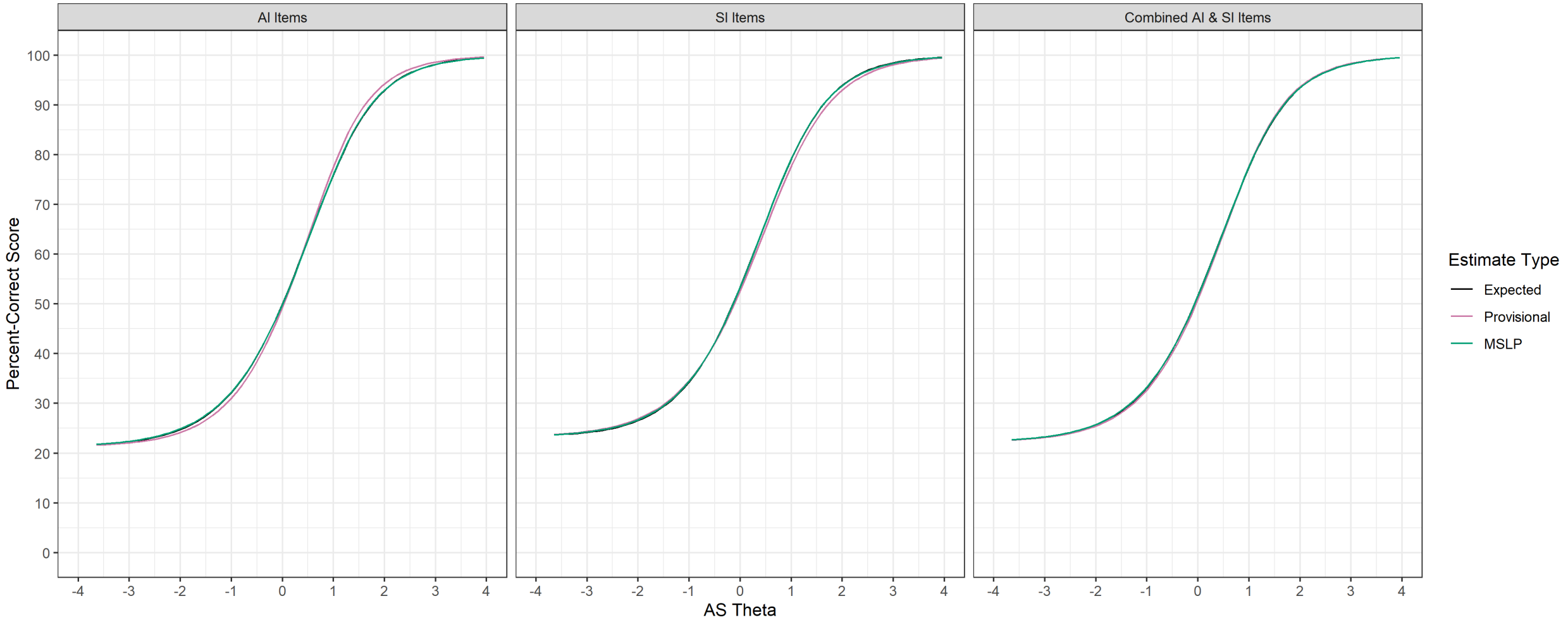
OPA
OFFICE OF PEOPLE ANALYTICS

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

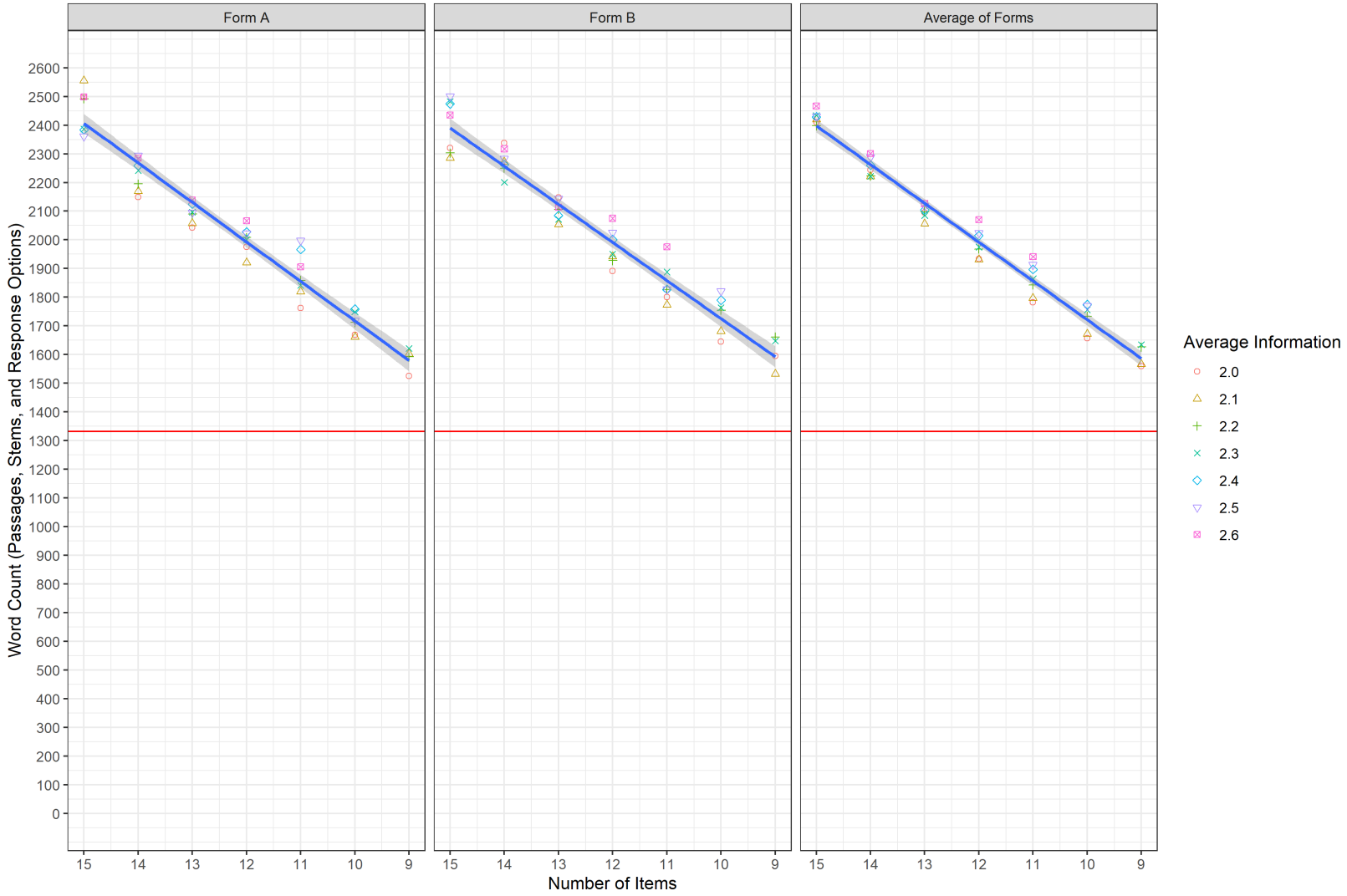
- Compared to past testlet-based PC sections, constructing new PC sections from items with stand-alone reading passages requires reducing the