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

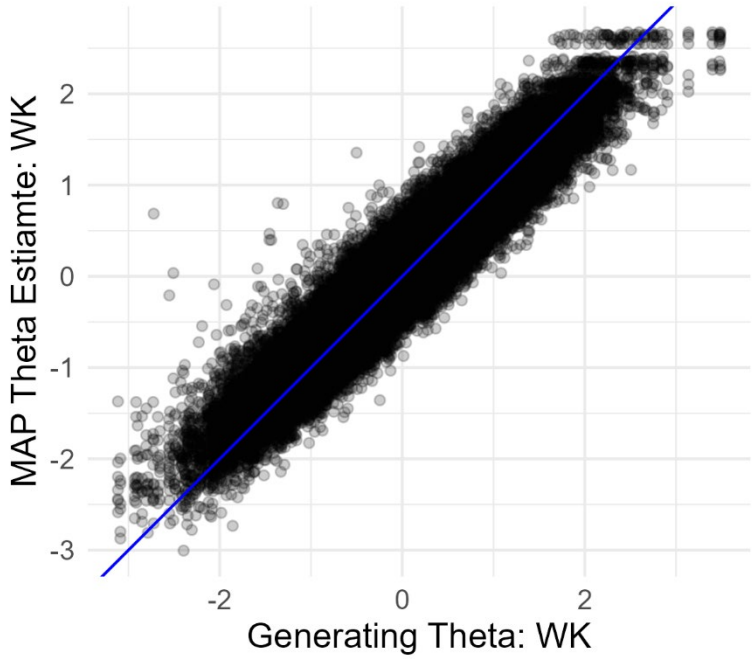
- How do empirically informed, copula-based deflections to item parameter estimates affect composite score distributions for CAT-ASVAB?
- How do biased difficulty parameter deflections affect composite score distributions for CAT-ASVAB?
- If effects are present, which composites and which ranges of those 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

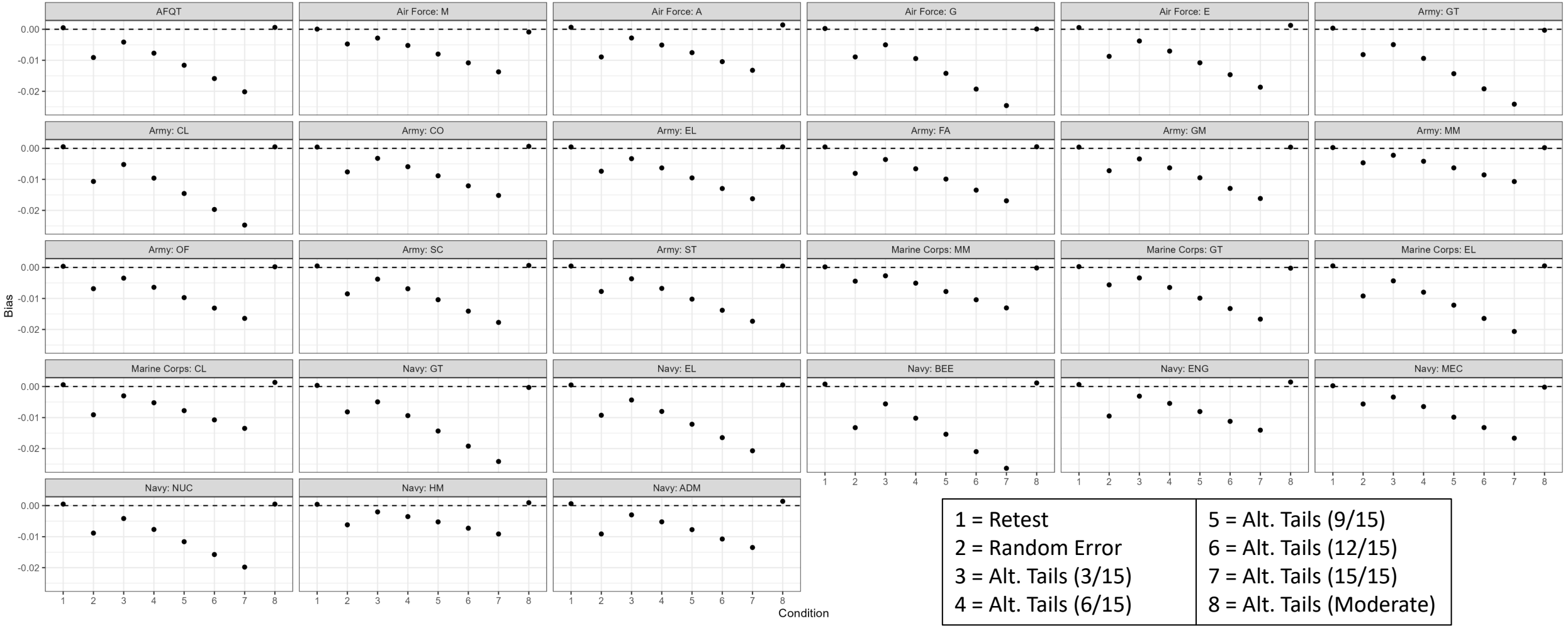
Conditions (Part 3)