number of items administered to control the reading load
- When shortening a test, there are two primary objectives to satisfy:
  - Maintain an acceptable level 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

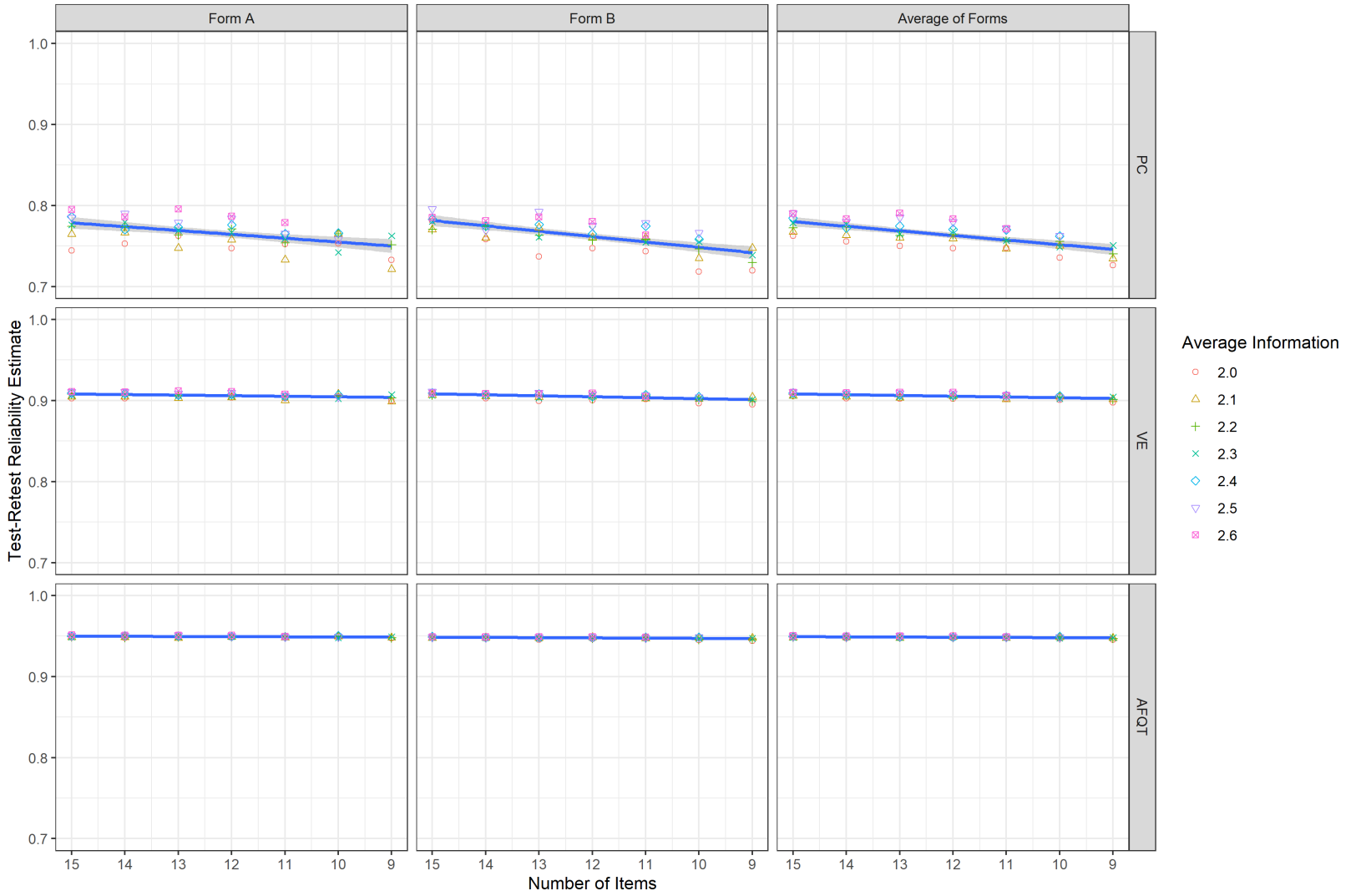
Word Counts for PC Item Sets with Varied Information



OPA
OFFICE OF PEOPLE ANALYTICS

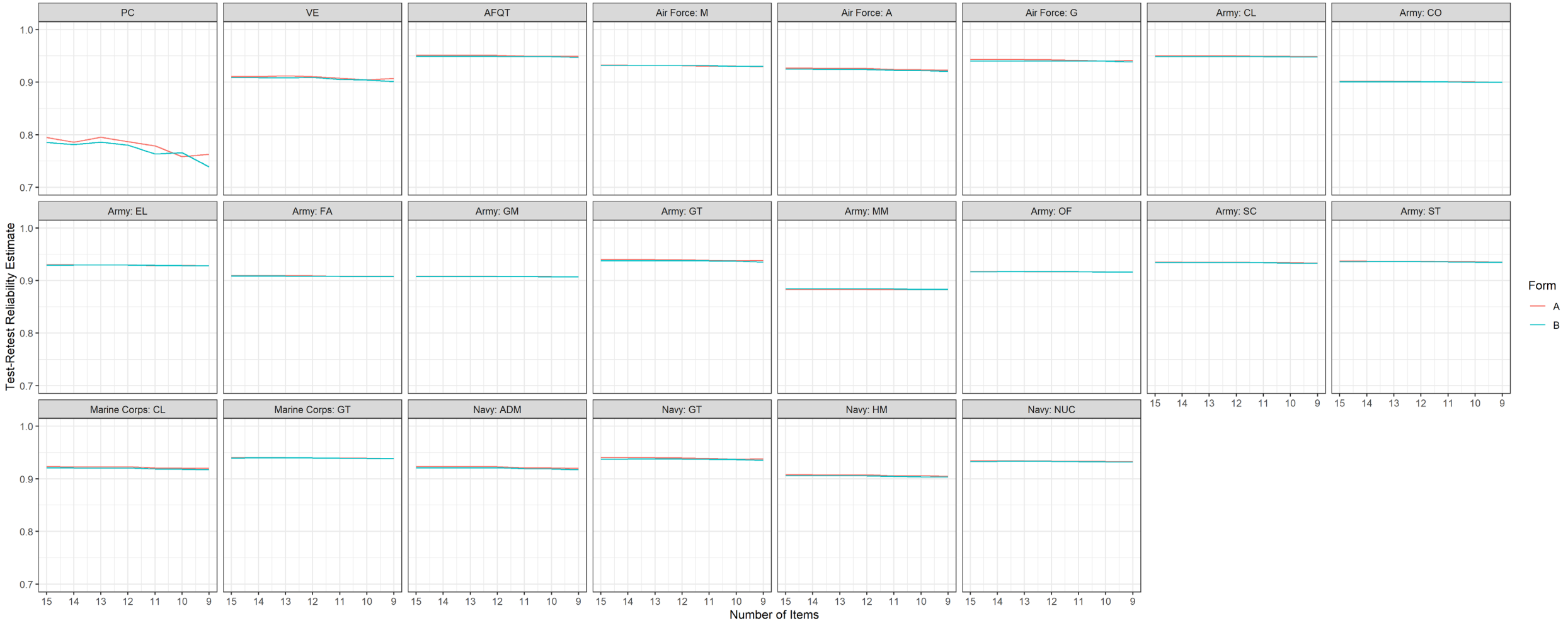
Note: Red horizontal line = average word count of PC subtests from P&P-ASVAB Forms 23 – 26