- 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

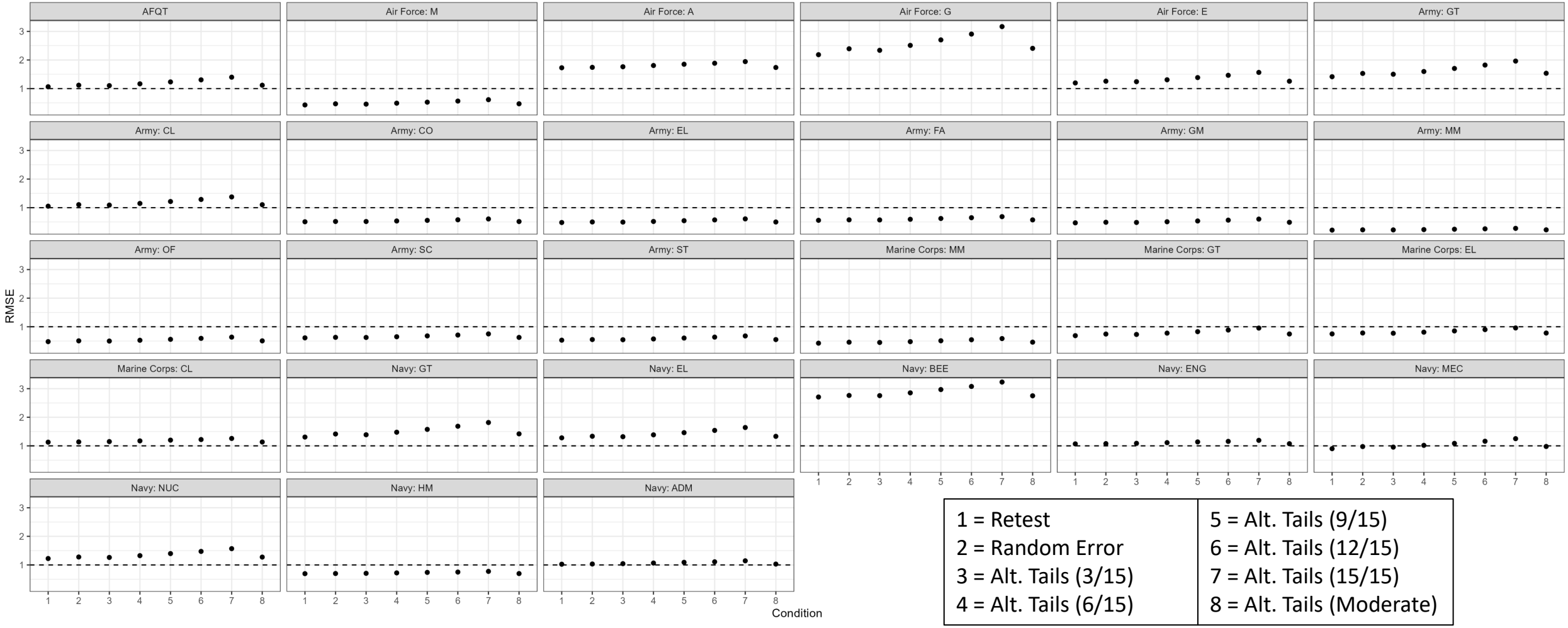
Bias per Composite and Condition



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

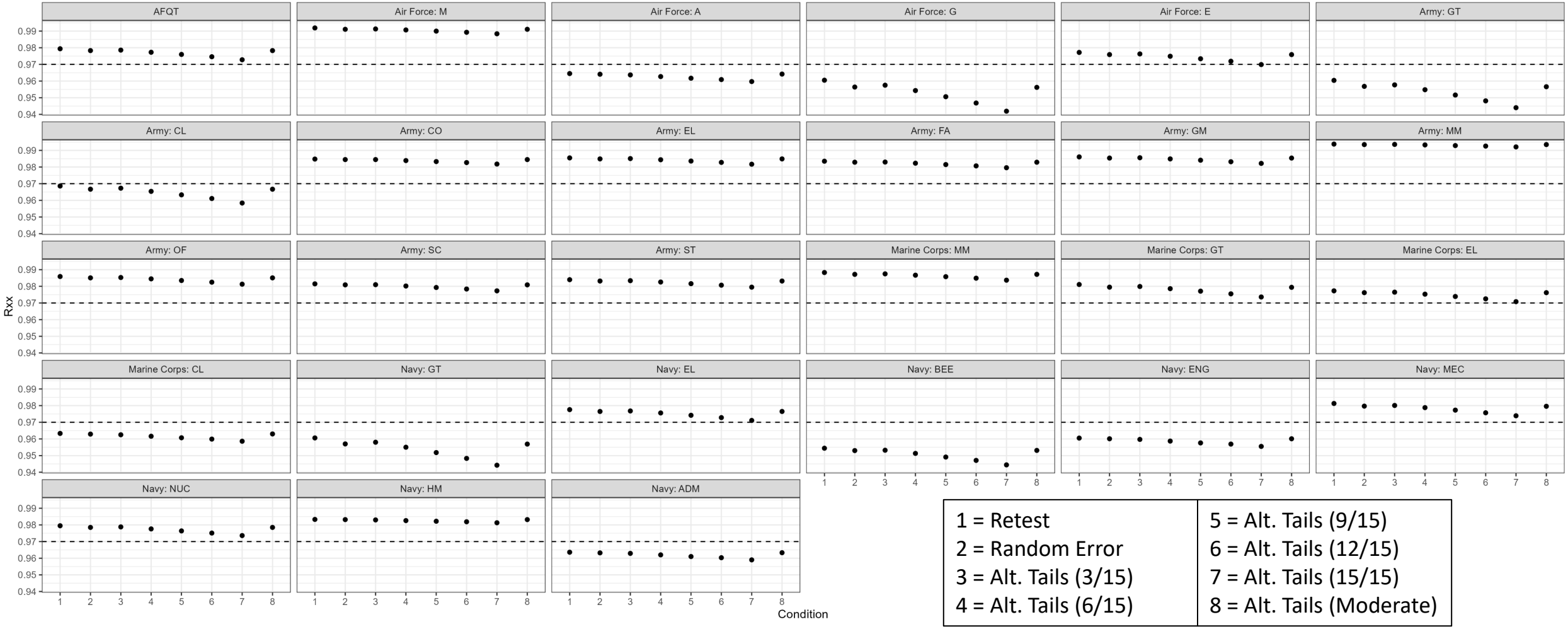
Root Mean Square Error (RMSE) per Composite and Condition



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)