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

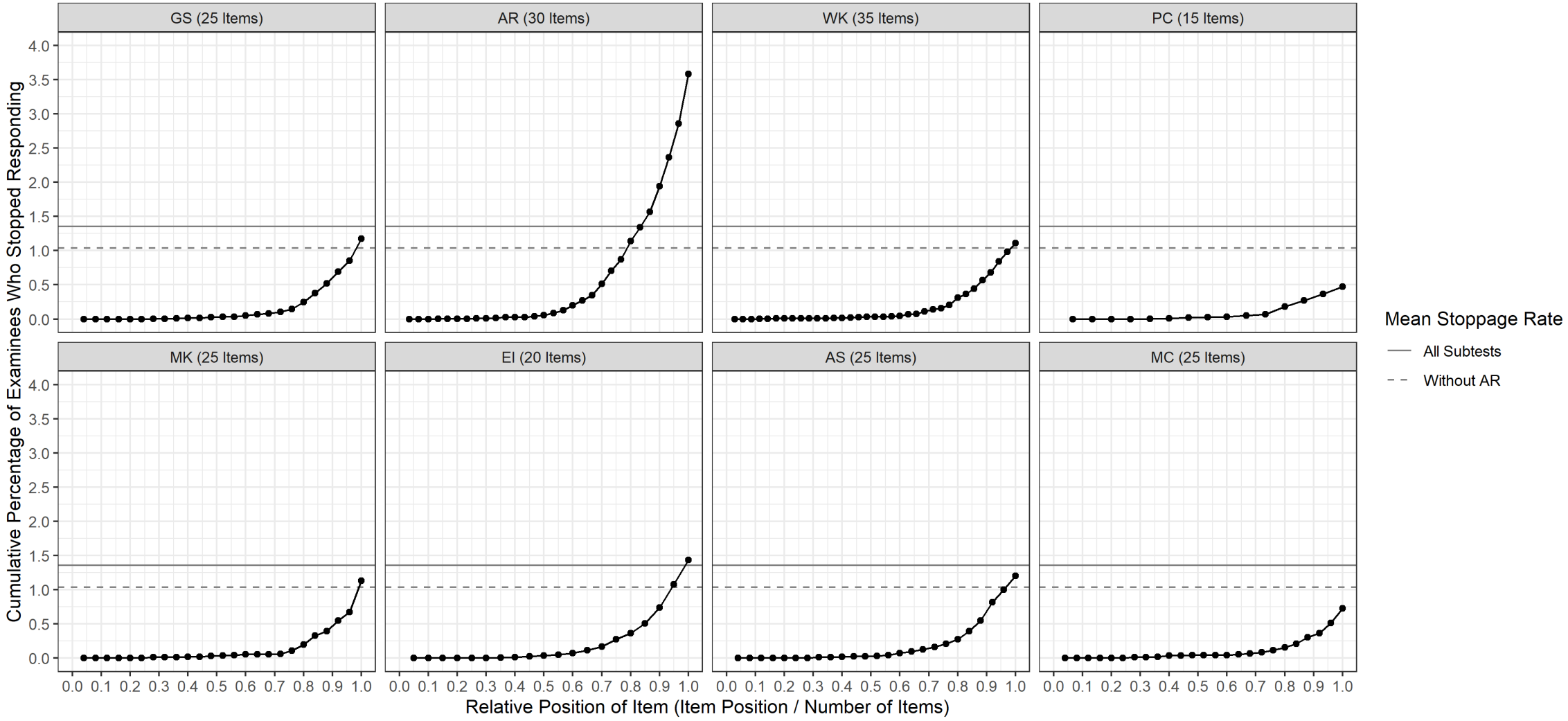
Length-Reduction Analyses for Arithmetic Reasoning (AR)

OPA
OFFICE OF PEOPLE ANALYTICS

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

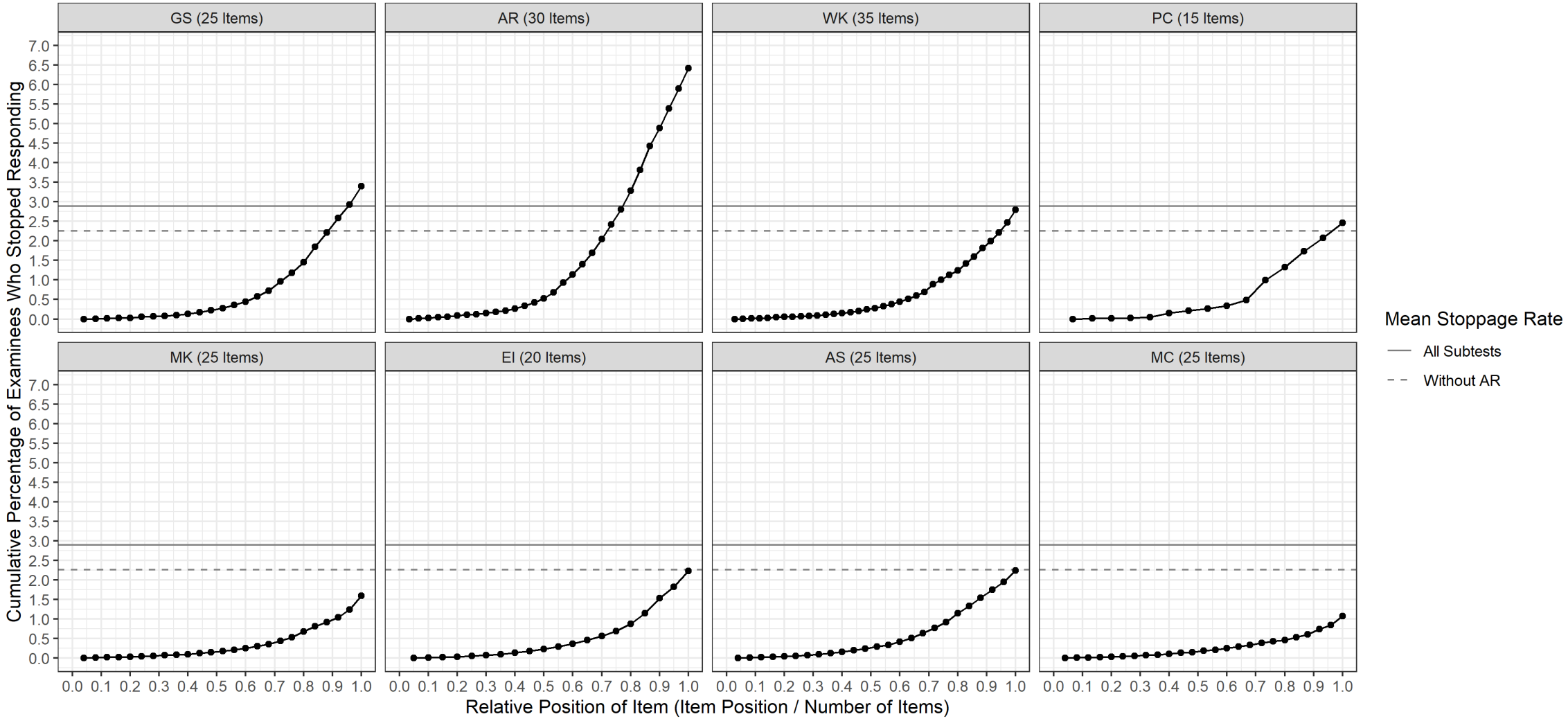
- As we explored the impact of the recommended changes to PC on time limits, we benchmarked whether past P&P-ASVAB PC sections appeared to have sufficient time limits
- We examined trends from all P&P-ASVAB subtests to provide context for the PC trends
- We found that AR appeared to be much more speeded 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