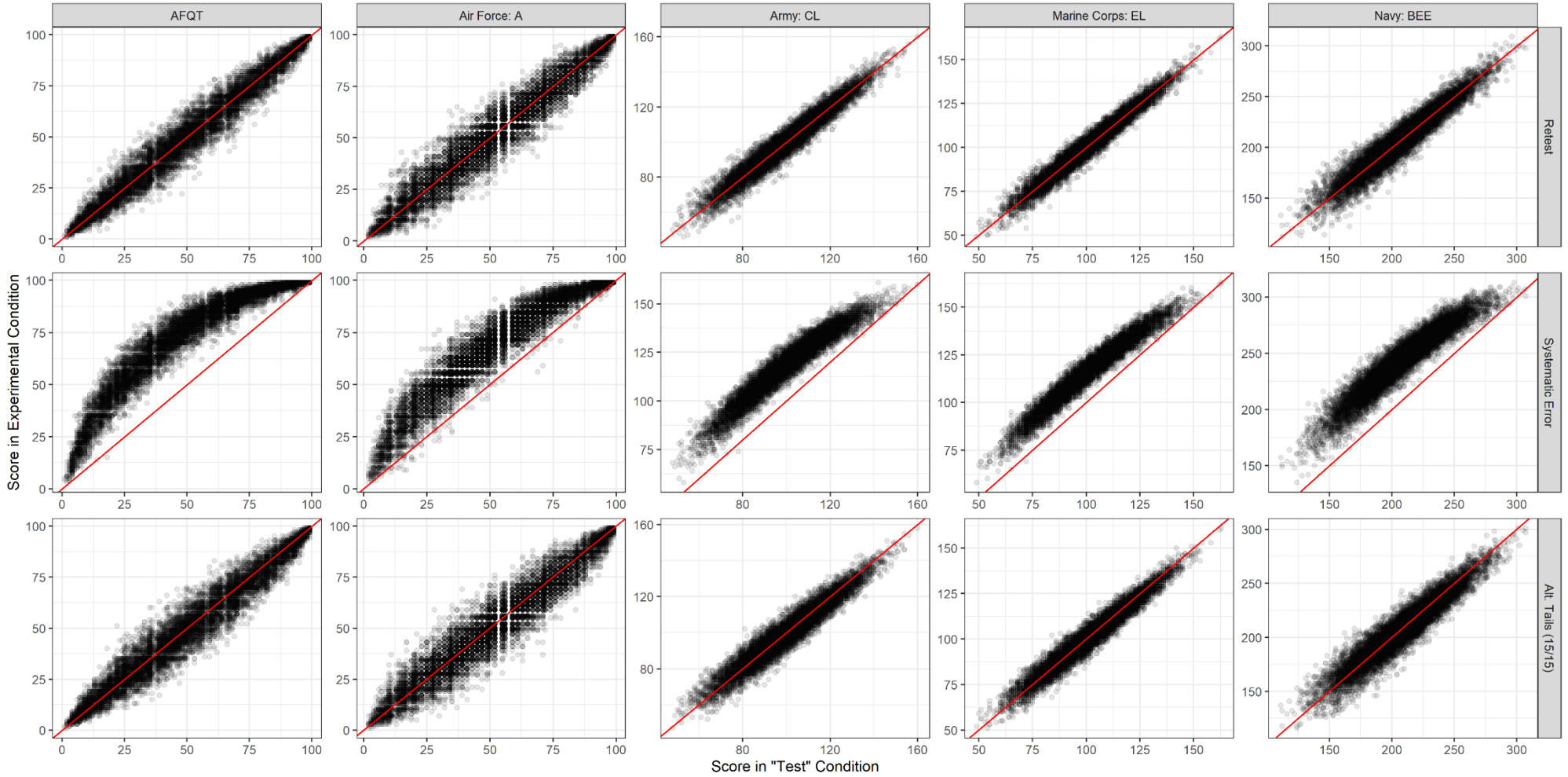
OPA
OFFICE OF PEOPLE ANALYTICS

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

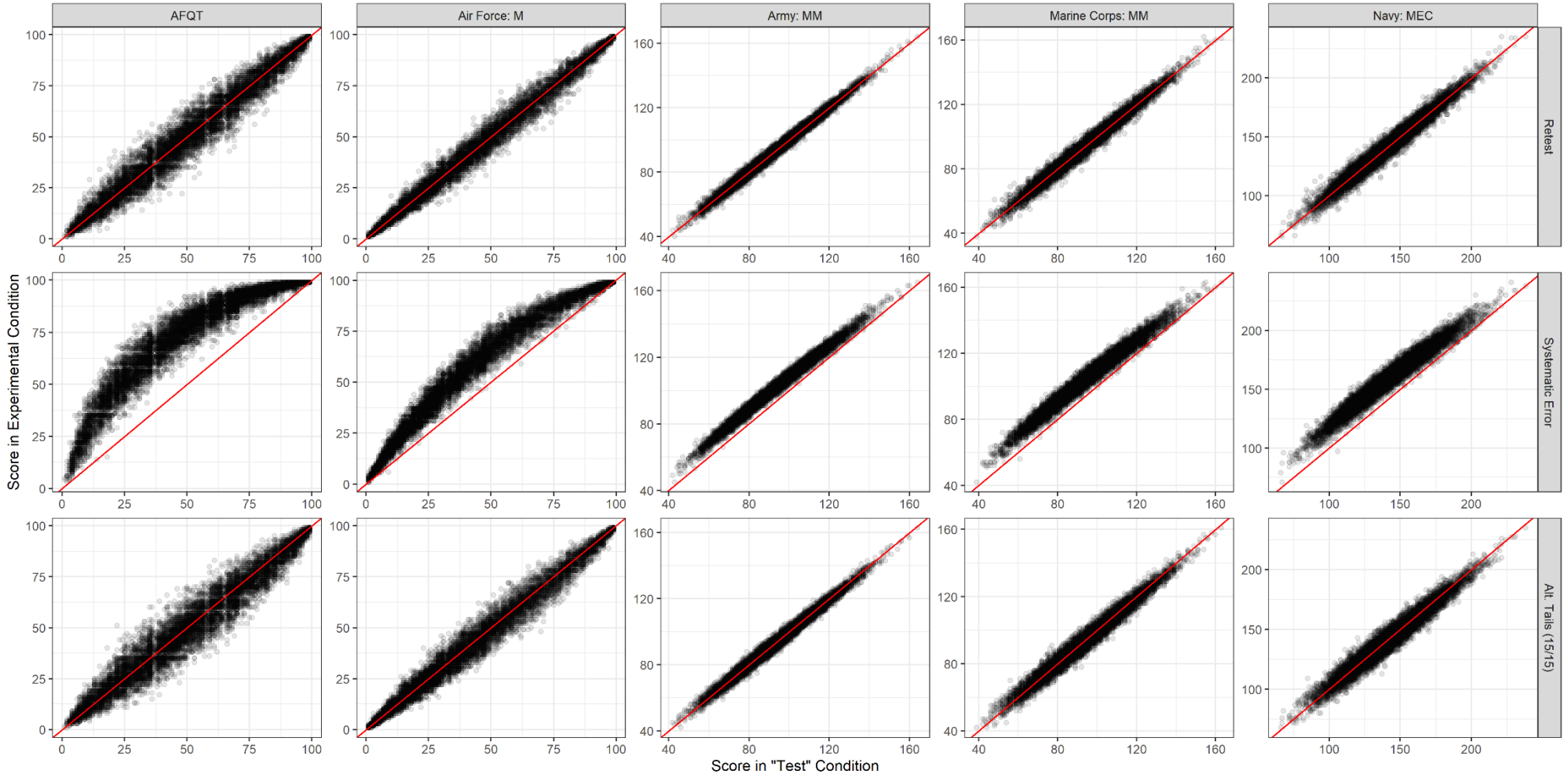
Scatterplots of Experimental Conditions vs. Test Condition for AFQT and the Most Math-Heavy Composites per Service