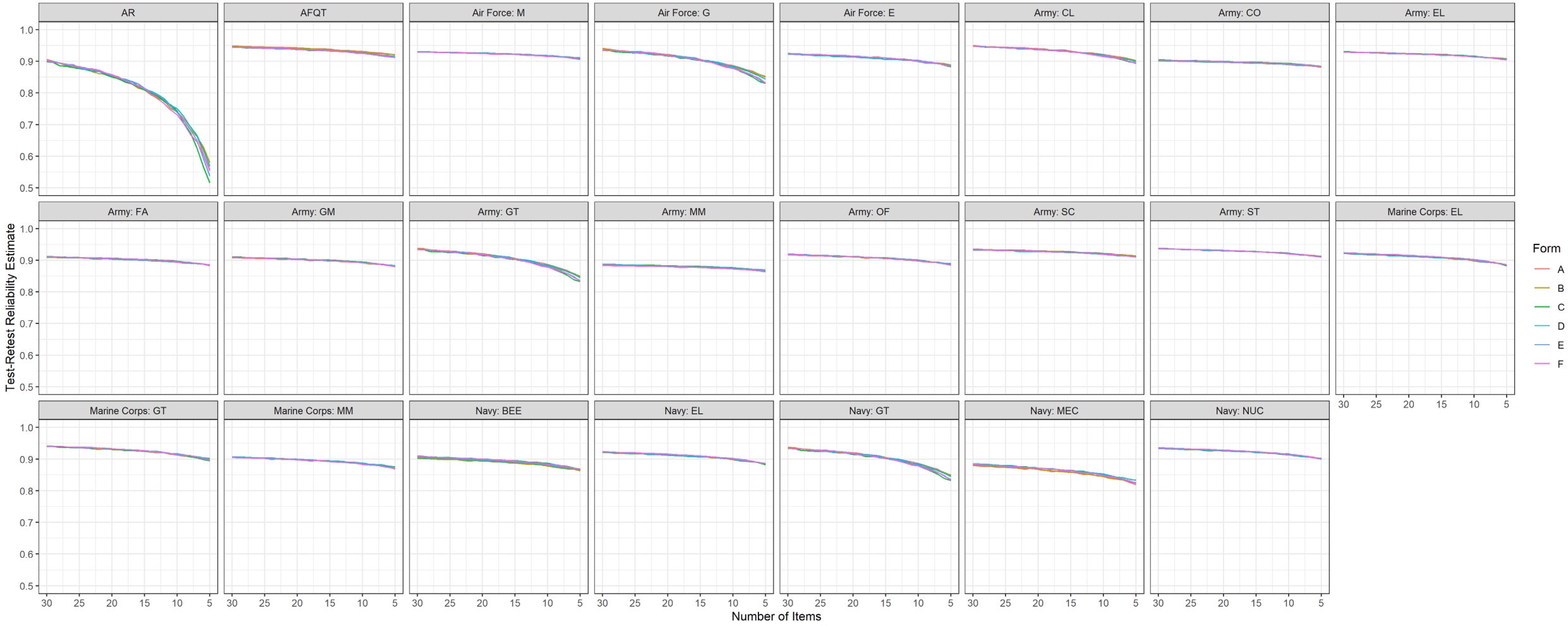


OPA
OFFICE OF PEOPLE ANALYTICS

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

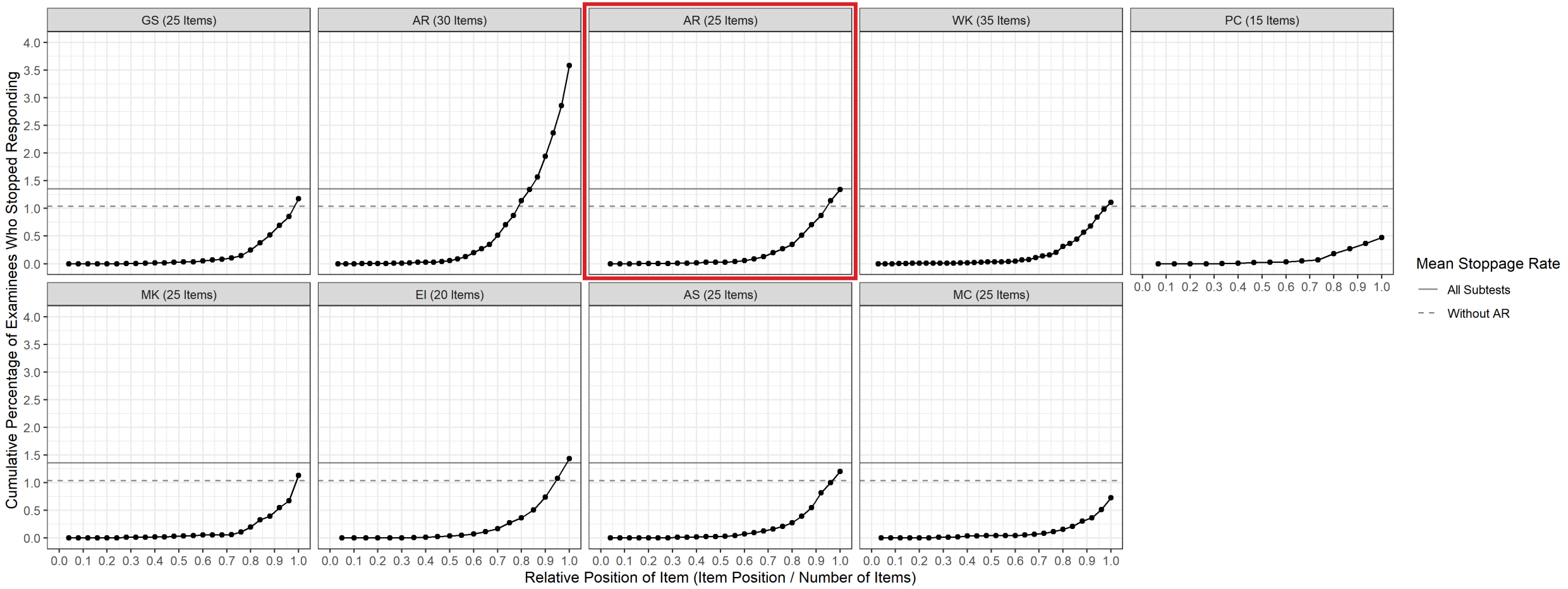


OPA
OFFICE OF PEOPLE ANALYTICS

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

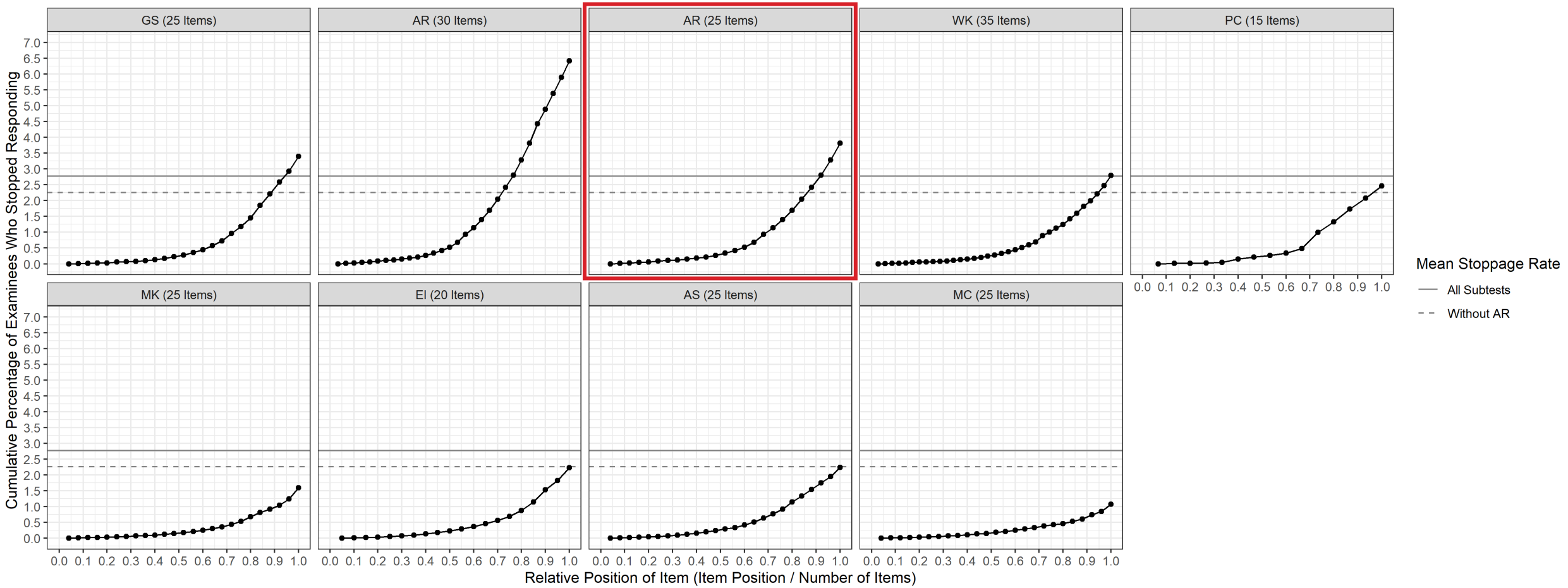
- Based on simulated reliability estimates, 25 items appear to preserve reliability across all scores we evaluated
  - This length also works well for covering all test blueprint categories
  - As with PC, AR scores are never used in isolation for selection or classification into the military, so the 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

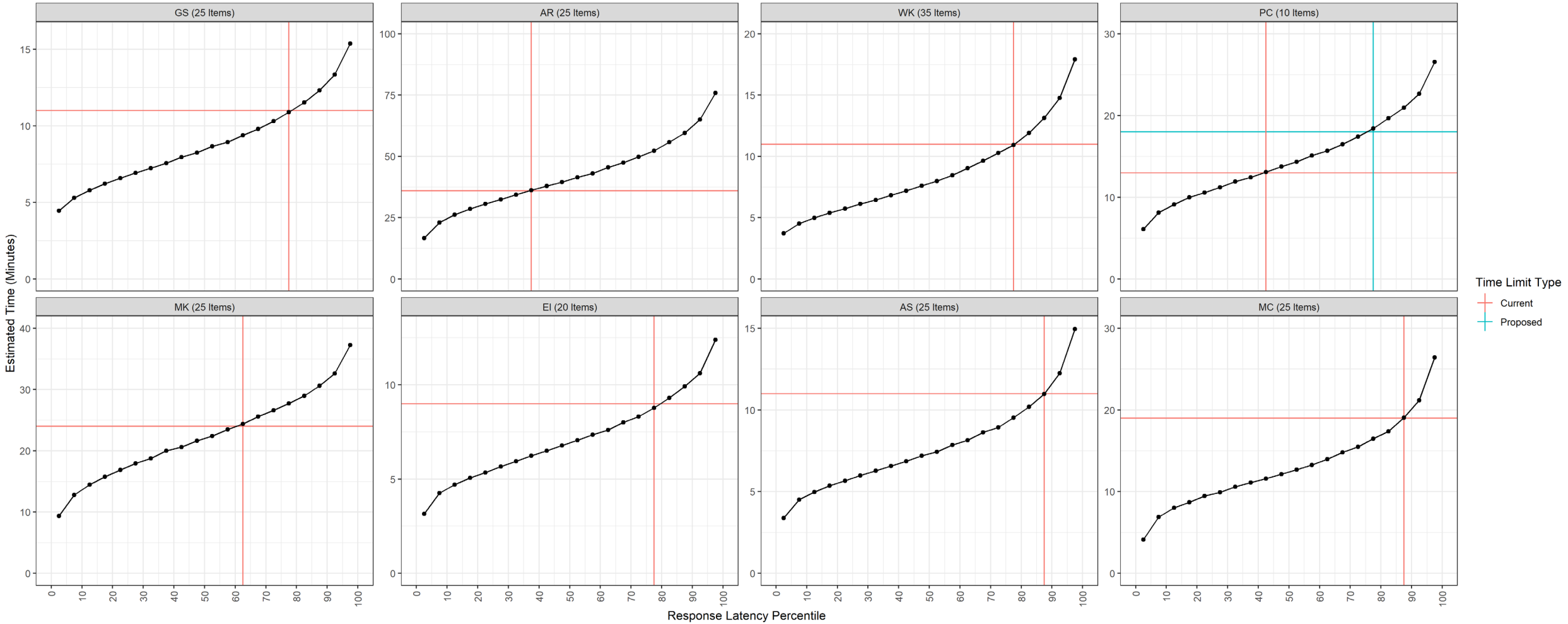
Context for Interpreting the Projected Time Requirements