OPA
OFFICE OF PEOPLE ANALYTICS

Note: Each plot shows a random sample of 10,000 cases

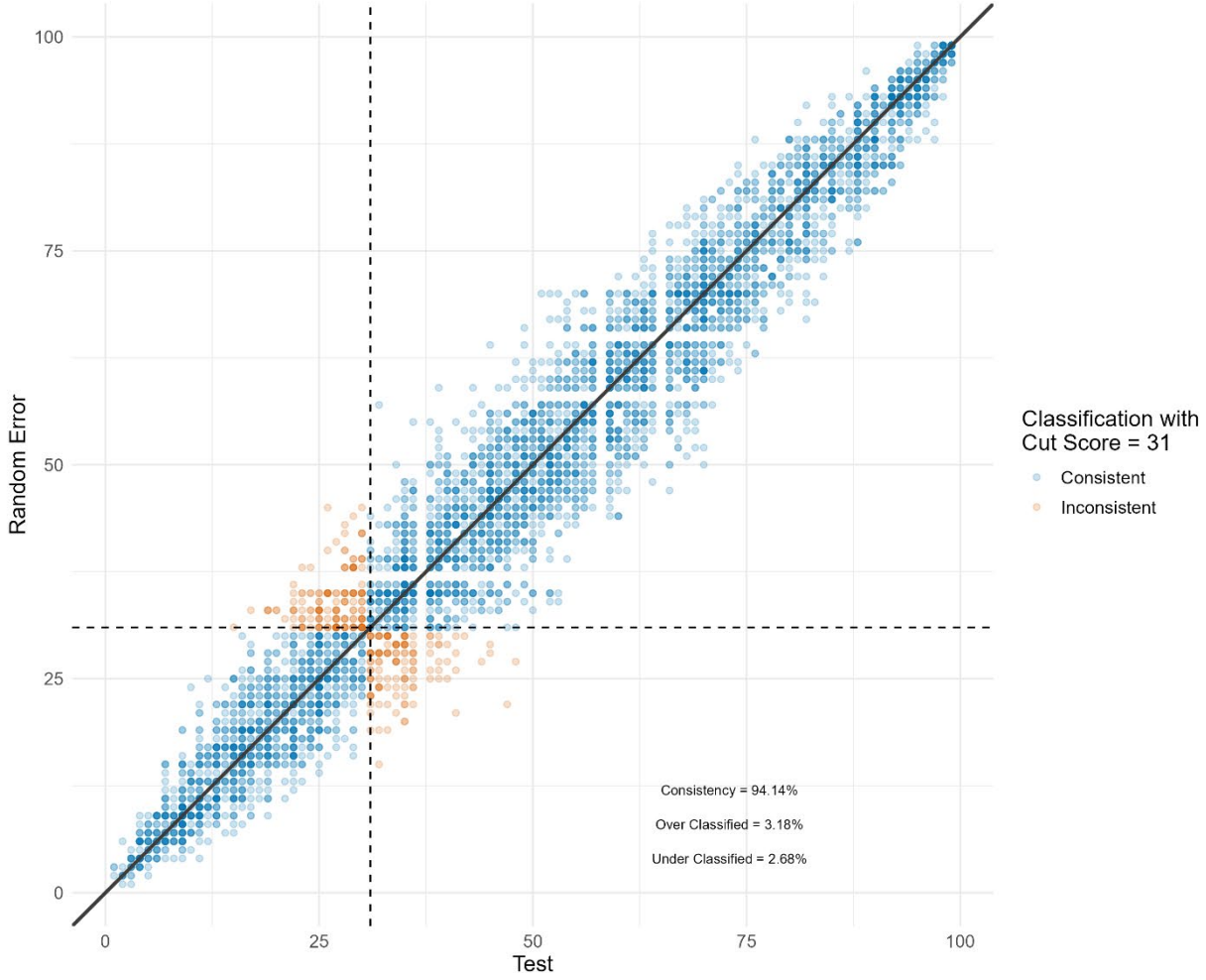
Scatterplots of Experimental Conditions vs. Test Condition for AFQT and the Least Math-Heavy Composites per Service

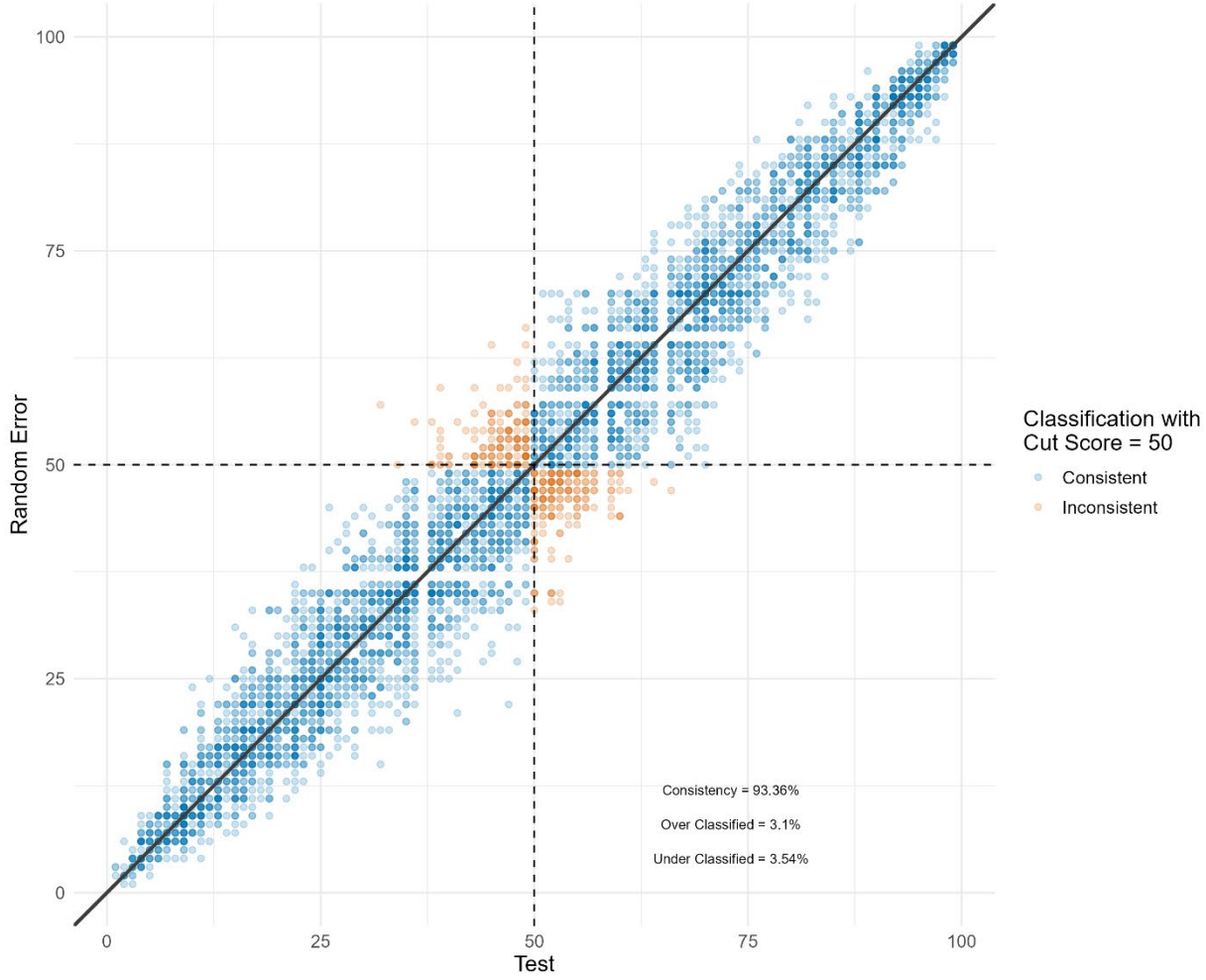


OPA
OFFICE OF PEOPLE ANALYTICS

Note: Each plot shows a random sample of 10,000 cases

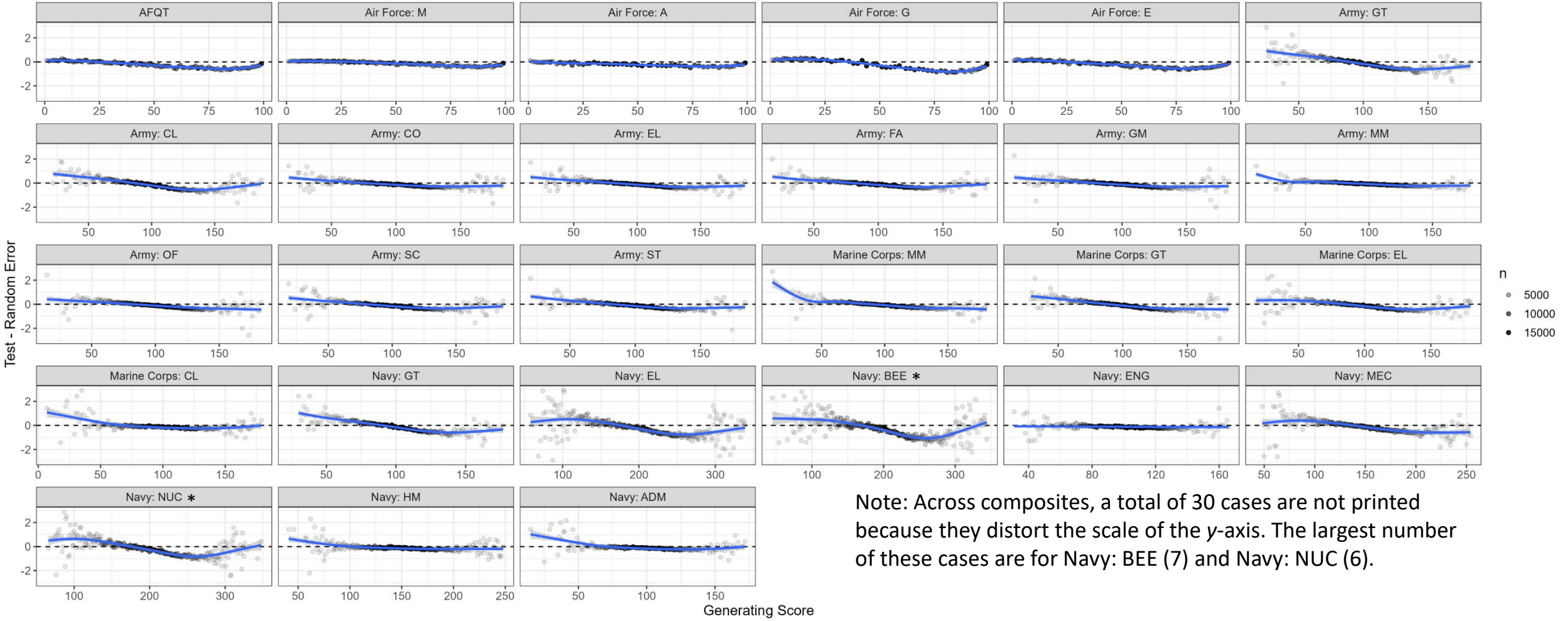
AFQT Classification for Test/Random Error for Two Cut Scores