- 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

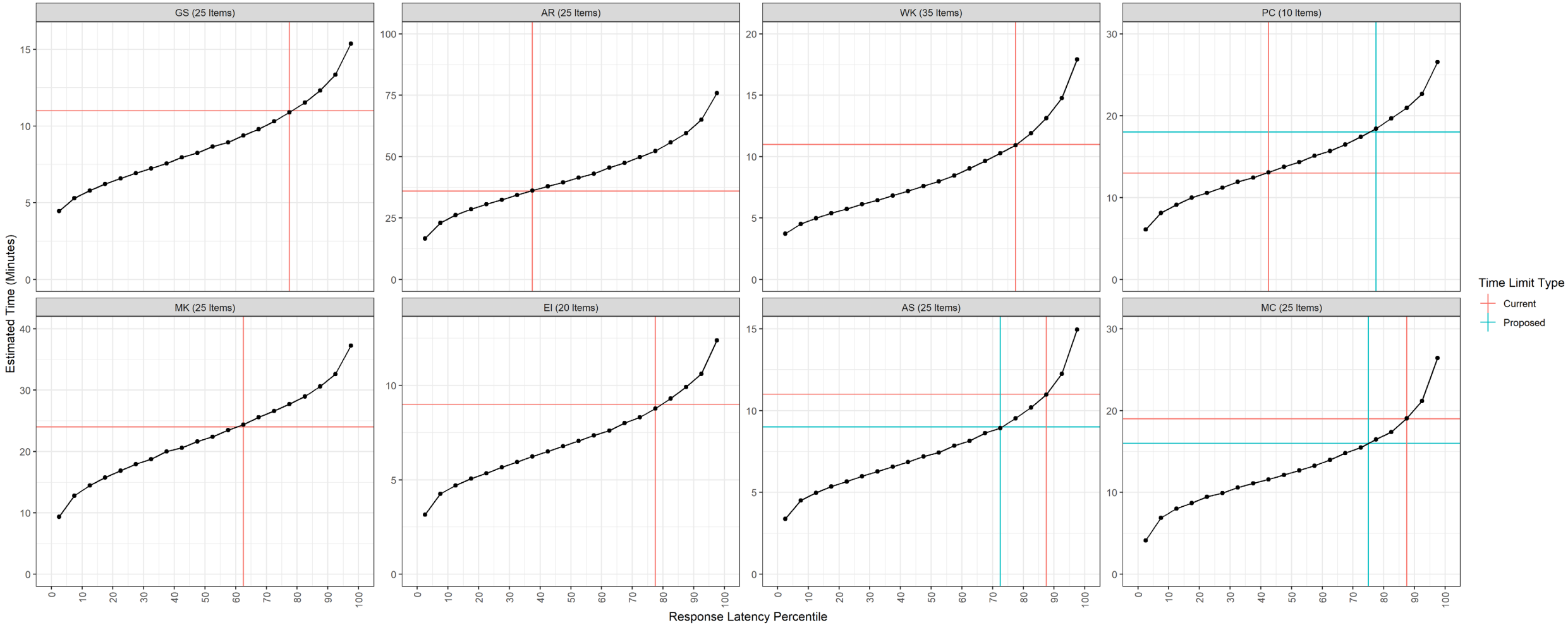
- **Option A**
  - Increase time limit for PC without altering other subtests' time limits
- **Option B (recommended)**
  - Offset the increased PC time limit by reducing time limits for AS and MC

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