OPA
OFFICE OF PEOPLE ANALYTICS

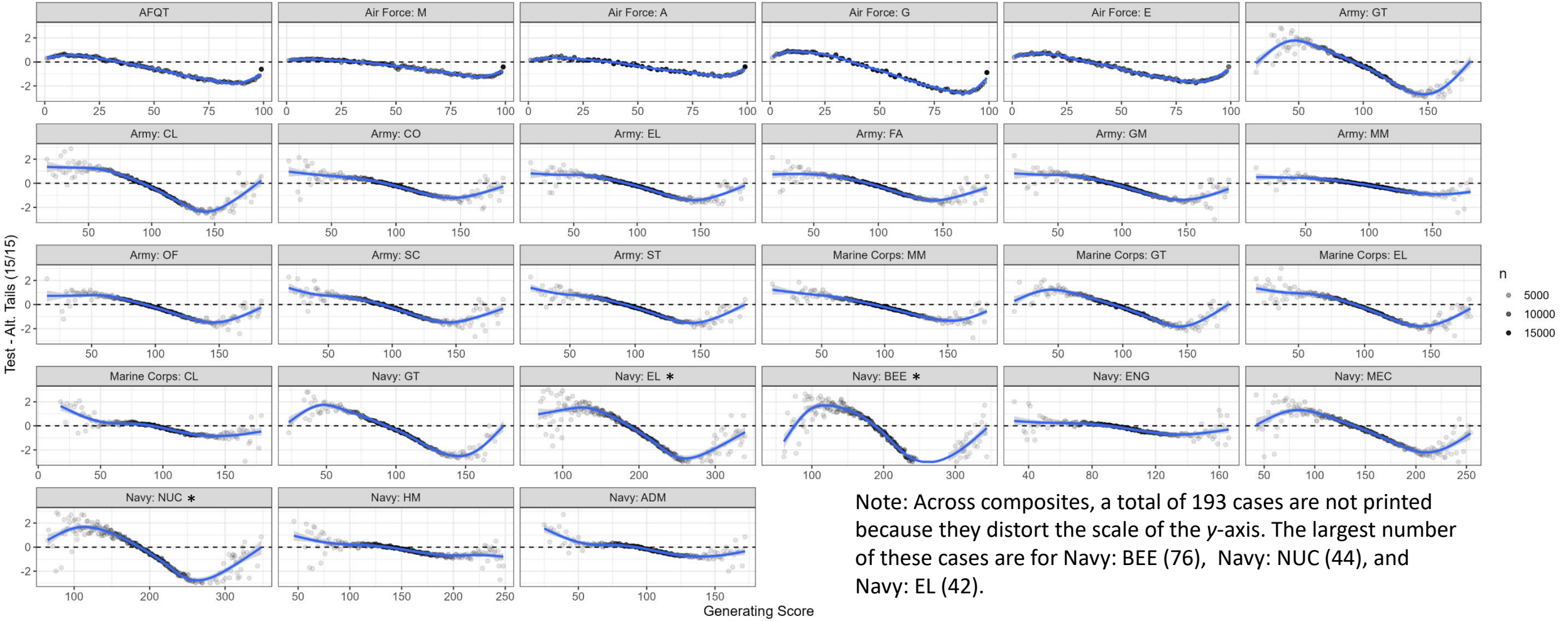
Mean Score Conditional Bias for All Composites: Random Error



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

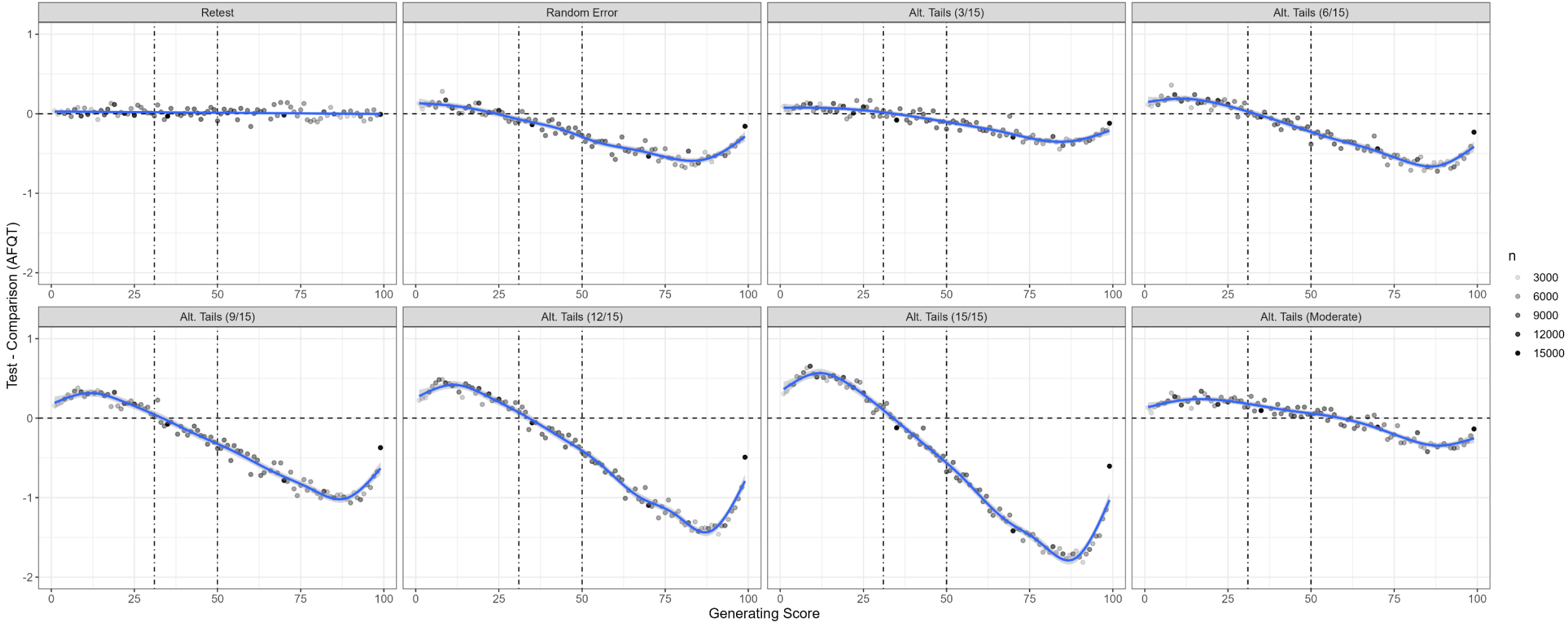
Mean Score Conditional Bias for All Composites: Alternating Tails 15/15



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).

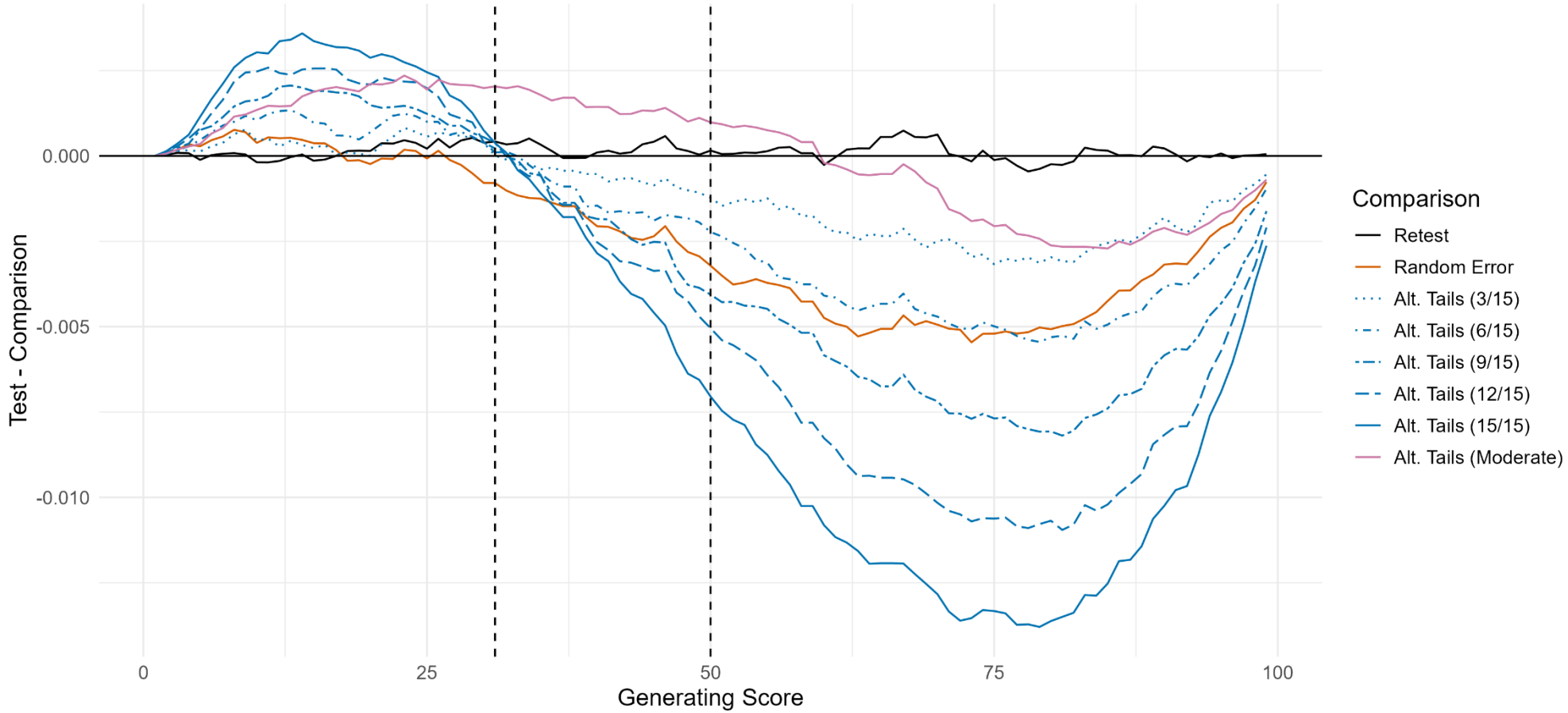
OPA
OFFICE OF PEOPLE ANALYTICS

Mean Score Conditional Bias for AFQT across Conditions



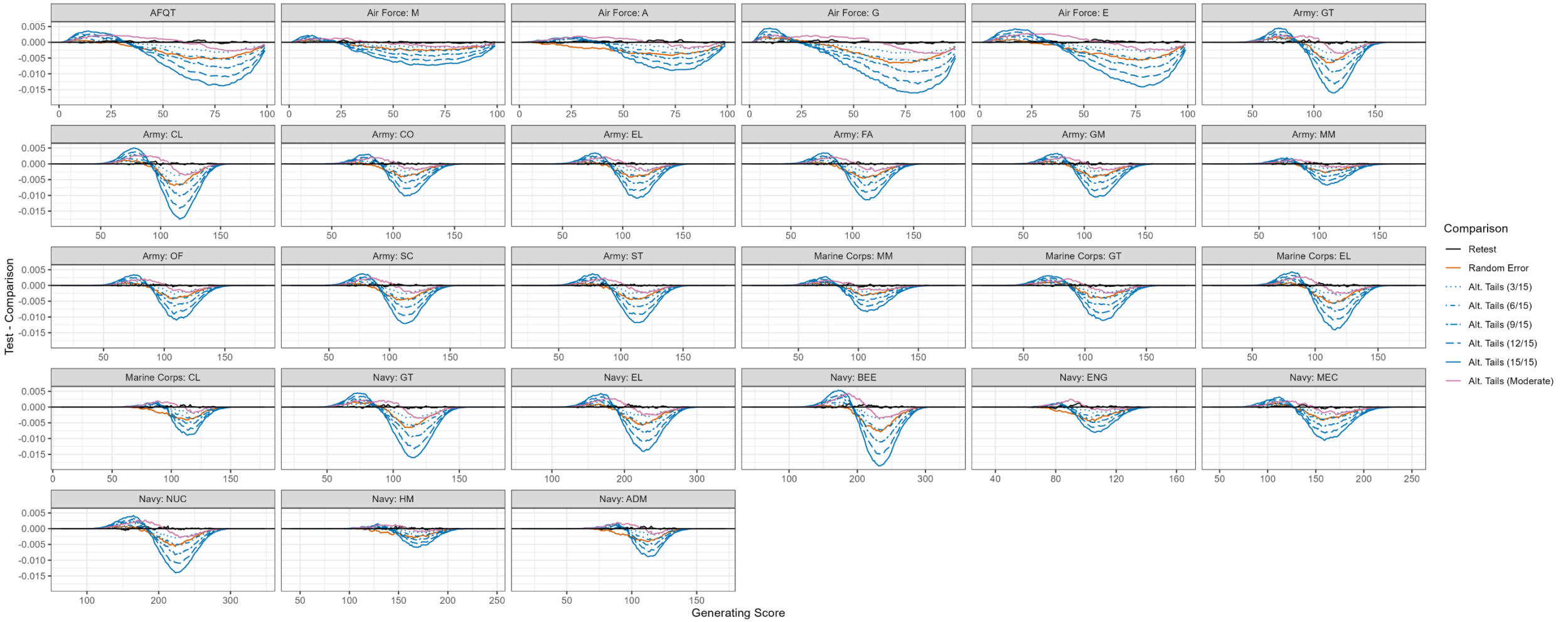
OPA
OFFICE OF PEOPLE ANALYTICS

Qualification Rate Differences per Condition for AFQT



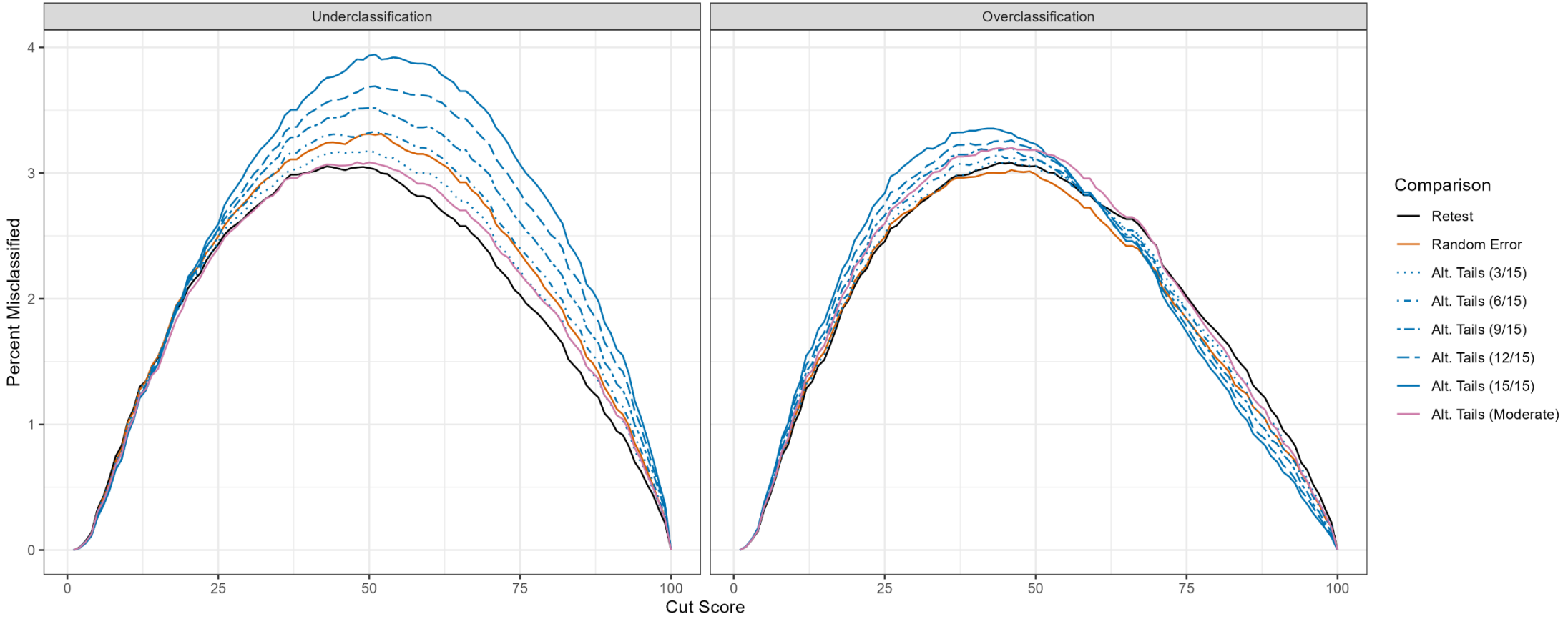
OPA
OFFICE OF PEOPLE ANALYTICS

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